Biostatistics for Health Sciences

Biostatistics for Health Sciences

Andrew Hayen

SPRINGER Med Press
www.springermedpress.com

SPRINGER Med Press

Published by Springer Med Press
1202 N. Orange Street Suit #600
Wilmington, DE19801 USA

© 2023 by Springer Med Press

ISBN: 979-8-88626-275-9

Biostatistics for Health Sciences

Andrew Hayen

This work is subject to copyright. All rights are reserved by the publisher. No part of this publication may be reproduced or transmitted in any form or by any means, electronic or mechanical, including photocopying, recording, or any information storage and retrieval system, without prior permission in writing from the publisher.

Notice

Practitioners and researchers must always rely on their experience and knowledge in evaluating and using information, methods, compounds, or experiments described herein. Because of rapid advances in the medical sciences, in particular, independent verification of diagnoses and drug dosages should be made. To the fullest extent of the law, no responsibility is assumed by Springer Med Press, authors, editors, or contributors for any injury and damage to persons or property as a matter of products liability, negligence, or otherwise, or from any use or operation of any methods, products, instructions, or ideas contained in the material herein.

Library of Congress Cataloging-in-Publication Data
A catalog record for this book is available from Library of Congress

For more information regarding Springer Med Press and its products, please visit the publisher's website
www.springermedpress.com

Contents

PREFACE

The use of statistics in a variety of biological areas is known as biostatistics, including the planning of biological experiments, particularly those used in agriculture, medicine, pharmacy, and fisheries; the gathering, summarising, and data analysis from such experiments; as well as the interpretation and drawing of conclusions from the findings.

Medical biostatistics, which is mostly focused on medicine and health, is a significant subfield. A significant component of many contemporary biological theories is biostatistical modelling. Since the outset, genetic researchers have employed statistical ideas to interpret the outcomes of their experiments. Even some genetics scientists helped develop new statistical techniques and tools, which advanced the field. Mendel employed statistics to interpret the data gathered from his genetic investigations of the genetic segregation patterns in families of peas. After Mendel's work on Mendelian inheritance was rediscovered in the early 1900s, there were knowledge gaps between genetics and Darwinian evolution. To supplement Mendel's findings with information about people, Francis Galton put up a new model that involved an unending succession of fractions of the inheritance that came from various ancestors. This hypothesis was dubbed the "Law of Ancestral Heredity" by the author. William Bateson, who adheres to Mendel's findings, strongly disagreed with him on the premise that genetic inheritance comes entirely from the parents, with half coming from each of them. This sparked a heated argument between the so-called mendelian, who support Bateson (and Mendel) ideas, and the biometricians who support Galton's ideas, people as Walter Weldon, Arthur Dukinfield Darbishire, and Karl Pearson. Later, Mendel's theories took hold since biometricians were unable to replicate Galton's findings in several trials. The 1930s saw the resolution of these disparities and the creation of the neo-Darwinian contemporary evolutionary synthesis thanks to models based on statistical reasoning.

-Author

1

Biostatistics

Biostatistics is the application of statistics to a wide range of topics in biology. It encompasses the design of biological experiments, especially in medicine, pharmacy, agriculture and fishery; the collection, summarization, and analysis of data from those experiments; and the interpretation of, and inference from, the results. A major branch is medical biostatistics, which is exclusively concerned with medicine and health.

HISTORY

Biostatistics and Genetics

Biostatistical modeling forms an important part of numerous modern biological theories. Genetics studies, since its beginning, used statistical concepts to understand observed experimental results. Some genetics scientists even contributed with statistical advances with the development of methods and tools. Gregor Mendel started the genetics studies investigating genetics segregation patterns in families of peas and used statistics to explain the collected data. In the early 1900s, after the rediscovery of Mendel's Mendelian inheritance work, there were gaps in understanding between genetics and evolutionary Darwinism. Francis Galton tried to expand Mendel's discoveries with human data and proposed a different model with fractions of the heredity coming from each ancestral composing a infinite series.

He called this theory of " Law of Ancestral Heredity". His ideas were strong disagreed by William Bateson, who follow Mendel's conclusions, that genetic inheritance were exclusively from the parents, half from each of them. This led to a vigorous debate between the ones called biometricians, who support Galton ideas, as Walter Weldon, Arthur Dukinfield Darbishire and Karl Pearson, and Mendelians, who support Bateson (and Mendel) ideas, such as Charles Davenport, and Wilhelm Johannsen. Later, biometricians could not reproduce Galton conclusions in different experiments, and Mendel's ideas prevailed. By the 1930s, models built on statistical reasoning had helped to resolve these differences and to produce the neo-Darwinian modern evolutionary synthesis.

Solving these differences also allowed to define the concept of population genetics and brought together genetics and evolution. The three leading figures in the establishment of population genetics and this synthesis all relied on statistics and developed its use in biology.

- Ronald Fisher developed several basic statistical methods in support of his work studying the crop experiments at Rothamsted Research, including in his books Statistical Methods for Research Workers (1925) end The Genetical Theory of Natural Selection (1930). He gave many contributions to genetics and statistics. Some of them include the ANOVA, p-value concepts, Fisher's exact test and Fisher's equation for population dynamics. He is credited for the sentence "Natural selection is a mechanism for generating an exceedingly high degree of improbability".

- Sewall G. Wright developed F-statistics and methods of computing them and defined inbreeding coefficient.
- J. B. S. Haldane's book, *The Causes of Evolution*, reestablished natural selection as the premier mechanism of evolution by explaining it in terms of the mathematical consequences of Mendelian genetics. Also developed the theory of primordial soup.

These and other biostatisticians, mathematical biologists, and statistically inclined geneticists helped bring together evolutionary biology and genetics into a consistent, coherent whole that could begin to be quantitatively modeled.

In parallel to this overall development, the pioneering work of D'Arcy Thompson in *On Growth and Form* also helped to add quantitative discipline to biological study.

Despite the fundamental importance and frequent necessity of statistical reasoning, there may nonetheless have been a tendency among biologists to distrust or deprecate results which are not qualitatively apparent. One anecdote describes Thomas Hunt Morgan banning the Friden calculator from his department at Caltech, saying "Well, I am like a guy who is prospecting for gold along the banks of the Sacramento River in 1849. With a little intelligence, I can reach down and pick up big nuggets of gold. And as long as I can do that, I'm not going to let any people in my department waste scarce resources in placer mining."

BIOSTATISTICS AND MEDICINE

Statistical concepts also are present in clinical trials and epidemiological studies. These fields have their own history of biostatistics developments. In the 18th century, statistical methods were used to decide whether the application of certain treatments were effective, such as the insertion of smallpox pustules under an individual's skin in the hope of creating a mild case of the disease that would induce later immunity. Since this actually put patients at risk of contracting a potentially fatal form of the disease, this treatment became the subject of much controversy. John Arbuthnot in 1722 studied the chances of people dying by naturally-occurring smallpox as compared to inoculation-induced smallpox.

On the basis of the statistical studies, it was concluded that inoculation was preferred. Later, Daniel Bernoulli and Jean d'Alembert developed more robust statistical methods for the same problem, both based on the proportion of people who died. James Lind were the first to propose groups of test of hypotheses. He applied this method to solve an outbreak of scurvy. For this, he is considered the "father" of clinical trial. In 1835, Pierre-Charles-Alexandre Louis proposed the "numerical method" to argue that the practice of bloodletting was actually doing more harm than good for the patients.

In 1840, Louis Denis Jules Gavarret publish the Principes Généraux de Statistique Médicale in which he pointed out that Louis's averages could vary between what he called "limits of oscillation" (or confidence interval) if multiple samples were taken from the same population. Karl Pearson also expanded his methods to medicine, despite his main goal was to explicit the statistical implications of Darwin's theory of natural selection.

RESEARCH PLANNING

Any research in life sciences is proposed to answer a scientific question we might have. To answer this question with a high certainty, we need accurate results. The correct definition of the main hypothesis and the research plan will reduce errors while taking a decision in understanding a phenomenon. The research plan might include the research question, the hypothesis to be tested, the experimental design, data collection methods, data analysis perspectives and costs evolved. It is essential to carry the study based on the three basic principles of experimental statistics: randomization, replication, and local control.

RESEARCH QUESTION

The research question will define the objective of a study. The research will be headed by the question, so it needs to be concise, at the same time it is focused on interesting and novel topics that may improve science and knowledge and that field. To define the way to ask the scientific question, an exhaustive literature review might be necessary. So, the research can be useful to add value to the scientific community.

HYPOTHESIS DEFINITION

Once the aim of the study is defined, the possible answers to the research question can be proposed, transforming this question into a hypothesis. The main propose is called null hypothesis (H_0) and is usually based on a permanent knowledge about the topic or an obvious occurrence of the phenomena, sustained by a deep literature review. We can say it is the standard expected answer for the data under the situation in test. In general, H_O assumes no association between treatments. On the other hand, the alternative hypothesis is the denial of H_O. It assumes some degree of association between the treatment and the outcome. Although, the hypothesis is sustained by question research and its expected and unexpected answers. As an example, consider groups of similar animals (mice, for example) under two different diet systems. The research question would be: what is the best diet? In this case, H_0 would be that there is no difference between the two diets in mice metabolism (H_0: $\mu_1 = \mu_2$) and the alternative hypothesis would be that the diets have different effects over animals metabolism (H_1: $\mu_1 \neq \mu_2$). The hypothesis is defined by the researcher, according to his/her interests in answering the main question. Besides that, the alternative hypothesis can be more than one hypothesis. It can assume not only differences across observed parameters, but their degree of differences (*i.e.* higher or shorter).

SAMPLING

Usually, a study aims to understand an effect of a phenomenon over a population. In biology, a population is defined as all the individuals of a given species, in a specific area at a given time. In biostatistics, this concept is extended to a variety of collections possible of study. Although, in biostatistics, a population is not only the individuals, but the total of one specific component of their organisms, as the whole genome, or all the sperm cells, for animals, or the total leaf area, for a plant, for example.

It is not possible to take the measures from all the elements of a population. Because of that, the sampling process is very important for statistical inference. Sampling is defined as to randomly get a representative part of the entire population, to make posterior inferences about the population. So, the sample might catch the most variability across a population. The sample size is determined by several things, since the scope of the research to the resources available. In clinical research, the trial type, as inferiority, equivalence, and superiority is a key in determining sample size.

EXPERIMENTAL DESIGN

Experimental designs sustain those basic principles of experimental statistics. There are three basic experimental designs to randomly allocate treatments in all plots of the experiment. They are completely randomized design, randomized block design, and factorial designs. Treatments can be arranged in many ways inside the experiment. In agriculture, the correct experimental design is the root of a good study and the arrangement of treatments within the study is essential because environment largely affects the plots (plants, livestock, microorganisms). These main arrangements can be found in the literature under the names of “lattices”, “incomplete blocks”, “split plot”, “augmented blocks”, and many others. All of the designs might include control plots, determined by the researcher, to provide an error estimation during inference.

In clinical studies, the samples are usually smaller than in other biological studies, and in most cases, the environment effect can be controlled or measured. It is common to use randomized controlled clinical trials, where results are usually compared with observational study designs such as case–control or cohort.

DATA COLLECTION

Data collection methods must be considered in research planning, because it highly influences the sample size and experimental design. Data collection varies according to type of data. For qualitative data, collection can be done with structured questionnaires or by observation, considering presence or intensity of disease, using score criterion to categorize levels of occurrence. For quantitative data, collection is done by measuring numerical information using instruments. In agriculture and biology studies, yield data and its components can be obtained by metric measures. However, pest and disease injuries in plats are obtained by observation, considering score scales for levels of damage. Especially, in genetic studies, modern methods for data collection in field and laboratory should be considered, as high-throughput platforms for phenotyping and genotyping. These tools allow bigger experiments, while turn possible evaluate many plots in lower time than a human-based only method for data collection. Finally, all data collected of interest must be stored in an organized data frame for further analysis.

STATISTICAL INFERENCE

Statistical inference is the process of using data analysis to deduce properties of an underlying probability distribution. Inferential statistical analysis infers properties of a population, for example by testing hypotheses and deriving estimates. It is assumed that the observed data set is sampled from a larger population. Inferential statistics can be contrasted with descriptive statistics. Descriptive statistics is solely concerned with properties of the observed data, and it does not rest on the assumption that the data come from a larger population. Statistical inference makes propositions about a population, using data drawn from the population with some form of sampling. Given a hypothesis about a population, for which we wish to draw inferences, statistical inference consists of (first) selecting a statistical model of the process that generates the data and (second) deducing propositions from the model.

Konishi and Kitagawa state, "The majority of the problems in statistical inference can be considered to be problems related to statistical modeling". Relatedly, Sir David Cox has said, "How [the] translation from subject-matter problem to statistical model is done is often the most critical part of an analysis".

The conclusion of a statistical inference is a statistical proposition. Some common forms of statistical proposition are the following:

- a point estimate, *i.e.* a particular value that best approximates some parameter of interest;
- an interval estimate, *e.g.* a confidence interval (or set estimate), *i.e.* an interval constructed using a dataset drawn from a population so that, under repeated sampling of such datasets, such intervals would contain the true parameter value with the probability at the stated confidence level;
- a credible interval, *i.e.* a set of values containing, for example, 95 per cent of posterior belief;
- rejection of a hypothesis;
- clustering or classification of data points into groups.

MODELS AND ASSUMPTIONS

Any statistical inference requires some assumptions. A statistical model is a set of assumptions concerning the generation of the observed data and similar data. Descriptions of statistical models usually emphasize the role of population quantities of interest, about which we wish to draw inference. Descriptive statistics are typically used as a preliminary step before more formal inferences are drawn.

Degree of Models/Assumptions

Statisticians distinguish between three levels of modeling assumptions;

- *Fully parametric:* The probability distributions describing the data-generation process are assumed to be fully described by a family of probability distributions involving only a finite number of unknown parameters. For example, one may assume that the distribution of population values is truly Normal, with unknown mean and variance, and that datasets are generated by 'simple' random sampling. The family of generalized linear models is a widely used and flexible class of parametric models.
- *Non-parametric:* The assumptions made about the process generating the data are much less than in parametric statistics and may be minimal. For example, every continuous probability distribution has a median, which may be estimated using the sample median or the Hodges–Lehmann–Sen estimator, which has good properties when the data arise from simple random sampling.
- *Semi-parametric:* This term typically implies assumptions 'in between' fully and non-parametric approaches. For example, one may assume that a population distribution has a finite mean. Furthermore, one may assume that the mean response level in the population depends in a truly linear manner on some covariate (a parametric assumption) but not make any parametric assumption describing the variance around that mean (*i.e.* about the presence or possible form of any heteroscedasticity). More generally, semi-parametric models can often be separated into 'structural' and 'random variation' components. One component is treated parametrically and the other non-parametrically. The well-known Cox model is a set of semi-parametric assumptions.

Importance of Valid Models/Assumptions

Whatever level of assumption is made, correctly calibrated inference in general requires these assumptions to be correct; *i.e.* that the data-generating mechanisms really have been correctly specified.

Incorrect assumptions of 'simple' random sampling can invalidate statistical inference. More complex semi- and fully parametric assumptions are also cause for concern. For example, incorrectly assuming the Cox model can in some cases lead to faulty conclusions. Incorrect assumptions of Normality in the population also invalidates some forms of regression-based inference. The use of any parametric model is viewed skeptically by most experts in sampling human populations: "most sampling statisticians, when they deal with confidence intervals at all, limit themselves to statements about [estimators] based on very large samples, where the central limit theorem ensures that these [estimators] will have distributions that are nearly normal." In particular, a normal distribution "would be a totally unrealistic and catastrophically unwise assumption to make if we were dealing with any kind of economic population." Here, the central limit theorem states that the distribution of the sample mean "for very large samples" is approximately normally distributed, if the distribution is not heavy tailed.

Approximate Distributions

Given the difficulty in specifying exact distributions of sample statistics, many methods have been developed for approximating these. With finite samples, approximation results measure how close a limiting distribution approaches the statistic's sample distribution: For example, with 10,000 independent samples the normal distribution approximates (to two digits of accuracy) the distribution of the sample mean for many population distributions, by the Berry–Esseen theorem. Yet for many practical purposes, the normal approximation provides a good approximation to the sample-mean's distribution when there are 10 (or more) independent samples, according to simulation studies and statisticians' experience. Following Kolmogorov's work in the 1950s, advanced statistics uses approximation theory and functional analysis to quantify the error of approximation. In this approach, the metric geometry of probability distributions is studied; this approach quantifies approximation

error with, for example, the Kullback–Leibler divergence, Bregman divergence, and the Hellinger distance. With indefinitely large samples, limiting results like the central limit theorem describe the sample statistic's limiting distribution, if one exists. Limiting results are not statements about finite samples, and indeed are irrelevant to finite samples. However, the asymptotic theory of limiting distributions is often invoked for work with finite samples. For example, limiting results are often invoked to justify the generalized method of moments and the use of generalized estimating equations, which are popular in econometrics and biostatistics. The magnitude of the difference between the limiting distribution and the true distribution (formally, the 'error' of the approximation) can be assessed using simulation. The heuristic application of limiting results to finite samples is common practice in many applications, especially with low-dimensional models with log-concave likelihoods (such as with one-parameter exponential families).

Randomization-based Models

For a given dataset that was produced by a randomization design, the randomization distribution of a statistic (under the null-hypothesis) is defined by evaluating the test statistic for all of the plans that could have been generated by the randomization design. In frequentist inference, randomization allows inferences to be based on the randomization distribution rather than a subjective model, and this is important especially in survey sampling and design of experiments. Statistical inference from randomized studies is also more straightforward than many other situations. In Bayesian inference, randomization is also of importance: in survey sampling, use of sampling without replacement ensures the exchangeability of the sample with the population; in randomized experiments, randomization warrants a missing at random assumption for covariate information.

Objective randomization allows properly inductive procedures. Many statisticians prefer randomization-based analysis of data that was generated by well-defined randomization procedures. (However, it is true that in fields of science with developed theoretical knowledge and experimental control, randomized experiments may increase the costs of experimentation without improving the quality of inferences.) Similarly, results from randomized experiments are recommended by leading statistical authorities as allowing inferences with greater reliability than do observational studies of the same phenomena. However, a good observational study may be better than a bad randomized experiment. The statistical analysis of a randomized experiment may be based on the randomization scheme stated in the experimental protocol and does not need a subjective model. However, at any time, some hypotheses cannot be tested using objective statistical models, which accurately describe randomized experiments or random samples. In some cases, such randomized studies are uneconomical or unethical.

Model-based Analysis of Randomized Experiments

It is standard practice to refer to a statistical model, often a linear model, when analyzing data from randomized experiments. However, the randomization scheme guides the choice of a statistical model. It is not possible to choose an appropriate model without knowing the randomization scheme. Seriously misleading results can be obtained analyzing data from randomized experiments while ignoring the experimental protocol; common mistakes include forgetting the blocking used in an experiment and confusing repeated measurements on the same experimental unit with independent replicates of the treatment applied to different experimental units.

PARADIGMS FOR INFERENCE

Different schools of statistical inference have become established. These schools—or "paradigms"—are not mutually exclusive, and methods that work well under one paradigm often have attractive interpretations under other paradigms.

Bandyopadhyay and Forster describe four paradigms:

- Classical statistics or error statistics,
- Bayesian statistics,
- Likelihood-based statistics, and
- The Akaikean-Information Criterion-based statistics". The classical (or frequentist) paradigm, the Bayesian paradigm, and the AIC-based paradigm are summarized below. The likelihood-based paradigm is essentially a sub-paradigm of the AIC-based paradigm.

Frequentist Inference

This paradigm calibrates the plausibility of propositions by considering (notional) repeated sampling of a population distribution to produce datasets similar to the one at hand. By considering the dataset's characteristics under repeated sampling, the frequentist properties of a statistical proposition can be quantified—although in practice this quantification may be challenging.

Examples of Frequentist Inference

- *p*-value
- Confidence interval

Frequentist Inference, Objectivity, and Decision Theory

One interpretation of frequentist inference (or classical inference) is that it is applicable only in terms of frequency probability; that is, in terms of repeated sampling from a population. However, the approach of Neyman develops these procedures in terms of pre-experiment probabilities. That is, before undertaking an experiment, one decides on a rule for coming to a conclusion such that the probability of being correct is controlled in a suitable way: such a probability need not have a frequentist or repeated sampling interpretation. In contrast, Bayesian inference works in terms of conditional probabilities (*i.e.* probabilities conditional on the observed data), compared to the marginal (but conditioned on unknown parameters) probabilities used in the frequentist approach. The frequentist procedures of significance testing and confidence intervals can be constructed without regard to utility functions. However, some elements of frequentist statistics, such as statistical decision theory, do incorporate utility functions. In particular, frequentist developments of optimal inference (such as minimum-variance unbiased estimators, or uniformly most powerful testing) make use of loss functions, which play the role of (negative) utility functions. Loss functions need not be explicitly stated for statistical theorists to prove that a statistical procedure has an optimality property. However, loss-functions are often useful for stating optimality properties: for example, median-unbiased estimators are optimal under absolute value loss functions, in that they minimize expected loss, and least squares estimators are optimal under squared error loss functions, in that they minimize expected loss.

While statisticians using frequentist inference must choose for themselves the parameters of interest, and the estimators/test statistic to be used, the absence of obviously explicit utilities and prior distributions has helped frequentist procedures to become widely viewed as 'objective'.

Bayesian Inference

The Bayesian calculus describes degrees of belief using the 'language' of probability; beliefs are positive, integrate to one, and obey probability axioms. Bayesian inference uses the available posterior beliefs as the basis for making statistical propositions. There are several different justifications for using the Bayesian approach.

Examples of Bayesian Inference

- Credible interval for interval estimation
- Bayes factors for model comparison

Bayesian Inference, Subjectivity and Decision Theory

Many informal Bayesian inferences are based on "intuitively reasonable" summaries of the posterior. For example, the posterior mean, median and mode, highest posterior density intervals, and Bayes Factors can all be motivated in this way. While a user's utility function need not be stated for this sort of inference, these summaries do all depend (to some extent) on stated prior beliefs, and are generally viewed as subjective conclusions. (Methods of prior construction which do not require external input have been proposed but not yet fully developed.)

Formally, Bayesian inference is calibrated with reference to an explicitly stated utility, or loss function; the 'Bayes rule' is the one which maximizes expected utility, averaged over the posterior uncertainty. Formal Bayesian inference therefore automatically provides optimal decisions in a decision theoretic sense. Given assumptions, data and utility, Bayesian inference can be made for essentially any problem, although not every statistical inference need have a Bayesian interpretation. Analyses which are not formally Bayesian can be (logically) incoherent; a feature of Bayesian procedures which use proper priors (*i.e.* those integrable to one) is that they are guaranteed to be coherent. Some advocates of Bayesian inference assert that inference *must* take place in this decision-theoretic framework, and that Bayesian inference should not conclude with the evaluation and summarization of posterior beliefs.

AIC-based Inference

The *Akaike information criterion* (AIC) is an estimator of the relative quality of statistical models for a given set of data. Given a collection of models for the data, AIC estimates the quality of each model, relative to each of the other models. Thus, AIC provides a means for model selection.

AIC is founded on information theory: it offers an estimate of the relative information lost when a given model is used to represent the process that generated the data. (In doing so, it deals with the trade-off between the goodness of fit of the model and the simplicity of the model.)

Other Paradigms for Inference

Minimum Description Length

The minimum description length (MDL) principle has been developed from ideas in information theory and the theory of Kolmogorov complexity. The (MDL) principle selects statistical models that maximally compress the data; inference proceeds without assuming counterfactual or non-falsifiable "data-generating mechanisms" or probability models for the data, as might be done in frequentist or Bayesian approaches. However, if a "data generating mechanism" does exist in reality, then according to Shannon's source coding theorem it provides the MDL description of the data, on average and asymptotically. In minimizing description length (or descriptive complexity), MDL estimation is similar to maximum likelihood estimation and maximum a posteriori estimation (using maximum-entropy Bayesian priors). However, MDL avoids assuming that the underlying probability model is known; the MDL principle can also be applied without assumptions that *e.g.* the data arose from independent sampling.

The MDL principle has been applied in communication-coding theory in information theory, in linear regression, and in data mining. The evaluation of MDL-based inferential procedures often uses techniques or criteria from computational complexity theory.

Fiducial Inference

Fiducial inference was an approach to statistical inference based on fiducial probability, also known as a "fiducial distribution". In subsequent work, this approach has been called ill-defined, extremely limited in applicability, and even fallacious. However this argument is the same as that which shows that a so-called confidence distribution is not a valid probability distribution and, since this has not invalidated the application of confidence intervals, it does not necessarily invalidate conclusions drawn from fiducial arguments. An attempt was made to reinterpret the early work of Fisher's fiducial argument as a special case of an inference theory using Upper and lower probabilities.

Structural Inference

Developing ideas of Fisher and of Pitman from 1938 to 1939, George A. Barnard developed "structural inference" or "pivotal inference", an approach using invariant probabilities on group families. Barnard reformulated the arguments behind fiducial inference on a restricted class of models on which "fiducial" procedures would be well-defined and useful.

Inference Topics

The topics below are usually included in the area of statistical inference.

- Statistical assumptions
- Statistical decision theory
- Estimation theory
- Statistical hypothesis testing
- Revising opinions in statistics
- Design of experiments, the analysis of variance, and regression
- Survey sampling
- Summarizing statistical data

STATISTICAL CONSIDERATIONS

Power and statistical error

When testing a hypothesis, there are two types of statistic errors possible: Type I error and Type II error. The type I error or false positive is the incorrect rejection of a true null hypothesis and the type II error or false negative is the failure to reject a false null hypothesis. The significance level denoted by α is the type I error rate and should be chosen before performing the test. The type II error rate is denoted by β and statistical power of the test is $1 - \beta$.

P-VALUE

The p-value is the probability of obtaining results as extreme as or more extreme than those observed, assuming the null hypothesis (H_0) is true. It is also called the calculated probability. It is common to confuse the p-value with the significance level (α), but, the α is a predefined threshold for calling significant results. If p is less than α, the null hypothesis (H_0) is rejected.

MULTIPLE TESTING

In multiple tests of the same hypothesis, the probability of the occurrence of falses positives (familywise error rate) increase and some strategy are used to control this occurrence. This is commonly achieved by using a more stringent threshold to reject null hypotheses. The Bonferroni correction defines an acceptable global significance level, denoted by α^* and each test is individually compared with a value of $\alpha = \alpha^*/m$. This ensures that the familywise error rate in all m tests, is less than or equal to α^*. When m is large, the Bonferroni correction may be overly conservative. An alternative to the Bonferroni correction is to control the false discovery rate (FDR). The FDR controls the expected proportion of the rejected null hypotheses (the so-called discoveries) that are false (incorrect rejections). This procedure ensures that, for independent tests, the false discovery rate is at most q^*. Thus, the FDR is less conservative than the Bonferroni correction and have more power, at the cost of more false positives.

MIS-SPECIFICATION AND ROBUSTNESS CHECKS

The main hypothesis being tested (*e.g.*, no association between treatments and outcomes) is often accompanied by other technical assumptions (*e.g.*, about the form of the probability distribution of the outcomes) that are also part of the null hypothesis. When the technical assumptions are violated in practice, then the null may be frequently rejected even if the main hypothesis is true. Such rejections are said to be due to model mis-specification. Verifying whether the outcome of a statistical test does not change when the technical assumptions are slightly altered (so-called robustness checks) is the main way of combating mis-specification.

MODEL SELECTION CRITERIA

Model criteria selection will select or model that more approximate true model. The Akaike's Information Criterion (AIC) and The Bayesian Information Criterion (BIC) are examples of asymptotically efficient criteria.

DEVELOPMENTS AND BIG DATA

Recent developments have made a large impact on biostatistics. Two important changes have been the ability to collect data on a high-throughput scale, and the ability to perform much more complex analysis using computational techniques. This comes from the development in areas as sequencing technologies, Bioinformatics and Machine learning (Machine learning in bioinformatics).

USE IN HIGH-THROUGHPUT DATA

New biomedical technologies like microarrays, next-generation sequencers (for genomics) and mass spectrometry (for proteomics) generate enormous amounts of data, allowing many tests to be performed simultaneously. Careful analysis with biostatistical methods is required to separate the signal from the noise. For example, a microarray could be used to measure many thousands of genes simultaneously, determining which of them have different expression in diseased cells compared to normal cells. However, only a fraction of genes will be differentially expressed. Multicollinearity often occurs in high-throughput biostatistical settings. Due to high intercorrelation between the predictors (such as gene expression levels), the information of one predictor might be contained in another one. It could be that only 5 per cent of the predictors are responsible for 90 per cent of the variability of the response. In such a case, one could apply the biostatistical technique of dimension reduction (for example via principal component analysis). Classical statistical techniques like linear or logistic regression and linear discriminant analysis do not work well for high dimensional data (*i.e.* when the number of observations n is smaller than the number of features or predictors p: $n < p$). As a matter of fact, one

can get quite high R-values despite very low predictive power of the statistical model. These classical statistical techniques (esp. least squares linear regression) were developed for low dimensional data (*i.e.* where the number of observations n is much larger than the number of predictors p: n >> p). In cases of high dimensionality, one should always consider an independent validation test set and the corresponding residual sum of squares (RSS) and R of the validation test set, not those of the training set. Often, it is useful to pool information from multiple predictors together. For example, Gene Set Enrichment Analysis (GSEA) considers the perturbation of whole (functionally related) gene sets rather than of single genes. These gene sets might be known biochemical pathways or otherwise functionally related genes. The advantage of this approach is that it is more robust: It is more likely that a single gene is found to be falsely perturbed than it is that a whole pathway is falsely perturbed. Furthermore, one can integrate the accumulated knowledge about biochemical pathways (like the JAK-STAT signaling pathway) using this approach.

BIOINFORMATICS ADVANCES IN DATABASES, DATA MINING, AND BIOLOGICAL INTERPRETATION

The development of biological databases enables storage and management of biological data with the possibility of ensuring access for users around the world. They are useful for researchers depositing data, retrieve information and files (raw or processed) originated from other experiments or indexing scientific articles, as PubMed. Another possibility is search for the desired term (a gene, a protein, a disease, an organism, and so on) and check all results related to this search. There are databases dedicated to SNPs (dbSNP), the knowledge on genes characterization and their pathways (KEGG) and the description of gene function classifying it by cellular component, molecular function and biological process (Gene Ontology). In addition to databases that contain specific molecular information, there are others that are ample in the sense that they store information about an organism or group of organisms. As an example of a database directed towards just one organism, but that contains lots of data about it, is the *Arabidopsis thaliana* genetic and molecular database – TAIR. Phytozome, in turn, stores the assemblies and annotation files of dozen of plant genomes, also containing visualization and analysis tools. Moreover, there is an interconnection between some databases in the information exchange/sharing and a major initiative was the International Nucleotide Sequence Database Collaboration (INSDC) which relates data from DDBJ, EMBL-EBI, and NCBI. Nowadays, increase in size and complexity of molecular datasets leads to use of powerful statistical methods provided by computer science algorithms which are developed by machine learning area. Therefore, data mining and machine learning allow detection of patterns in data with a complex structure, as biological ones, by using methods of supervised and unsupervised learning, regression, detection of clusters and association rule mining, among others. To indicate some of them, self-organizing maps and *k*-means are examples of cluster algorithms; neural networks implementation and support vector machines models are examples of common machine learning algorithms.

Collaborative work among molecular biologists, bioinformaticians, statisticians and computer scientist is important to perform an experiment correctly, going from planning, passing through data generation and analysis, and ending with biological interpretation of the results.

USE OF COMPUTATIONALLY INTENSIVE METHODS

On the other hand, the advent of modern computer technology and relatively cheap computing resources have enabled computer-intensive biostatistical methods like bootstrapping and re-sampling methods. In recent times, random forests have gained popularity as a method for performing statistical classification. Random forest techniques generate a panel of decision trees. Decision trees have the advantage that you can draw them and interpret them (even with a basic understanding of mathematics and statistics). Random Forests have thus been used for clinical decision support systems.

APPLICATIONS

Public health

Public health, including epidemiology, health services research, nutrition, environmental health and health care policy and management. In these medicine contents, it's important to consider the design and analysis of the clinical trials. As one example, there is the assessment of severity state of a patient with a prognosis of an outcome of a disease. With new technologies and genetics knowledge, biostatistics are now also used for Systems medicine, which consists in a more personalized medicine. For this, is made a integration of data from different sources, including conventional patient data, clinico-pathological parameters, molecular and genetic data as well as data generated by additional new-omics technologies.

QUANTITATIVE GENETICS

The study of Population genetics and Statistical genetics in order to link variation in genotype with a variation in phenotype. In other words, it is desirable to discover the genetic basis of a measurable trait, a quantitative trait, that is under polygenic control. A genome region that is responsible for a continuous trait is called Quantitative trait locus (QTL). The study of QTLs become feasible by using molecular markers and measuring traits in populations, but their mapping needs the obtaining of a population from an experimental crossing, like an F2 or Recombinant inbred strains/lines (RILs). To scan for QTLs regions in a genome, a gene map based on linkage have to be built. Some of the best-known QTL mapping algorithms are Interval Mapping, Composite Interval Mapping, and Multiple Interval Mapping. However, QTL mapping resolution is impaired by the amount of recombination assayed, a problem for species in which it is difficult to obtain large offspring. Furthermore, allele diversity is restricted to individuals originated from contrasting parents, which limit studies of allele diversity when we have a panel of individuals representing a natural population. For this reason, the Genome-wide association study was proposed in order to identify QTLs based on linkage disequilibrium, that is the non-random association between traits and molecular markers. It was leveraged by the development of high-throughput SNP genotyping.

In animal and plant breeding, the use of markers in selection aiming for breeding, mainly the molecular ones, collaborated to the development of marker-assisted selection. While QTL mapping is limited due resolution, GWAS does not have enough power when rare variants of small effect that are also influenced by environment. So, the concept of Genomic Selection (GS) arises in order to use all molecular markers in the selection and allow the prediction of the performance of candidates in this selection. The proposal is to genotype and phenotype a training population, develop a model that can obtain the genomic estimated breeding values (GEBVs) of individuals belonging to a genotyped and but not phenotyped population, called testing population. This kind of study could also include a validation population, thinking in the concept of cross-validation, in which the real phenotype results measured in this population are compared with the phenotype results based on the prediction, what used to check the accuracy of the model.

As a summary, some points about the application of quantitative genetics are:

- This has been used in agriculture to improve crops (Plant breeding) and livestock (Animal breeding).
- In biomedical research, this work can assist in finding candidates gene alleles that can cause or influence predisposition to diseases in human genetics

EXPRESSION DATA

Studies for differential expression of genes from RNA-Seq data, as for RT-qPCR and microarrays, demands comparison of conditions. The goal is to identify genes which have a significant change in abundance between

different conditions. Then, experiments are designed appropriately, with replicates for each condition/treatment, randomization and blocking, when necessary. In RNA-Seq, the quantification of expression uses the information of mapped reads that are summarized in some genetic unit, as exons that are part of a gene sequence. As microarray results can be approximated by a normal distribution, RNA-Seq counts data are better explained by other distributions. The first used distribution was the Poisson one, but it underestimate the sample error, leading to false positives. Currently, biological variation is considered by methods that estimate a dispersion parameter of a negative binomial distribution. Generalized linear models are used to perform the tests for statistical significance and as the number of genes is high, multiple tests correction have to be considered. Some examples of other analysis on genomics data comes from microarray or proteomics experiments. Often concerning diseases or disease stages.

OTHER STUDIES

- Ecology, ecological forecasting
- Biological sequence analysis
- Systems biology for gene network inference or pathways analysis.
- Population dynamics, especially in regards to fisheries science.
- Phylogenetics and evolution

TOOLS

CycDesigN

CycDesigN is a computer package for generating optimal or near-optimal experimental designs. Four general classifications of designs can be constructed: these are resolvable, non-resolvable, partially replicated and crossover designs. The first three classes can be set out in blocks or in rows and columns. Resolvable designs can be latinized, t-latinized or partially latinized, while non-resolvable designs can also be unequally replicated. In addition, spatial resolvable and non-resolvable designs, using the linear variance or exponential variance models, can be constructed. Partially replicated designs have test treatments set out in a number of locations where each test treatment appears zero, once or twice in each location; standard treatments can be included with multiple replication at each location. Crossover designs are used when sequences of treatments are applied to several subjects over a number of time periods. The direct effect of the treatment applied in the current period and the carry-over effects of the treatments applied in one or more previous periods can be modelled in various ways, and numerous options are provided in CycDesigN. Crossover designs are also known as change-over or carry-over designs.

CycAnalysis is a separate module in CycDesigN that allows output from a CycDesigN session to be tailored in a form ready for analysis. For example, a spreadsheet of the design blocking and treatment structures is automatically generated. CycDesigN also lets you generate Genstat and SAS code for the analysis of most designs.

CycDesigN provides the most comprehensive design generation package yet available for experimenters. In particular, resolvable and partially replicated designs are used by experimenters involved in field variety trials. For smaller experiments, say in the glasshouse or laboratory, non-resolvable row-column designs are often employed and can typically include spatial enhancement. Crossover designs are frequently used in areas such as clinical trials, taste and psychological testing and psycho-physical experiments. The authors are leading researchers and, as a result, the algorithms incorporate the most recent developments in the construction of experimental designs.

New features:

- New and improved licensing system.
- A substantial revision of the algorithms used to generate resolvable, non-resolvable and partially replicated block designs. The algorithm now seamlessly constructs designs taking advantage of the speed of cyclic and alpha design construction, and automatically switches to a more general algorithm to search for more efficient designs.
- The construction of non-resolvable designs with unequal treatment replication.
- Spatial enhancement of both equal and unequal replicate non-resolvable designs.
- The construction of partially replicated row-column designs.
- The incorporation of standard treatments into the generation of efficient partially replicated block and row-column designs.
- Genstat or SAS code for block and row-column partially replicated designs, with or without standard treatments.
- An updated and improved Help facility that is now more self-contained.

SAS

SAS (previously "Statistical Analysis System") is a software suite developed by SAS Institute for advanced analytics, multivariate analyses, business intelligence, data management, and predictive analytics.

SAS was developed at North Carolina State University from 1966 until 1976, when SAS Institute was incorporated. SAS was further developed in the 1980s and 1990s with the addition of new statistical procedures, additional components and the introduction of JMP. A point-and-click interface was added in version 9 in 2004. A social media analytics product was added in 2010.

Technical Overview and Terminology

SAS is a software suite that can mine, alter, manage and retrieve data from a variety of sources and perform statistical analysis on it. SAS provides a graphical point-and-click user interface for non-technical users and more advanced options through the SAS language.

SAS programmes have DATA steps, which retrieve and manipulate data, and PROC steps, which analyze the data. Each step consists of a series of statements.

The DATA step has executable statements that result in the software taking an action, and declarative statements that provide instructions to read a data set or alter the data's appearance. The DATA step has two phases: compilation and execution. In the compilation phase, declarative statements are processed and syntax errors are identified. Afterwards, the execution phase processes each executable statement sequentially. Data sets are organized into tables with rows called "observations" and columns called "variables". Additionally, each piece of data has a descriptor and a value.

The PROC step consists of PROC statements that call upon named procedures. Procedures perform analysis and reporting on data sets to produce statistics, analyses, and graphics. There are more than 300 procedures and each one contains a substantial body of programming and statistical work. PROC statements can also display results, sort data or perform other operations.

SAS macros are pieces of code or variables that are coded once and referenced to perform repetitive tasks.

SAS data can be published in HTML, PDF, Excel and other formats using the Output Delivery System, which was first introduced in 2007. The SAS Enterprise Guide is SAS's point-and-click interface. It generates code to manipulate data or perform analysis automatically and does not require SAS programming experience to use.

The SAS software suite has more than 200 components Some of the SAS components include:

- Base SAS – Basic procedures and data management
- SAS/STAT – Statistical analysis
- SAS/GRAPH – Graphics and presentation
- SAS/OR – Operations research
- SAS/ETS – Econometrics and Time Series Analysis
- SAS/IML – Interactive matrix language
- SAS/AF – Applications facility
- SAS/QC – Quality control
- SAS/INSIGHT – Data mining
- SAS/PH – Clinical trial analysis
- Enterprise Miner – data mining
- Enterprise Guide - GUI based code editor and project manager
- SAS EBI - Suite of Business Intelligence Applications
- SAS Grid Manager - Manager of SAS grid computing environment

History

Origins

The development of SAS began in 1966 after North Carolina State University re-hired Anthony Barr to programme his analysis of variance and regression software so that it would run on IBM System/360 computers. The project was funded by the National Institute of Health and was originally intended to analyze agricultural data to improve crop yields. Barr was joined by student James Goodnight, who developed the software's statistical routines, and the two became project leaders. In 1968, Barr and Goodnight integrated new multiple regression and analysis of variance routines. In 1972, after issuing the first release of SAS, the project lost its funding. According to Goodnight, this was because NIH only wanted to fund projects with medical applications. Goodnight continued teaching at the university for a salary of $1 and access to mainframe computers for use with the project, until it was funded by the University Statisticians of the Southern Experiment Stations the following year. John Sall joined the project in 1973 and contributed to the software's econometrics, time series, and matrix algebra. Another early participant, Caroll G. Perkins, contributed to SAS' early programming. Jolayne W. Service and Jane T. Helwig created SAS' first documentation. The first versions of SAS were named after the year in which they were released. In 1971, SAS 71 was published as a limited release. It was used only on IBM mainframes and had the main elements of SAS programming, such as the DATA step and the most common procedures in the PROC step. The following year a full version was released as SAS 72, which introduced the MERGE statement and added features for handling missing data or combining data sets. In 1976, Barr, Goodnight, Sall, and Helwig removed the project from North Carolina State and incorporated it into SAS Institute, Inc.

Development

SAS was re-designed in SAS 76 with an open architecture that allowed for compilers and procedures. The INPUT and INFILE statements were improved so they could read most data formats used by IBM mainframes.

Generating reports was also added through the PUT and FILE statements. The ability to analyze general linear models was also added as was the FORMAT procedure, which allowed developers to customize the

appearance of data. In 1979, SAS 79 added support for the CMS operating system and introduced the DATASETS procedure. Three years later, SAS 82 introduced an early macro language and the APPEND procedure.

SAS version 4 had limited features, but made SAS more accessible. Version 5 introduced a complete macro language, array subscripts, and a full-screen interactive user interface called Display Manager. In 1985, SAS was rewritten in the C programming language. This allowed for the SAS' Multivendor Architecture that allows the software to run on UNIX, MS-DOS, and Windows. It was previously written in PL/I, Fortran, and assembly language. In the 1980s and 1990s, SAS released a number of components to complement Base SAS. SAS/GRAPH, which produces graphics, was released in 1980, as well as the SAS/ETS component, which supports econometric and time series analysis. A component intended for pharmaceutical users, SAS/PH-Clinical, was released in the 1990s. The Food and Drug Administration standardized on SAS/PH-Clinical for new drug applications in 2002. Vertical products like SAS Financial Management and SAS Human Capital Management (then called CFO Vision and HR Vision respectively) were also introduced. JMP was developed by SAS co-founder John Sall and a team of developers to take advantage of the graphical user interface introduced in the 1984 Apple Macintosh and shipped for the first time in 1989. Updated versions of JMP were released continuously after 2002 with the most recent release being from 2016.

SAS version 6 was used throughout the 1990s and was available on a wider range of operating systems, including Macintosh, OS/2, Silicon Graphics, and Primos. SAS introduced new features through dot-releases. From 6.06 to 6.09, a user interface based on the windows paradigm was introduced and support for SQL was added. Version 7 introduced the Output Delivery System (ODS) and an improved text editor. ODS was improved upon in successive releases. For example, more output options were added in version 8. The number of operating systems that were supported was reduced to UNIX, Windows and z/OS, and Linux was added. SAS version 8 and SAS Enterprise Miner were released in 1999.

Recent History

In 2002, the Text Miner software was introduced. Text Miner analyzes text data like emails for patterns in Business Intelligence applications. In 2004, SAS Version 9.0 was released, which was dubbed "Project Mercury" and was designed to make SAS accessible to a broader range of business users. Version 9.0 added custom user interfaces based on the user's role and established the point-and-click user interface of SAS Enterprise Guide as the software's primary graphical user interface (GUI). The Customer Relationship Management (CRM) features were improved in 2004 with SAS Interaction Management. In 2008 SAS announced Project Unity, designed to integrate data quality, data integration and master data management.

SAS sued World Programming, the developers of a competing implementation, World Programming System, alleging that they had infringed SAS's copyright in part by implementing the same functionality. This case was referred from the United Kingdom's High Court of Justice to the European Court of Justice on 11 August 2010. In May 2012, the European Court of Justice ruled in favour of World Programming, finding that "the functionality of a computer programme and the programming language cannot be protected by copyright." A free version was introduced for students in 2010. SAS Social Media Analytics, a tool for social media monitoring, engagement and sentiment analysis, was also released that year. SAS Rapid Predictive Modeler (RPM), which creates basic analytical models using Microsoft Excel, was introduced that same year. JMP 9 in 2010 added a new interface for using the R programming language from JMP and an add-in for Excel. The following year, a High Performance Computing appliance was made available in a partnership with Teradata and EMC Greenplum. In 2011, the company released Enterprise Miner 7.1. The company introduced 27 data management products from October 2013 to October 2014 and updates to 160 others. At the 2015 SAS Global Forum, it announced several new products that were specialized for different industries, as well as new training software.

Releases Date

Table. SAS had many releases since 1972. Since release 9.3, SAS/STAT has its own release numbering.

Release	Date	Comment
72	January 1972	
76	July 1976	
79.5	April 1981	
82.4	January 1983	
4.06	March 1984	
5.03	July 1986	
6.01	January 1985	PC-DOS
6.03	March 1988	
6.06	March 1990	
6.07	April 1991	
6.08	March 1993	
6.09	October 1993	
6.10	October 1994	
6.11	October 1995	
6.12	November 1996	
7.0	October 1998	
8.0	November 1999	
8.1	July 2000	
8.2	March 2001	
9.0	October 2002	
9.1	December 2003	
9.1.3	August 2004	
9.2	March 2008	STAT 9.2
9.2m2	April 2010	STAT 9.22
9.3	July 2011	STAT 9.3
9.3m2	August 2012	STAT 12.1
9.4	July 2013	STAT 12.3
9.4m1	December 2013	STAT 13.1
9.4m2	August 2014	STAT 13.2
9.4m3	July 2015	STAT 14.1
9.4m4	November 2016	STAT 14.2
9.4m5	September 2017	STAT 14.3

Software Products

As of 2011 SAS's largest set of products is its line for customer intelligence. Numerous SAS modules for web, social media and marketing analytics may be used to profile customers and prospects, predict their behaviors and manage and optimize communications. SAS also provides the SAS Fraud Framework. The framework's primary functionality is to monitor transactions across different applications, networks and partners and use analytics to identify anomalies that are indicative of fraud. SAS Enterprise GRC (Governance, Risk and Compliance) provides risk modeling, scenario analysis and other functions in order to manage and visualize risk, compliance and corporate policies. There is also a SAS Enterprise Risk Management product-set designed primarily for banks and financial services organizations.

SAS' products for monitoring and managing the operations of IT systems are collectively referred to as SAS IT Management Solutions. SAS collects data from various IT assets on performance and utilization, then creates reports and analyses. SAS' Performance Management products consolidate and provide graphical displays for

key performance indicators (KPIs) at the employee, department and organizational level. The SAS Supply Chain Intelligence product suite is offered for supply chain needs, such as forecasting product demand, managing distribution and inventory and optimizing pricing. There is also a "SAS for Sustainability Management" set of software to forecast environmental, social and economic effects and identify causal relationships between operations and an impact on the environment or ecosystem.

SAS has product sets for specific industries, such as government, retail, telecommunications and aerospace and for marketing optimization or high-performance computing.

Comparison to other Products

In a 2005 article for the *Journal of Marriage and Family* comparing statistical packages from SAS and its competitors Stata and SPSS, Alan C. Acock wrote that SAS programmes provide "extraordinary range of data analysis and data management tasks," but were difficult to use and learn. SPSS and Stata, meanwhile, were both easier to learn (with better documentation) but had less capable analytic abilities, though these could be expanded with paid (in SPSS) or free (in Stata) add-ons. Acock concluded that SAS was best for power users, while occasional users would benefit most from SPSS and Stata. A comparison by the University of California, Los Angeles, gave similar results. Competitors such as Revolution Analytics and Alpine Data Labs advertise their products as considerably cheaper than SAS'. In a 2011 comparison, Doug Henschen of *InformationWeek* found that start-up fees for the three are similar, though he admitted that the starting fees were not necessarily the best basis for comparison. SAS' business model is not weighted as heavily on initial fees for its programmes, instead focusing on revenue from annual subscription fees.

Adoption

According to IDC, SAS is the largest market-share holder in "advanced analytics" with 35.4 percent of the market as of 2013. It is the fifth largest market-share holder for business intelligence (BI) software with a 6.9 per cent share and the largest independent vendor. It competes in the BI market against conglomerates, such as SAP BusinessObjects, IBM Cognos, SPSS Modeler, Oracle Hyperion, and Microsoft BI. SAS has been named in the Gartner Leader's Quadrant for Data Integration Tool and for Business Intelligence and Analytical Platforms. A study published in 2011 in *BMC Health Services Research* found that SAS was used in 42.6 percent of data analyses in health service research, based on a sample of 1,139 articles drawn from three journals.

R (PROGRAMMING LANGUAGE)

R is a programming language and free software environment for statistical computing and graphics supported by the R Foundation for Statistical Computing. The R language is widely used among statisticians and data miners for developing statistical software and data analysis. Polls, data mining surveys and studies of scholarly literature databases, show substantial increases in popularity in recent years. As of August 2018, R ranks 18th in the TIOBE index, a measure of popularity of programming languages. A GNU package, source code for the R software environment is written primarily in C, Fortran and R itself and is freely available under the GNU General Public License. Pre-compiled binary versions are provided for various operating systems. Although R has a command line interface, there are several graphical user interfaces, such as RStudio, an Integrated development environment.

History

R is an implementation of the S programming language combined with lexical scoping semantics, inspired by Scheme. S was created by John Chambers in 1976, while at Bell Labs. There are some important differences,

but much of the code written for S runs unaltered. R was created by Ross Ihaka and Robert Gentleman at the University of Auckland, New Zealand, and currently developed by the *R Development Core Team* (of which Chambers is a member). R is named partly after the first names of the first two R authors and partly as a play on the name of S. The project was conceived in 1992, with an initial version released in 1995 and a stable beta version in 2000.

Statistical Features

R and its libraries implement a wide variety of statistical and graphical techniques, including linear and nonlinear modelling, classical statistical tests, time-series analysis, classification, clustering, and others. R is easily extensible through functions and extensions, and the R community is noted for its active contributions in terms of packages. Many of R's standard functions are written in R itself, which makes it easy for users to follow the algorithmic choices made. For computationally intensive tasks, C, C++, and Fortran code can be linked and called at run time. Advanced users can write C, C++, Java,.NET or Python code to manipulate R objects directly. R is highly extensible through the use of user-submitted packages for specific functions or specific areas of study. Due to its S heritage, R has stronger object-oriented programming facilities than most statistical computing languages. Extending R is also eased by its lexical scoping rules. Another strength of R is static graphics, which can produce publication-quality graphs, including mathematical symbols. Dynamic and interactive graphics are available through additional packages.

R has Rd, its own LaTeX-like documentation format, which is used to supply comprehensive documentation, both online in a number of formats and in hard copy.

Programming Features

R is an interpreted language; users typically access it through a command-line interpreter. If a user types 2+2 at the R command prompt and presses enter, the computer replies with 4, as shown below:

```
>2+2
[1] 4
```

This calculation is interpreted as the sum of two single-element vectors, resulting in a single-element vector. The prefix [1] indicates that the list of elements following it on the same line starts with the *first* element of the vector (a feature that is useful when the output extends over multiple lines).

Like other similar languages such as APL and MATLAB, R supports matrix arithmetic. R's data structures include vectors, matrices, arrays, data frames (similar to tables in a relational database) and lists. R's extensible object system includes objects for (among others): regression models, time-series and geo-spatial coordinates. The scalar data type was never a data structure of R. Instead, a scalar is represented as a vector with length one.

R supports procedural programming with functions and, for some functions, object-oriented programming with generic functions. A generic function acts differently depending on the classes of arguments passed to it. In other words, the generic function dispatches the function (method) specific to that class of object. For example, R has a generic print function that can print almost every class of object in R with a simple print(objectname) syntax. Although used mainly by statisticians and other practitioners requiring an environment for statistical computation and software development, R can also operate as a general matrix calculation toolbox – with performance benchmarks comparable to GNU Octave or MATLAB. Arrays are stored in column-major order.

Packages

The capabilities of R are extended through user-created *packages*, which allow specialised statistical techniques, graphical devices, import/export capabilities, reporting tools (knitr, Sweave), etc. These packages

are developed primarily in R, and sometimes in Java, C, C++, and Fortran. The R packaging system is also used by researchers to create compendia to organise research data, code and report files in a systematic way for sharing and public archiving. A core set of packages is included with the installation of R, with more than 15,000 additional packages (as of September 2018) available at the Comprehensive R Archive Network (CRAN), Bioconductor, Omegahat, GitHub, and other repositories.

The "Task Views" page (subject list) on the CRAN web site lists a wide range of tasks (in fields such as Finance, Genetics, High Performance Computing, Machine Learning, Medical Imaging, Social Sciences and Spatial Statistics) to which R has been applied and for which packages are available. R has also been identified by the FDA as suitable for interpreting data from clinical research. Other R package resources include Crantastic, a community site for rating and reviewing all CRAN packages, and R-Forge, a central platform for the collaborative development of R packages, R-related software, and projects. R-Forge also hosts many unpublished beta packages, and development versions of CRAN packages.

The Bioconductor project provides R packages for the analysis of genomic data, such as Affymetrix and cDNA microarray object-oriented data-handling and analysis tools, and has started to provide tools for analysis of data from next-generation high-throughput sequencing methods.

Milestones

Table. A list of changes in R releases is maintained in various "news" files at CRAN. Some highlights are listed below for several major releases.

Release	Date	Description
0.16		This is the last alpha version developed primarily by Ihaka and Gentleman. Much of the basic functionality from the "White Book" was implemented. The mailing lists commenced on April 1, 1997.
0.49	1997-04-23	This is the oldest source release which is currently available on CRAN. CRAN is started on this date, with 3 mirrors that initially hosted 12 packages. Alpha versions of R for Microsoft Windows and the classic Mac OS are made available shortly after this version.
0.60	1997-12-05	R becomes an official part of the GNU Project. The code is hosted and maintained on CVS.
0.65.1	1999-10-07	First versions of update.packages and install.packages functions for downloading and installing packages from CRAN.
1.0	2000-02-29	Considered by its developers stable enough for production use.
1.4	2001-12-19	S4 methods are introduced and the first version for Mac OS X is made available soon after.
2.0	2004-10-04	Introduced lazy loading, which enables fast loading of data with minimal expense of system memory.
2.1	2005-04-18	Support for UTF-8 encoding, and the beginnings of internationalization and localization for different languages.
2.11	2010-04-22	Support for Windows 64 bit systems.
2.13	2011-04-14	Adding a new compiler function that allows speeding up functions by converting them to byte-code.
2.14	2011-10-31	Added mandatory namespaces for packages. Added a new parallel package.
2.15	2012-03-30	New load balancing functions. Improved serialisation speed for long vectors.
3.0	2013-04-03	Support for numeric index values 2 and larger on 64 bit systems.
3.4	2017-04-21	Just-in-time compilation (JIT) of functions and loops to byte-code enabled by default.
3.5	2018-04-23	Packages byte-compiled on installation by default. Compact internal representation of integer sequences. Added a new serialisation format to support compact internal representations.

Interfaces

The most commonly used graphical integrated development environment for R is RStudio. A similar development interface is R Tools for Visual Studio. Interfaces with more of a point-and-click approach include Rattle GUI, R Commander, and RKWard. Some of the more common editors with varying levels of support for

R include Eclipse, Emacs (Emacs Speaks Statistics), Kate, LyX, Notepad++, Visual Studio Code, WinEdt, and Tinn-R. R functionality is accessible from several scripting languages such as Python, Perl, Ruby, F#, and Julia. Interfaces to other, high-level programming languages, like Java and.NET C# are available as well.

Implementations

The main R implementation is written in R, C, and Fortran, and there are several other implementations aimed at improving speed or increasing extensibility. A closely related implementation is pqR (pretty quick R) by Radford M. Neal with improved memory management and support for automatic multithreading. Renjin and FastR are Java implementations of R for use in a Java Virtual Machine. CXXR, rho, and Riposte are implementations of R in C++. Renjin, Riposte, and pqR attempt to improve performance by using multiple processor cores and some form of deferred evaluation. Most of these alternative implementations are experimental and incomplete, with relatively few users, compared to the main implementation maintained by the R Development Core Team.

TIBCO built a runtime engine called TERR, which is part of Spotfire.

Microsoft R Open is a fully compatible R distribution with modifications for multi-threaded computations.

R Communities

R has vibrant and active local communities worldwide for users to network, share ideas and learn.

There are regular R-user meetups and a more focused R-Ladies groups which promotes gender diversity.

User! Conferences

The official annual gathering of R users is called "useR!". The first such event was useR! 2004 in May 2004, Vienna, Austria. After skipping 2005, the useR! conference has been held annually, usually alternating between locations in Europe and North America.

Subsequent conferences have included:

- useR! 2006, Vienna, Austria
- useR! 2007, Ames, Iowa, USA
- useR! 2008, Dortmund, Germany
- useR! 2009, Rennes, France
- useR! 2010, Gaithersburg, Maryland, USA
- useR! 2011, Coventry, United Kingdom
- useR! 2012, Nashville, Tennessee, USA
- useR! 2013, Albacete, Spain
- useR! 2014, Los Angeles, California, USA
- useR! 2015, Aalborg, Denmark
- useR! 2016, Stanford, California, USA
- useR! 2017, Brussels, Belgium
- useR! 2018, Brisbane, Australia

Future conferences planned are as follows:

- useR! 2019, Toulouse, France
- useR! 2020, Boston, Massachusetts, USA

R Journal

The R Journal is the open access, refereed journal of the R project for statistical computing. It features short to medium length articles on the use, and development of R, including packages, programming tips, CRAN news, and foundation news.

Comparison with SAS, SPSS, and Stata

R is comparable to popular commercial statistical packages, such as SAS, SPSS, and Stata, but R is available to users at no charge under a free software license. In January 2009, the *New York Times* ran an article charting the growth of R, the reasons for its popularity among data scientists and the threat it poses to commercial statistical packages such as SAS.

Commercial Support for R

Although R is an open-source project supported by the community developing it, some companies strive to provide commercial support and/or extensions for their customers. This section gives some examples of such companies. In 2007 Richard Schultz, Martin Schultz, Steve Weston and Kirk Mettler founded Revolution Analytics to provide commercial support for Revolution R, their distribution of R, which also includes components developed by the company. Major additional components include: ParallelR, the R Productivity Environment IDE, RevoScaleR (for big data analysis), RevoDeployR, web services framework, and the ability for reading and writing data in the SAS file format. Revolution Analytics also offer a distribution of R designed to comply with established IQ/OQ/PQ criteria which enables clients in the pharmaceutical sector to validate their installation of REvolution R. In 2015, Microsoft Corporation completed the acquisition of Revolution Analytics. and has since integrated the R programming language into SQL Server 2016, SQL Server 2017, Power BI, Azure SQL Database, Azure Cortana Intelligence, Microsoft R Server and Visual Studio 2017. In October 2011 Oracle announced the *Big Data Appliance*, which integrates R, Apache Hadoop, Oracle Linux, and a NoSQL database with Exadata hardware. As of 2012, Oracle R Enterprise became one of two components of the "Oracle Advanced Analytics Option" (alongside Oracle Data Mining).

IBM offers support for in-Hadoop execution of R, and provides a programming model for massively parallel in-database analytics in R. Other major commercial software systems supporting connections to or integration with R include: JMP, Mathematica, MATLAB, Microsoft Power BI, Pentaho, Spotfire, SPSS, Statistica, Platform Symphony, SAS, Tableau Software, Esri ArcGis, Dundas and Statgraphics.

Tibco offers a runtime-version R as a part of Spotfire. Mango offers a validation package for R, ValidR, to make it compliant with drug approval agencies, like FDA. These agencies allow for the use of any statistical software in submissions, if only the software is validated, either by the vendor or sponsor itself.

Examples

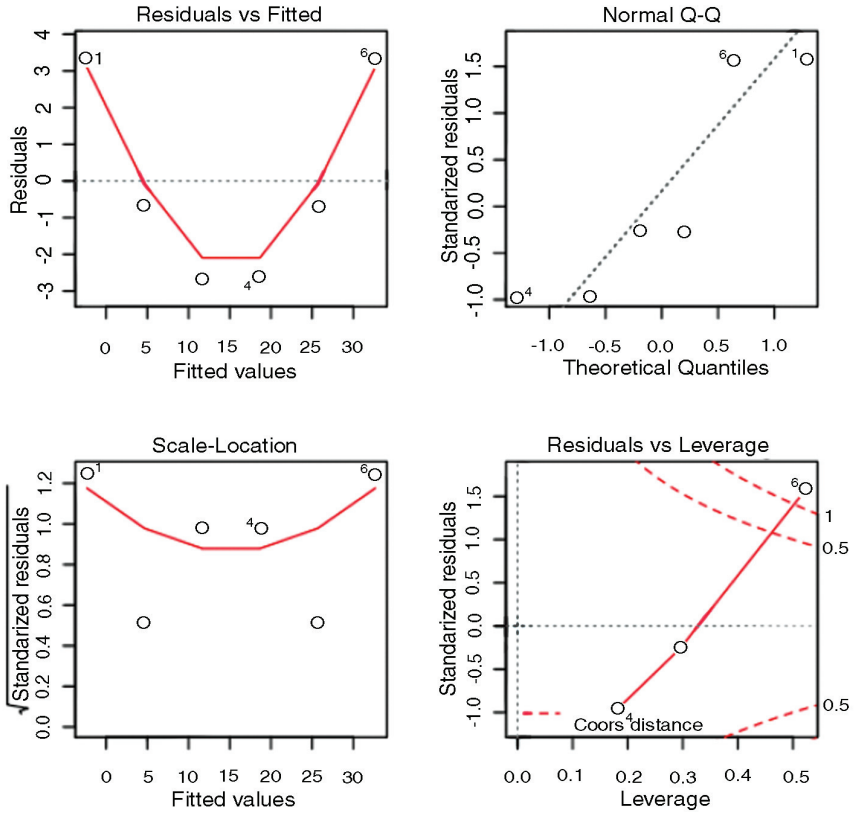
Basic Syntax

The following examples illustrate the basic syntax of the language and use of the command-line interface.

In R, the generally preferred assignment operator is an arrow made from two characters <-, although = can usually be used instead.

```
>x <-1:6# Create vector.
>y <- x^2# Create vector by formula.
>print(y)# Print the vector's contents.
[1] 1 4 9 16 25 36
>mean(y)# Arithmetic mean of vector.
[1] 15.16667
>var(y)# Sample variance of vector.
[1] 178.9667
```

```
>model <- lm(y ~ x)# Linear regression model y = A + B * x.
>print(model)# Print the model's results.
Call:
lm(formula = y ~ x)
Coefficients:
(Intercept) x
-9.333 7.000
>summary(model)# Display an in-depth summary of the model.
Call:
lm(formula = y ~ x)
Residuals:
1 2 3 4 5 6
3.3333 -0.6667 -2.6667 -2.6667 -0.6667 3.3333
Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept) -9.3333 2.8441 -3.282 0.030453 *
x 7.0000 0.7303 9.585 0.000662 ***
—-
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Residual standard error: 3.055 on 4 degrees of freedom
Multiple R-squared: 0.9583,  Adjusted R-squared: 0.9478
F-statistic: 91.88 on 1 and 4 DF, p-value: 0.000662
>par(mfrow =c(2,2))# Create a 2 by 2 layout for figures.
>plot(model)# Output diagnostic plots of the model.
```



Structure of a Function

One of R's strengths is the ease of creating new functions. Objects in the function body remain local to the function, and any data type may be returned. Here is an example user-created function:

```
# Declare function "f" with parameters "x", "y"
# that returns a linear combination of x and y.
f <-function(x, y){
z <-3* x +4* y
return(z)
}
>f(1,2)
[1] 11
>f(c(1,2,3),c(5,3,4))
[1] 23 18 25
>f(1:3,4)
[1] 19 22 25
```

Mandelbrot Set

Short R code calculating Mandelbrot set through the first 20 iterations of equation z = z + c plotted for different complex constants c.

This example demonstrates:

- Use of community-developed external libraries (called packages), in this case caTools package
- Handling of complex numbers
- Multidimensional arrays of numbers used as basic data type, see variables C, Z and X.

```
install.packages("caTools") # install external package
library(caTools) # external package providing write.gif function
jet.colors <- colorRampPalette(c("red", "blue", "#007FFF", "cyan", "#7FFF7F",
"yellow", "#FF7F00", "red", "#7F0000"))
dx <- 1500 # define width
dy <- 1400 # define height
C <- complex(real = rep(seq(-2.2, 1.0, length.out = dx), each = dy),
imag = rep(seq(-1.2, 1.2, length.out = dy), dx))
C <- matrix(C, dy, dx) # reshape as square matrix of complex numbers
Z <- 0 # initialize Z to zero
X <- array(0, c(dy, dx, 20)) # initialize output 3D array
for (k in 1:20) { # loop with 20 iterations
Z <- Z^2 + C # the central difference equation
X[,, k] <- exp(-abs(Z)) # capture results
}
write.gif(X, "Mandelbrot.gif", col = jet.colors, delay = 100)
```

ASREML

ASReml is a statistical software package for fitting linear mixed models using restricted maximum likelihood, a technique commonly used in plant and animal breeding and quantitative genetics as well as other fields.

It is notable for its ability to fit very large and complex data sets efficiently, due to its use of the average information algorithm and sparse matrix methods. It was originally developed by Arthur Gilmour. ASREML can be used in Windows, Linux, and as an add-on to S-PLUS and R.

WEKA

Waikato Environment for Knowledge Analysis (Weka) is a suite of machine learning software written in Java, developed at the University of Waikato, New Zealand. It is free software licensed under the GNU General Public License.

Weka contains a collection of visualization tools and algorithms for data analysis and predictive modeling, together with graphical user interfaces for easy access to these functions. The original non-Java version of Weka was a Tcl/Tk front-end to (mostly third-party) modeling algorithms implemented in other programming languages, plus data preprocessing utilities in C, and a Makefile-based system for running machine learning experiments. This original version was primarily designed as a tool for analyzing data from agricultural domains, but the more recent fully Java-based version (Weka 3), for which development started in 1997, is now used in many different application areas, in particular for educational purposes and research. Advantages of Weka include:

- Free availability under the GNU General Public License.
- Portability, since it is fully implemented in the Java programming language and thus runs on almost any modern computing platform.
- A comprehensive collection of data preprocessing and modeling techniques.
- Ease of use due to its graphical user interfaces.

Weka supports several standard data mining tasks, more specifically, data preprocessing, clustering, classification, regression, visualization, and feature selection. All of Weka's techniques are predicated on the assumption that the data is available as one flat file or relation, where each data point is described by a fixed number of attributes (normally, numeric or nominal attributes, but some other attribute types are also supported). Weka provides access to SQL databases using Java Database Connectivity and can process the result returned by a database query. Weka provides access to deep learning with Deeplearning4j. It is not capable of multi-relational data mining, but there is separate software for converting a collection of linked database tables into a single table that is suitable for processing using Weka. Another important area that is currently not covered by the algorithms included in the Weka distribution is sequence modeling.

Weka's main user interface is the *Explorer*, but essentially the same functionality can be accessed through the component-based *Knowledge Flow* interface and from the command line. There is also the *Experimenter*, which allows the systematic comparison of the predictive performance of Weka's machine learning algorithms on a collection of datasets.

The Explorer interface features several panels providing access to the main components of the workbench:

- The *Preprocess* panel has facilities for importing data from a database, a comma-separated values (CSV) file, etc., and for preprocessing this data using a so-called *filtering* algorithm. These filters can be used to transform the data (*e.g.*, turning numeric attributes into discrete ones) and make it possible to delete instances and attributes according to specific criteria.
- The *Classify* panel enables applying classification and regression algorithms (indiscriminately called *classifiers* in Weka) to the resulting dataset, to estimate the accuracy of the resulting predictive model, and to visualize erroneous predictions, receiver operating characteristic (ROC) curves, etc., or the model itself (if the model is amenable to visualization like, *e.g.*, a decision tree).
- The *Associate* panel provides access to association rule learners that attempt to identify all important interrelationships between attributes in the data.

- The *Cluster* panel gives access to the clustering techniques in Weka, *e.g.*, the simple k-means algorithm. There is also an implementation of the expectation maximization algorithm for learning a mixture of normal distributions.
- The *Select attributes* panel provides algorithms for identifying the most predictive attributes in a dataset.
- The *Visualize* panel shows a scatter plot matrix, where individual scatter plots can be selected and enlarged, and analyzed further using various selection operators.

Weka has a large number of regression and classification tools. Native packages are the ones included in the executible Weka software, while other non-native ones can be downloaded and used within R.Weka environment. Among the native packages, the most famous tool is the M5p model tree package. The full list of tools is available here. Some of the regression tools are:

- M5Rules (M5' algorithm presented in terms of mathematical function without a tree)
- DecisionStump (same as M5' but with a single number output in each node)
- M5P (splitting domain into successive binary regions and then fit linear models to each tree node)
- RandomForest (several model trees combined)
- RepTree (several model trees combined)
- ZeroR (the average value of outputs)
- DecisionRules (splits data into several regions based on a single independent variable and provides a single output value for each range)
- LinearRegression
- SMOreg (support vector regression)
- SimpleLinearRegression (uses an intercept and only 1 input variable for multivariate data)
- MultiLayerPerceptron (neural network)
- GaussianProcesses

In version 3.7.2, a package manager was added to allow the easier installation of extension packages. Some functionality that used to be included with Weka prior to this version has since been moved into such extension packages, but this change also makes it easier for others to contribute extensions to Weka and to maintain the software, as this modular architecture allows independent updates of the Weka core and individual extensions.

Related tools:

- Auto-WEKA is an automated machine learning system for Weka.
- Environment for DeveLoping KDD-Applications Supported by Index-Structures (ELKI) is a similar project to Weka with a focus on cluster analysis, *i.e.*, unsupervised methods.
- KNIME is a machine learning and data mining software implemented in Java.
- Massive Online Analysis (MOA) is an open-source project for large scale mining of data streams, also developed at the University of Waikato in New Zealand.
- Neural Designer is a data mining software based on deep learning techniques written in C++.
- Orange is a similar open-source project for data mining, machine learning and visualization written in Python and C++.
- RapidMiner is a commercial machine learning framework implemented in Java which integrates Weka.

ORANGE

Orange is an open-source data visualization, machine learning and data mining toolkit. It features a visual programming front-end for explorative data analysis and interactive data visualization, and can also be used as a Python library.

Orange is a component-based visual programming software package for data visualization, machine learning, data mining, and data analysis.

Orange components are called widgets and they range from simple data visualization, subset selection, and preprocessing, to empirical evaluation of learning algorithms and predictive modeling. Visual programming is implemented through an interface in which workflows are created by linking predefined or user-designed widgets, while advanced users can use Orange as a Python library for data manipulation and widget alteration.

Orange is an open-source software package released under GPL. Versions up to 3.0 include core components in C++ with wrappers in Python are available on github. From version 3.0 onwards, Orange uses common Python open-source libraries for scientific computing, such as numpy, scipy and scikit-learn, while its graphical user interface operates within the cross-platform Qt framework. Orange3 has a separate github. The default installation includes a number of machine learning, preprocessing and data visualization algorithms in 6 widget sets (data, visualize, classify, regression, evaluate and unsupervised). Additional functionalities are available as add-ons (bioinformatics, data fusion and text-mining).

Orange is supported on macOS, Windows and Linux and can also be installed from the Python Package Index repository (*pip install Orange3*). As of May 2018 the stable version is 3.13 and runs with Python 3, while the legacy version 2.7 that runs with Python 2.7 is still available.

Orange consists of a canvas interface onto which the user places widgets and creates a data analysis workflow. Widgets offer basic functionalities such as reading the data, showing a data table, selecting features, training predictors, comparing learning algorithms, visualizing data elements, etc. The user can interactively explore visualizations or feed the selected subset into other widgets.

- Canvas: graphical front-end for data analysis
- Widgets:
  - (a) Data: widgets for data input, data filtering, sampling, imputation, feature manipulation and feature selection
  - (b) Visualize: widgets for common visualization (box plot, histograms, scatter plot) and multivariate visualization (mosaic display, sieve diagram).
  - (c) Classify: a set of supervised machine learning algorithms for classification
  - (d) Regression: a set of supervised machine learning algorithms for regression
  - (e) Evaluate: cross-validation, sampling-based procedures, reliability estimation and scoring of prediction methods
  - (f) Unsupervised: unsupervised learning algorithms for clustering (k-means, hierarchical clustering) and data projection techniques (multidimensional scaling, principal component analysis, correspondence analysis).
  - (g) Add-ons:
    - Associate: widgets for mining frequent itemsets and association rule learning
    - Bioinformatics: widgets for gene set analysis, enrichment, and access to pathway libraries
    - Data fusion: widgets for fusing different data sets, collective matrix factorization, and exploration of latent factors
    - Educational: widgets for teaching machine learning concepts, such as k-means clustering, polynomial regression, stochastic gradient descent,...
    - Geo: widgets for working with geospatial data
    - Image analytics: widgets for working with images and ImageNet embeddings
    - Network: widgets for graph and network analysis
    - Text mining: widgets for natural language processing and text mining
    - Time series: widgets for time series analysis and modeling

The programme provides a platform for experiment selection, recommendation systems, and predictive modeling and is used in biomedicine, bioinformatics, genomic research, and teaching. In science, it is used as a platform for testing new machine learning algorithms and for implementing new techniques in genetics and bioinformatics. In education, it was used for teaching machine learning and data mining methods to students of biology, biomedicine, and informatics.

SCOPE AND TRAINING PROGRAMMES

Almost all educational programmes in biostatistics are at postgraduate level. They are most often found in schools of public health, affiliated with schools of medicine, forestry, or agriculture, or as a focus of application in departments of statistics.

In the United States, where several universities have dedicated biostatistics departments, many other top-tier universities integrate biostatistics faculty into statistics or other departments, such as epidemiology. Thus, departments carrying the name "biostatistics" may exist under quite different structures. For instance, relatively new biostatistics departments have been founded with a focus on bioinformatics and computational biology, whereas older departments, typically affiliated with schools of public health, will have more traditional lines of research involving epidemiological studies and clinical trials as well as bioinformatics. In larger universities around the world, where both a statistics and a biostatistics department exist, the degree of integration between the two departments may range from the bare minimum to very close collaboration. In general, the difference between a statistics programme and a biostatistics programme is twofold:

- Statistics departments will often host theoretical/methodological research which are less common in biostatistics programmes and
- Statistics departments have lines of research that may include biomedical applications but also other areas such as industry (quality control), business and economics and biological areas other than medicine.

2

Characteristics of Biostatistics

ATTACK RATE

In epidemiology, the attack rate is the biostatistical measure of frequency of morbidity, or speed of spread, in an at risk population. It is used in hypothetical predictions and during actual outbreaks of disease. An at risk population is defined as one that has no immunity to the attacking pathogen which can be either a novel pathogen or an established pathogen. It is used to project the number of victims to expect during an epidemic. This aids in marshalling resources for delivery of medical care as well as production of vaccines and/or anti-viral and anti-bacterial medicines. The rate is arrived at by taking the number of new cases in the population at risk and dividing by the number of persons at risk in the population.

$$\frac{\text{number of new cases in the population at risk}}{\text{number of persons at risk in the populatino}} = \text{Rate}$$

Rates are determined from the beginning of the outbreak to its end. The term should probably not be described as a rate because its time dimension is uncertain. While the duration of an epidemic can be predicted given other variables such as early intervention, it cannot be known in absolute terms. In epidemiology, a rate requires a defined unit change (in this instance, time) over which the rate applies. For this reason, it is often referred to as an attack ratio.

ACCURACY AND PRECISION

Precision is a description of *random errors*, a measure of statistical variability.

Accuracy has two definitions:

- More commonly, it is a description of *systematic errors*, a measure of statistical bias; as these cause a difference between a result and a "true" value, ISO calls this *trueness*.
- Alternatively, ISO defines accuracy as describing a combination of both types of observational error above (random and systematic), so high accuracy requires both high precision and high trueness.

In simplest terms, given a set of data points from repeated measurements of the same quantity, the set can be said to be *precise* if the values are close to each other, while the set can be said to be *accurate* if their average is close to the *true value* of the quantity being measured. In the first, more common definition above, the two concepts are independent of each other, so a particular set of data can be said to be either accurate, or precise, or both, or neither.

COMMON TECHNICAL DEFINITION

In the fields of science and engineering, the accuracy of a measurement system is the degree of closeness of measurements of a quantity to that quantity's true value. The precision of a measurement system, related to

reproducibility and repeatability, is the degree to which repeated measurements under unchanged conditions show the same results. Although the two words precision and accuracy can be synonymous in colloquial use, they are deliberately contrasted in the context of the scientific method. The field of statistics, where the interpretation of measurements plays a central role, prefers to use the terms *bias* and *variability* instead of accuracy and precision: bias is the amount of inaccuracy and variability is the amount of imprecision.

A measurement system can be accurate but not precise, precise but not accurate, neither, or both. For example, if an experiment contains a systematic error, then increasing the sample size generally increases precision but does not improve accuracy. The result would be a consistent yet inaccurate string of results from the flawed experiment. Eliminating the systematic error improves accuracy but does not change precision.

A measurement system is considered *valid* if it is both *accurate* and *precise*. Related terms include *bias* (non-random or directed effects caused by a factor or factors unrelated to the independent variable) and *error* (random variability). The terminology is also applied to indirect measurements—that is, values obtained by a computational procedure from observed data. In addition to accuracy and precision, measurements may also have a measurement resolution, which is the smallest change in the underlying physical quantity that produces a response in the measurement. In numerical analysis, accuracy is also the nearness of a calculation to the true value; while precision is the resolution of the representation, typically defined by the number of decimal or binary digits. In military terms, accuracy refers primarily to the accuracy of fire (or "justesse de tir"), the precision of fire expressed by the closeness of a grouping of shots at and around the centre of the target.

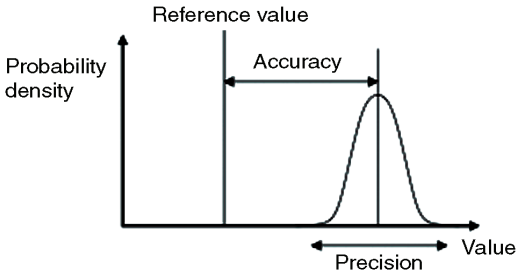


Fig. Accuracy is the proximity of measurement results to the true value; precision, the repeatability, or reproducibility of the measurement.

Quantification

In industrial instrumentation, accuracy is the measurement tolerance, or transmission of the instrument and defines the limits of the errors made when the instrument is used in normal operating conditions.

Ideally a measurement device is both accurate and precise, with measurements all close to and tightly clustered around the true value. The accuracy and precision of a measurement process is usually established by repeatedly measuring some traceable reference standard. Such standards are defined in the International System of Units (abbreviated SI from French: Système international d'unités) and maintained by national standards organizations such as the National Institute of Standards and Technology in the United States.

This also applies when measurements are repeated and averaged. In that case, the term standard error is properly applied: the precision of the average is equal to the known standard deviation of the process divided by the square root of the number of measurements averaged. Further, the central limit theorem shows that the probability distribution of the averaged measurements will be closer to a normal distribution than that of individual measurements.

With regard to accuracy we can distinguish:

- The difference between the mean of the measurements and the reference value, the bias. Establishing and correcting for bias is necessary for calibration.
- The combined effect of that and precision.

A common convention in science and engineering is to express accuracy and/or precision implicitly by means of significant figures. Here, when not explicitly stated, the margin of error is understood to be one-half the value of the last significant place. For instance, a recording of 843.6 m, or 843.0 m, or 800.0 m would imply a margin of 0.05 m (the last significant place is the tenths place), while a recording of 8436 m would imply a margin of error of 0.5 m (the last significant digits are the units).

A reading of 8,000 m, with trailing zeroes and no decimal point, is ambiguous; the trailing zeroes may or may not be intended as significant figures. To avoid this ambiguity, the number could be represented in scientific notation: 8.0 × 10 m indicates that the first zero is significant (hence a margin of 50 m) while 8.000 × 10 m indicates that all three zeroes are significant, giving a margin of 0.5 m. Similarly, it is possible to use a multiple of the basic measurement unit: 8.0 km is equivalent to 8.0 × 10 m. In fact, it indicates a margin of 0.05 km (50 m). However, reliance on this convention can lead to false precision errors when accepting data from sources that do not obey it. For example, a source reporting a number like 153,753 with precision +/- 5,000 looks like it has precision +/- 0.5. Under the convention it would have been rounded to 154,000.

Precision includes:

- *Repeatability* — the variation arising when all efforts are made to keep conditions constant by using the same instrument and operator, and repeating during a short time period; and
- *Reproducibility* — the variation arising using the same measurement process among different instruments and operators, and over longer time periods.

ISO DEFINITION (ISO 5725)

A shift in the meaning of these terms appeared with the publication of the ISO 5725 series of standards in 1994, which is also reflected in the 2008 issue of the "BIPM International Vocabulary of Metrology" (VIM), items 2.13 and 2.14.

According to ISO 5725-1, the general term "accuracy" is used to describe the closeness of a measurement to the true value. When the term is applied to sets of measurements of the same *measure*and, it involves a component of random error and a component of systematic error. In this case trueness is the closeness of the mean of a set of measurement results to the actual (true) value and precision is the closeness of agreement among a set of results.

ISO 5725-1 and VIM also avoid the use of the term "bias", previously specified in BS 5497-1, because it has different connotations outside the fields of science and engineering, as in medicine and law.

Accuracy of a target grouping according to BIPM and ISO 5725

Fig. Low accuracy due to poor precision.

Fig. Low accuracy due to poor trueness.

IN BINARY CLASSIFICATION

Accuracy is also used as a statistical measure of how well a binary classification test correctly identifies or excludes a condition. That is, the accuracy is the proportion of true results (both true positives and true negatives) among the total number of cases examined. To make the context clear by the semantics, it is often referred to as the "Rand accuracy" or "Rand index". It is a parameter of the test.

IN PSYCHOMETRICS AND PSYCHOPHYSICS

In psychometrics and psychophysics, the term *accuracy* is interchangeably used with validity and *constant error*. *Precision* is a synonym for reliability and *variable error*. The validity of a measurement instrument or psychological test is established through experiment or correlation with behaviour. Reliability is established with a variety of statistical techniques, classically through an internal consistency test like Cronbach's alpha to ensure sets of related questions have related responses, and then comparison of those related question between reference and target population.

IN LOGIC SIMULATION

In logic simulation, a common mistake in evaluation of accurate models is to compare a logic simulation model to a transistor circuit simulation model. This is a comparison of differences in precision, not accuracy. Precision is measured with respect to detail and accuracy is measured with respect to reality.

IN INFORMATION SYSTEMS

Information retrieval systems, such as databases and web search engines, are evaluated by many different metrics, some of which are derived from the confusion matrix, which divides results into true positives (documents correctly retrieved), true negatives (documents correctly not retrieved), false positives (documents incorrectly retrieved), and false negatives (documents incorrectly not retrieved). Commonly used metrics include the notions of precision and recall. In this context, precision is defined as the fraction of retrieved documents which are relevant to the query (true positives divided by true+false positives), using a set of ground truth relevant results selected by humans. Recall is defined as the fraction of relevant documents retrieved compared to the total number of relevant documents (true positives divided by true positives+false negatives). Less commonly, the metric of accuracy is used, is defined as the total number of correct classifications (true positives plus true

negatives) divided by the total number of documents. None of these metrics take into account the ranking of results. Ranking is very important for web search engines because readers seldom go past the first page of results, and there are too many documents on the web to manually classify all of them as to whether they should be included or excluded from a given search. Adding a cutoff at a particular number of results takes ranking into account to some degree. The measure precision at k, for example, is a measure of precision looking only at the top ten (k=10) search results. More sophisticated metrics, such as discounted cumulative gain, take into account each individual ranking, and are more commonly used where this is important.

ANALYSIS OF RHYTHMIC VARIANCE

In statistics, analysis of rhythmic variance (ANORVA) is a method for detecting rhythms in biological time series, published by Peter Celec (Biol Res. 2004, 37(4 Suppl A):777–82). It is a procedure for detecting cyclic variations in biological time series and quantification of their probability. ANORVA is based on the premise that the variance in groups of data from rhythmic variables is low when a time distance of one period exists between the data entries.

BEVERTON–HOLT MODEL

The Beverton–Holt model is a classic discrete-time population model which gives the expected number n_{t+1} (or density) of individuals in generation $t + 1$ as a function of the number of individuals in the previous generation,

$$n_{t+1} = \frac{R_0 n_t}{1 + n_t / M}.$$

Here R_0 is interpreted as the proliferation rate per generation and $K = (R_0 - 1) M$ is the carrying capacity of the environment. The Beverton–Holt model was introduced in the context of fisheries by Beverton and Holt (1957). Subsequent work has derived the model under other assumptions such as contest competition (Brännström and Sumpter 2005), within-year resource limited competition (Geritz and Kisdi 2004) or even as the outcome of a source-sink Malthusian patches linked by density-dependent dispersal (Bravo de la Parra et al. 2013). The Beverton–Holt model can be generalized to include scramble competition. It is also possible to include a parameter reflecting the spatial clustering of individuals.

Despite being nonlinear, the model can be solved explicitly, since it is in fact an inhomogeneous linear equation in $1/n$. The solution is

$$n_1 = \frac{Kn_0}{n_0 + (K - n_0)R_0^{-t}}$$

Because of this structure, the model can be considered as the discrete-time analogue of the continuous-time logistic equation for population growth introduced by Verhulst; for comparison, the logistic equation is

$$\frac{dN}{dt} = rN\left(1 - \frac{N}{K}\right),$$

and its solution is

$$N(t) = \frac{KN(0)}{N(0) + (K - N(0))e^{-rt}}$$

BILLS OF MORTALITY

Bills of mortality were the weekly mortality statistics in London, designed to monitor burials from 1592 to 1595 and then continuously from 1603. The responsibility to produce the statistics was chartered in 1611 to the

Worshipful Company of Parish Clerks. The bills covered an area that started to expand as London grew from the City of London, before reaching its maximum extent in 1636. New parishes were then only added where ancient parishes within the area were divided. Factors such as the use of suburban cemeteries outside the area, the exemption of extra-parochial places within the area, the wider growth of the metropolis, and that they recorded burials rather than deaths, rendered their data incomplete. Production of the bills went into decline from 1819 as parishes ceased to provide returns, with the last surviving weekly bill dating from 1858. They were superseded by the weekly returns of the Registrar General from 1840, taking in further parishes until 1847. This area became the district of the Metropolitan Board of Works in 1855, the County of London in 1889 and Inner London in 1965.

HISTORY

Bills of mortality were produced intermittently in the several parishes of the City of London during outbreaks of plague. The first bill of mortality is believed to date from November 1532. The first regular weekly collection and publishing of the number of burials in the parishes of London began on 21 December 1592 and continued until 18 December 1595. The practice was abandoned and then revived on 21 December 1603 when there was another outbreak. In 1611 the duty to produce the bills was imposed on the members of the Worshipful Company of Parish Clerks by a charter granted by James I. Annual returns were made on 21 December (the feast of St Thomas), to coincide with the city calendar. New charters were granted by Charles I in 1636 and 1639. The bills covered 129 parishes at the granting of the 1639 charter.

By 1570 the bills included baptisms; in 1629 the cause of death was given, and in the early 18th century the age at death.

In 1819 the bills ceased to be published under the authority of the Corporation of London, coming directly from the Worshipful Company of Parish Clerks. The clerk of St George Hanover Square ceased to provide returns from 1823. From then until 1858 the practice of producing bills of mortality was in decline, as parishes ceased to provide returns to the Worshipful Company of Parish Clerks. The last surviving bill of mortality is believed to be from 28 September 1858.

PROBLEMS WITH THE BILLS

The area fixed in 1636, adding only St Mary le Strand in 1726 which was already within the outer boundary of the bills. The area quickly became much smaller than the growing metropolis. The bills recorded burials in Church of England churchyards and not deaths. The bills did not include the English Dissenters, Roman Catholics or those of other faiths. From 1830 burials started to take place outside the bills area in the large suburban cemeteries. Extra-parochial places and certain churches within the area failed to give returns because they were outside the normal parish system. For example, the Church of St Peter ad Vincula in the Tower of London was added in 1729, but was excluded in 1730 because of a successful claim of being extra-parochial. These defects meant that the bills failed to record approximately a third of deaths in the Metropolis.

PLACES WITHIN THE BILLS

Formed 1767 by separating the Middlesex portion of the parish St Andrew Holborn from the remainder in the City of London and merging with the parish of St George the Martyr. Formed from part of Stepney in 1743. Formed from part of Stepney in 1729. The remainder of the parish lay in the Liberty of Westminster. The parish of St John was formed from part of St James in 1723. The two parishes of St Giles and St George were united in 1774. Formed from Stepney in 1725. Parish created 1733 from the part of St Giles Cripplegate outside the City of London.

Table. These places were within the boundaries of the bills of mortality.

County	Parts there of
City of London	Entire, comprising: 97 parishes within the Walls; 16 parishes without the Walls; Inns of Court and Chancery
Middlesex	The City and Liberty of Westminster; The Tower and its Liberty (including the Old Artillery Ground); St Andrew Holborn above Bars with St George the Martyr; St Matthew, Bethnal Green; St Botolph without Aldgate; The Charterhouse; Christchurch, Spitalfields; St Clement Danes (part); St James and St John, Clerkenwell; Liberty of the Duchy of Lancaster (part); Ely Place; St Giles in the Fields and St George, Bloomsbury; St George in the East; Liberty of Glasshouse Yard; St John, Hackney; St Mary, Islington; St Katherine near the Tower; St Ann, Limehouse; St Luke, Middlesex; Liberty of the Rolls; Liberty of Saffron Hill and Hatton Garden; St John the Baptist in the Savoy; St Sepulchre (part); St Paul, Shadwell; St Leonard, Shoreditch; St Dunstan, Stepney (the hamlets of Ratcliffe, Mile End Old Town and Mile End New Town); St John, Wapping; St Mary, Whitechapel
Surrey	The Borough of Southwark (the parishes of St George the Martyr; St John Horsleydown; St Olave; St Saviour and St Thomas and Christchurch) St Mary, Rotherhithe; St Mary, Bermondsey; St Mary, Newington Butts; St Mary, Lambeth

The remainder of the parish lay in the City of London. Formed from part of Stepney in 1670. Formed from part of Stepney in the early 17th century. Parish of Christchurch, Southwark formed 1670: originally the Liberty of Paris Garden.

POPULATION

Table. The population of the parishes in Bills of mortality area, as it was fixed in 1726, consisting of some 21,587 acres (87.36 km), was.

Year	1801	1811	1821	1831	1841
Population	746,233	856,412	1,011,948	1,180,292	1,353,345

REGISTRAR GENERAL RETURNS

Under the direction of John Rickman, the Bills of mortality area and the "five villages beyond the Bills" consisting of the parishes of Chelsea, Kensington, Marylebone, Paddington and St Pancras were designated the "Metropolis" in the 1801 to 1831 censuses.

From 11 January 1840, the bills were superseded by the Registrar General's weekly returns for the Metropolis, following the Births and Deaths Registration Act 1836. The weekly returns were based on death certificates, and therefore much more accurate than the bills of mortality based on burials. When the Registrar General began weekly returns in 1840 to the Metropolis defined in the 1831 census were added the parishes of Bow, Camberwell, Fulham, Hammersmith and the Greenwich Poor Law Union. This area was used for annual returns from 1837 and was the definition of the Metropolis in the 1841 census.

In 1844 the Wandsworth and Clapham Poor Law Union was added and in 1847 the parish of Hampstead and the Lewisham Poor Law Union were added to the weekly returns. This was the definition of the Metropolis used in the 1851 census. This area, with minor adjustments, became the district of the Metropolitan Board of Works in 1855, the County of London in 1889 and Inner London in 1965.

GUIDANCES FOR STATISTICS IN REGULATORY AFFAIRS

Guidances for statistics in regulatory affairs are applicable to the pharmaceutical industry and medical devices industry. These Guidances represent the current thinking of regulatory agencies on a particular subject. It is to be noted that the term "Guidances" is used in the USA, whereas the term "Guidelines" is used in Europe.

Regulatory affairs, also called government affairs, is a profession within regulated industries, such as pharmaceutical and medical devices, where professionals such as statisticians are expected to implement regulatory guidance into their work practices. Statisticians working in a regulated environment (*e.g.* the pharmaceutical and health care industry) are obliged to have a sound knowledge and understanding of the regulatory requirements that affect the design, conduct, analysis and reporting of their studies.

Regulatory guidance for the pharmaceutical and medical devices industry can be found at the international level (*e.g.* ICH - International Council on Harmonisation of Technical Requirements for Registration of Pharmaceuticals for Human Use), as well as at the regional/national level; for example:

- EMA - European Medicines Agency in Europe,
- MHRA - Medicines and Health care products Regulatory Agency in UK,
- IQWiG - Institute for Quality and Efficiency in Health Care in Germany,
- FDA - Food and Drug Administration in USA and
- PMDA - Pharmaceuticals and Medical Devices Agency in Japan.

Furthermore, statistical regulatory guidance is found under general topics (*e.g.* Good Clinical Practice - ICH E6(R2)) and specific ones explicitly related to statistics (*e.g.* Statistical Principles for Clinical Trials - ICH E9) or not explicitly (*e.g.* Special Populations: Geriatrics [ICH E7] or Clinical Trial Endpoints in Oncology [FDA]). This large volume and diversity of regulatory guidances (draft and/or final) is subject to revisions. Therefore, users of the guidances are advised to consult the original web site to check for the latest version. Users are also encouraged to update the Wikipedia content

HISTORY

Regulation in the USA started in 1906 with the Food and Drugs Act and further USA regulation came in 1938 with the Federal Food, Drug, and Cosmetic Act following the deaths due to Elixir sulfanilamide in 1937. Further tragedies resulted from the use of Thalidomide that was marketed in 1957, Germany, without adequate testing.

The Thalidomide catastrophe tightened the regulatory pressure in the US with the Kefauver Harris Amendment (1962) to the Federal Food, Drug, and Cosmetic Act and the European regulation appeared with UK's "Medicines Act 1968". Further developments lead to a large volume and variety of regulatory "guidances" by the Food and Drug Administration (FDA) in the USA and "guidelines" by the European Medicines Agency (EMA) in Europe. Also, other regions of the world issued regulatory guidance, for example, Pharmaceuticals and Medical Devices Agency (PMDA) in Japan. In 1989 a plan for harmonization of guidance across Europe, Japan and the USA was started and the first meeting of International Conference on Harmonisation of Technical Requirements for Registration of Pharmaceuticals for Human Use (ICH) was held in 1990, Brussels.

Regarding applications in Health technology assessment (HTA) a number of national guidance papers are available from the local HTA organizations: the Institute for Quality and Efficiency in Health Care (IQWiG) in Germany, the National Institute for Health and Care Excellence (NICE) in UK, the Agency for Health care Research and Quality (AHRQ) in USA or the Canadian Agency for Drugs and Technologies in Health (CADTH) in Canada. In Europe, the European Network for Health Technology Assessment (EUnetHTA) was established in 2005 to create an effective and sustainable network for HTA to support collaboration between European HTA organizations. EUnetHTA Guidelines have been developed to help the assessors of evidence to process, analyse and interpret the data.

GENERAL GUIDANCE

General guidance covers statistical topics that relate to good clinical practice; study design, monitoring and reporting, and market authorization of medical products or medical devices.

Good Clinical Practice

The good clinical practice (GCP) is an international ethical and scientific quality standard for designing, conducting, recording and reporting trials that involve the participation of human subjects. It was issued by ICH under Good Clinical Practice Directive (Directive 2005/28/EC) of 8 April 2005. A similar guideline for clinical trials of medical devices is the international standard ISO 14155, that is valid in the European Union as a harmonized standard. Compliance with the GCP standard provides public assurance that the rights, safety and well-being of trial subjects are protected, consistent with the principles that have their origin in the Declaration of Helsinki, and that the clinical trial data are credible.

BMDP

BMDP was a statistical package developed in 1965 by Wilfrid Dixon at the University of California, Los Angeles. The acronym stands for Bio-Medical Data Package, the word package was added by Dixon as the software consisted of a series of programmes (subroutines) which performed different parametric and nonparametric statistical analyses.

BMDP was originally distributed for free. It was later sold by Statsols, who originally was a subsidiary of BMDP, but through a management buy-out formed the now independent company Statistical Solutions Ltd, known as Statsols. BMDP is no longer available as of 2017. The company decided to only offer its other statistical product nQuery Sample Size Software.

C-PROBABILITY

In statistics, a c-probability is the probability that a contrast variable obtains a positive value. Using a replication probability, the c-probability is defined as follows: if we get a random draw from each group (or factor level) and calculate the sampled value of the contrast variable based on the random draws, then the c-probability is the chance that the sampled values of the contrast variable are greater than 0 when the random drawing process is repeated infinite times. The c-probability is a probabilistic index accounting for distributions of compared groups (or factor levels). The c-probability and SMCV are two characteristics of a contrast variable. There is a link between SMCV and c-probability. The SMCV and c-probability provides a consistent interpretation to the strength of comparisons in contrast analysis. When only two groups are involved in a comparison, the c-probability becomes d-probability which is the probability that the difference of values from two groups is positive. To some extent, the d-probability (especially in the independent situations) is equivalent to the well-established probabilistic index $P(X > Y)$. Historically, the index $P(X > Y)$ has been studied and applied in many areas. The c-probability and d-probability have been used for data analysis in high-throughput experiments and biopharmaceutical research.

CELLULAR NOISE

Cellular noise is random variability in quantities arising in cellular biology. For example, cells which are genetically identical, even within the same tissue, are often observed to have different expression levels of proteins, different sizes and structures. These apparently random differences can have important biological and medical consequences.

Cellular noise was originally, and is still often, examined in the context of gene expression levels – either the concentration or copy number of the products of genes within and between cells. As gene expression levels are responsible for many fundamental properties in cellular biology, including cells' physical appearance, behaviour in response to stimuli, and ability to process information and control internal processes, the presence of noise in gene expression has profound implications for many processes in cellular biology.

INTRINSIC AND EXTRINSIC NOISE

Cellular noise is often investigated in the framework of *intrinsic* and *extrinsic* noise. Intrinsic noise refers to variation in identically-regulated quantities within a single cell: for example, the intra-cell variation in expression levels of two identically-controlled genes. Extrinsic noise refers to variation in identically-regulated quantities between different cells: for example, the cell-to-cell variation in expression of a given gene. Intrinsic and extrinsic noise levels are often compared in dual reporter studies, in which the expression levels of two identically-regulated genes (often fluorescent reporters like GFP and YFP) are plotted for each cell in a population.

SOURCES

Note: These lists are illustrative, not exhaustive, and identification of noise sources is an active and expanding area of research.

Intrinsic noise:

- *Low copy-number effects (including discrete birth and death events)*: the random (stochastic) nature of production and degradation of cellular components means that noise is high for components at low copy number (as the magnitude of these random fluctuations is not negligible with respect to the copy number);
- *Diffusive cellular dynamics*: many important cellular processes rely on collisions between reactants (for example, RNA polymerase and DNA) and other physical criteria which, given the diffusive dynamic nature of the cell, occur stochastically.
- *Noise propagation*: Low copy-number effects and diffusive dynamics result in each of the biochemical reactions in a cell occurring randomly. Stochasticity of reactions can be either attenuated or amplified. Contribution each reaction makes to the intrinsic variability in copy numbers can be quantified via Van Kampen's system size expansion.

Extrinsic noise:

- *Cellular age/ cell cycle stage*: cells in a dividing population that is not synchronised will, at a given snapshot in time, be at different cell cycle stages, with corresponding biochemical and physical differences;
- *Physical environment (temperature, pressure,...)*: physical quantities and chemical concentrations (particularly in the case of cell-to-cell signalling) may vary spatially across a population of cells, provoking extrinsic differences as a function of position;
- *Organelle distributions*: random factors in the quantity and quality of organelles (for example, the number and functionality of mitochondria) lead to significant cell-to-cell differences in a range of processes (as, for example, mitochondria play a central role in the energy budget of eukaryotic cells);
- *Inheritance noise*: uneven partitioning of cellular components between daughter cells at mitosis can result in large extrinsic differences in a dividing population.

Note that extrinsic noise can affect levels and types of intrinsic noise: for example, extrinsic differences in the mitochondrial content of cells lead, through differences in ATP levels, to some cells transcribing faster than others, affecting the rates of gene expression and the magnitude of intrinsic noise across the population.

EFFECTS

- *Gene expression levels*: noise in gene expression causes differences in the fundamental properties of cells, limits their ability to biochemically control cellular dynamics, and directly or indirectly induce many of the specific effects below;

- *Energy levels and transcription rate*: noise in transcription rate, arising from sources including transcriptional bursting, is a significant source of noise in expression levels of genes. Extrinsic noise in mitochondrial content has been suggested to propagate to differences in the ATP concentrations and transcription rates (with functional relationships implied between these three quantities) in cells, affecting cells' energetic competence and ability to express genes;
- *Phenotype selection*: bacterial populations exploit extrinsic noise to choose a population subset to enter a quiescent state. In a bacterial infection, for example, this subset will not propagate quickly but will be more robust when the population is threatened by antibiotic treatment: the rapidly replicating, infectious bacteria will be killed more quickly than the quiescent subset, which may be capable of restarting the infection. This phenomenon is why courses of antibiotics should be finished even when symptoms seem to have disappeared;
- *Development and stem cell differentiation*: developmental noise in biochemical processes which need to be tightly controlled (for example, patterning of gene expression levels that develop into different body parts) during organismal development can have dramatic consequences, necessitating the evolution of robust cellular machinery. Stem cells differentiate into different cell types depending on the expression levels of various characteristic genes: noise in gene expression can clearly perturb and influence this process, and noise in transcription rate can affect the structure of the dynamic landscape that differentiation occurs on;
- *Cancer treatments*: recent work has found extrinsic differences, linked to gene expression levels, in the response of cancer cells to anti-cancer treatments, potentially linking the phenomenon of fractional killing (whereby each treatment kills some but not all of a tumour) to noise in gene expression. Because individual cells could repeatedly and stochastically perform transitions between states associated with differences in responsiveness to a therapeutic modality (chemotherapy, targeted agent, radiation, etc.), therapy might need to be administered frequently (to ensure cells are treated soon after entering a therapy-responsive state, before they can rejoin the therapy-resistant subpopulation and proliferate) and over long times (to treat even those cells emerging late from the final residue of the therapy-resistant subpopulation).
- *Information processing*: as cellular regulation is performed with components that are themselves subject to noise, the ability of cells to process information and perform control is fundamentally limited by intrinsic noise

ANALYSIS

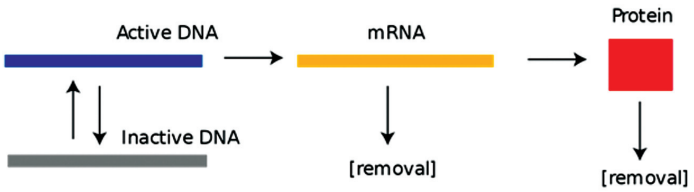


Fig. A canonical model for stochastic gene expression, known as the two-state or telegraph model. DNA flips between "inactive" and "active" states (involving, for example, chromatin remodelling and transcription factor binding). Active DNA is transcribed to produce mRNA which is translated to produce protein, both of which are degraded. All processes are Poissonian with given rates.

As many quantities of cell biological interest are present in discrete copy number within the cell (single DNAs, dozens of mRNAs, hundreds of proteins), tools from discrete stochastic mathematics are often used to analyse and model cellular noise. In particular, master equation treatments – where the probabilities P(x),t) of observing a system in a state x *at t* are linked through ODEs – have proved particularly fruitful. A canonical

model for noise gene expression, where the processes of DNA activation, transcription and translation are all represented as Poisson processes with given rates, gives a master equation which may be solved exactly (with generating functions) under various assumptions or approximated with stochastic tools like Van Kampen's system size expansion. Numerically, the Gillespie algorithm or stochastic simulation algorithm is often used to create realisations of stochastic cellular processes, from which statistics can be calculated. The problem of inferring the values of parameters in stochastic models (parametric inference) for biological processes, which are typically characterised by sparse and noisy experimental data, is an active field of research, with methods including Bayesian MCMC and approximate Bayesian computation proving adaptable and robust. Regarding the two-state model, a moment-based method was described for parameters inference from mRNAs distributions.

CIT PROGRAMME TUMOR IDENTITY CARDS

The "Cartes d'Identité des Tumeurs (CIT)" programme, launched and financed by the French charity "Ligue Nationale contre le Cancer", aims at refining the molecular knowledge of multiple types of tumors with the prospect of improving or developing better targeted therapeutic approaches. The CIT programme mainly relies on the large-scale and systematic profiling of large cohorts of tumors at various molecular levels including at least the genome, the epigenome, and the transcriptome.

CLINICAL SIGNIFICANCE

In medicine and psychology, clinical significance is the practical importance of a treatment effect—whether it has a real genuine, palpable, noticeable effect on daily life.

TYPES OF SIGNIFICANCE

Statistical Significance

Statistical significance is used in hypothesis testing, whereby the null hypothesis (that there is no relationship between variables) is tested. A level of significance is selected (most commonly $\alpha = 0.05$ or 0.01), which signifies the probability of incorrectly rejecting a true null hypothesis. If there is a significant difference between two groups at $\alpha = 0.05$, it means that there is only a 5 per cent probability of obtaining the observed results under the assumption that the difference is entirely due to chance (*i.e.*, the null hypothesis is true); it gives no indication of the magnitude or clinical importance of the difference. When statistically significant results are achieved, they favour rejection of the null hypothesis, but they do not prove that the null hypothesis is false. Likewise, non-significant results do not prove that the null hypothesis is true; they also give no evidence of the truth or falsity of the hypothesis the researcher has generated. Statistical significance relates only to the compatibility between observed data and what would be expected under the assumption that the null hypothesis is true.

Practical Significance

In broad usage, the "practical clinical significance" answers the question, *how effective* is the intervention or treatment, or how much change does the treatment causes. In terms of testing clinical treatments, practical significance optimally yields quantified information about the importance of a finding, using metrics such as effect size, number needed to treat (NNT), and preventive fraction. Practical significance may also convey semi-quantitative, comparative, or feasibility assessments of utility. Effect size is one type of practical significance. It quantifies the extent to which a sample diverges from expectations. Effect size can provide important

information about the results of a study, and are recommended for inclusion in addition to statistical significance. Effect sizes have their own sources of bias, are subject to change based on population variability of the dependent variable, and tend to focus on group effects, not individual changes. Although clinical significance and practical significance are often used synonymously, a more technical restrictive usage denotes this as erroneous. This technical use within psychology and psychotherapy not only results from a carefully drawn precision and particularity of language, but it enables a shift in perspective from group effects to the specifics of change(s) within an individual.

Specific Usage

In contrast, when used as a technical term within psychology and psychotherapy, clinical significance yields information on whether a treatment was effective enough to change a patient's diagnostic label. In terms of clinical treatment studies, clinical significance answers the question "Is a treatment effective enough to cause the patient to be normal [with respect to the diagnostic criteria in question]?"

For example, a treatment might significantly change depressive symptoms (statistical significance), the change could be a large decrease in depressive symptoms (practical significance- effect size), and 40 per cent of the patients no longer met the diagnostic criteria for depression (clinical significance). It is very possible to have a treatment that yields a significant difference and medium or large effect sizes, but does not move a patient from dysfunctional to functional. Within psychology and psychotherapy, clinical significance was first proposed by Jacobson, Follette, and Revenstorf as a way to answer the question, is a therapy or treatment effective enough such that a client does not meet the criteria for a diagnosis? Jacobson and Truax later defined clinical significance as "the extent to which therapy moves someone outside the range of the dysfunctional population or within the range of the functional population." They proposed two components of this index of change: the status of a patient or client after therapy has been completed, and "how much change has occurred during the course of therapy."

Clinical significance is also a consideration when interpreting the results of the psychological assessment of an individual. Frequently, there will be a difference of scores or subscores that is statistically significant, unlikely to have occurred purely by chance. However, not all of those statistically significant differences are clinically significant, in that they do not either explain existing information about the client, or provide useful direction for intervention. Differences that are small in magnitude typically lack practical relevance and are unlikely to be clinically significant. Differences that are common in the population are also unlikely to be clinically significant, because they may simply reflect a level of normal human variation. Additionally, clinicians look for information in the assessment data and the client's history that corroborates the relevance of the statistical difference, to establish the connection between performance on the specific test and the individual's more general functioning.

CALCULATION OF CLINICAL SIGNIFICANCE

Just as there are many ways to calculate statistical significance and practical significance, there are a variety of ways to calculate clinical significance. Five common methods are the Jacobson-Truax method, the Gulliksen-Lord-Novick method, the Edwards-Nunnally method, the Hageman-Arrindell method, and hierarchical linear modeling.

Jacobson-Truax

Jacobson-Truax is common method of calculating clinical significance. It involves calculating a Reliability Change Index (RCI). The RCI equals the difference between a participant's pre-test and post-test scores, divided

by the standard error of the difference. Cutoff scores are established for placing participants into one of four categories: recovered, improved, unchanged, or deteriorated, depending on the directionality of the RCI and whether the cutoff score was met.

Gulliksen-Lord-Novick

The Gulliksen-Lord-Novick method is similar to Jacobson-Truax, except that it takes into account regression to the mean. This is done by subtracting the pre-test and post-test scores from a population mean, and dividing by the standard deviation of the population.

Edwards-Nunnally

The Edwards-Nunnally method of calculating clinical significance is a more stringent alternative to the Jacobson-Truax method. Reliability scores are used to bring the pre-test scores closer to the mean, and then a confidence interval is developed for this adjusted pre-test score. Confidence intervals are used when calculating the change from pre-test to post-test, so greater actual change in scores is necessary to show clinical significance, compared to the Jacobson-Truax method.

Hageman-Arrindell

The Hageman-Arrindell calculation of clinical significance involves indices of group change and of individual change. The reliability of change indicates whether a patient has improved, stayed the same, or deteriorated. A second index, the clinical significance of change, indicates four categories similar to those used by Jacobson-Truax: deteriorated, not reliably changed, improved but not recovered, and recovered.

Hierarchical Linear Modeling (HLM)

HLM involves growth curve analysis instead of pre-test post-test comparisons, so three data points are needed from each patient, instead of only two data points (pre-test and post-test). A computer programme, such as Hierarchical Linear and Nonlinear Modeling is used to calculate change estimates for each participant. HLM also allows for analysis of growth curve models of dyads and groups.

COHEN'S H

In statistics, Cohen's h, popularized by Jacob Cohen, is a measure of distance between two proportions or probabilities. Cohen's h has several related uses:

- It can be used to describe the difference between two proportions as "small", "medium", or "large".
- It can be used to determine if the difference between two proportions is "meaningful".
- It can be used in calculating the sample size for a future study.

When measuring differences between proportions, Cohen's h can be used in conjunction with hypothesis testing. A "statistically significant" difference between two proportions is understood to mean that, given the data, it is likely that there is a difference in the population proportions. However, this difference might be too small to be meaningful—the statistically significant result does not tell us the size of the difference. Cohen's h, on the other hand, quantifies the size of the difference, allowing us to decide if the difference is meaningful.

USES

Researchers have used Cohen's h as follows.

- Describe the differences in proportions using the rule of thumb criteria set out by Cohen. Namely, $h = 0.2$ is a "small" difference, $h = 0.5$ is a "medium" difference, and $h = 0.8$ is a "large" difference.

- Only discuss differences that have *h* greater than some threshold value, such as 0.2.
- When the sample size is so large that many differences are likely to be statistically significant, Cohen's h identifies "meaningful", "clinically meaningful", or "practically significant" differences.

CALCULATION

Given a probability or proportion *p*, between 0 and 1, its "arcsine transformation" is

$$\phi = 2 \arcsin \sqrt{p} \quad \phi = 2 \arcsin \sqrt{p}$$

Given two proportions, p_1 and p_2, *h* is defined as the difference between their arcsine transformations. Namely,

$$h = \phi_1 - \phi_2$$

This is also sometimes called "directional h" because, in addition to showing the magnitude of the difference, it shows which of the two proportions is greater.

Often, researchers mean "nondirectional h", which is just the absolute value of the directional h:

$$h = |\phi_1 - \phi_2|$$

In R, Cohen's h can be calculated using the ES.h function in the pwr package.

INTERPRETATION

Cohen provides the following descriptive interpretations of h as a rule of thumb:

- *h* = 0.20: "small effect size".
- *h* = 0.50: "medium effect size".
- *h* = 0.80: "large effect size".

Cohen cautions that:

- As before, the reader is counseled to avoid the use of these conventions, if he can, in favour of exact values provided by theory or experience in the specific area in which he is working.

Nevertheless, many researchers do use these conventions as given.

COHORT

In statistics, marketing and demography, a cohort is a group of subjects who share a defining characteristic (typically subjects who experienced a common event in a selected time period, such as birth or graduation). Cohort data can oftentimes be more advantageous to demographers than period data. Because cohort data is honed to a specific time period, it is usually more accurate. It is more accurate because it can be tuned to retrieve custom data for a specific study. In addition, cohort data is not affected by tempo effects, unlike period data. On the contrary; cohort data can be disadvantageous in the sense that it can take a long amount of time to collect the data necessary for the cohort study. Another disadvantage of cohort studies is that it can be extremely costly to carry out, since the study will go on for a long period of time, demographers often require sufficient funds to fuel the study.

Demography often contrasts cohort perspectives and period perspectives. For instance, the total cohort fertility rate is an index of the average completed family size for cohorts of women, but since it can only be known for women who have finished child-bearing, it cannot be measured for currently fertile women. It can be calculated as the sum of the cohort's age-specific fertility rates that obtain as it ages through time. In contrast, the total period fertility rate uses current age-specific fertility rates to calculate the completed family size for a notional woman, were she to experience these fertility rates through her life.

A study on a cohort is a cohort study.

Two important aspects of cohort studies are:

- *Prospective Cohort Study:* In this type of study, there is a collection of exposure data (baseline data) from the subjects recruited before development of the outcomes of interest. The subjects are then followed through time (future) to record when the subject develops the outcome of interest. Ways to follow-up with subjects of the study include: phone interviews, face-to-face interviews, physical exams, medical/laboratory tests, and mail questionnaires. An example of a prospective cohort study is, for instance, if a demographer wanted to measure all the males births in the year 2018. The demographer would have to wait for the event to be over, the year 2018 must come to an end in order for the demographer to have all the necessary data.
- *Retrospective Cohort Study:* Retrospective Studies start with subjects that are at risk to have the outcome or disease of interest and identifies the outcome starting from where the subject is when the study starts to the past of the subject to identify the exposure. Retrospective use records: clinical, educational, birth certificates, death certificates, etc. but that may be difficult because there may not be data for the study that is being initiated. These studies may have multiple exposures which may make this study difficult. On the other hand, an example of a retrospective cohort study is, if a demographer was examining a group of people born in year 1970 who have type 1 diabetes. The demographer would begin by looking at historical data. However, if the demographer was looking at ineffective data in attempts to deduce the source of type 1 diabetes, the demographers results would not be accurate.

COLONY-FORMING UNIT

In microbiology, a colony-forming unit (CFU, cfu, Cfu) is a unit used to estimate the number of viable bacteria or fungal cells in a sample. Viable is defined as the ability to multiply via binary fission under the controlled conditions. Counting with colony-forming units requires culturing the microbes and counts only viable cells, in contrast with microscopic examination which counts all cells, living or dead. The visual appearance of a colony in a cell culture requires significant growth, and when counting colonies it is uncertain if the colony arose from one cell or a group of cells. Expressing results as colony-forming units reflects this uncertainty.

THEORY

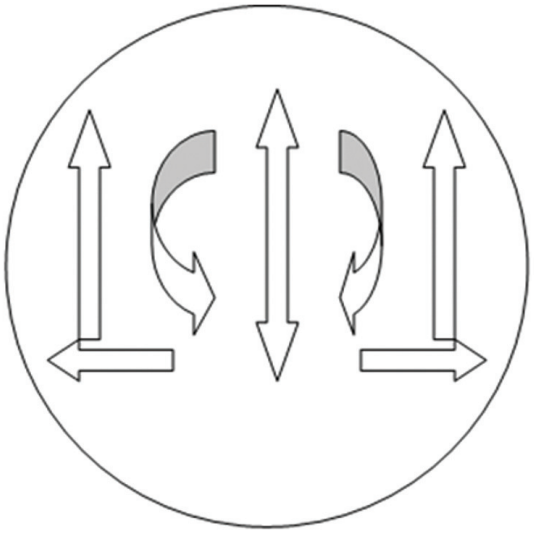


Fig. A dilution made with bacteria and peptoned water is placed in an Agar plate (*Agar plate count* for food samples or *Trypticase soy agar* for clinic samples) and spread over the plate by tipping in the pattern shown.

The purpose of plate counting is to estimate the number of cells present based on their ability to give rise to colonies under specific conditions of nutrient medium, temperature and time. Theoretically, one viable cell can give rise to a colony through replication. However, solitary cells are the exception in nature, and most likely the progenitor of the colony was a mass of cells deposited together. In addition, many bacteria grow in chains (*e.g.* Streptococcus) or clumps (*e.g.* Staphylococcus). Estimation of microbial numbers by CFU will, in most cases, undercount the number of living cells present in a sample for these reasons. This is because the counting of CFU assumes that every colony is separate and founded by a single viable microbial cell.

The plate count is linear for *E. coli* over the range of 30 - 300 CFU on a standard sized Petri dish. Therefore, to ensure that a sample will yield CFU in this range requires dilution of the sample and plating of several dilutions. Typically ten-fold dilutions are used, and the dilution series is plated in replicates of 2 or 3 over the chosen range of dilutions. The CFU/plate is read from a plate in the linear range, and then the CFU/g (or CFU/mL) of the original is deduced mathematically, factoring in the amount plated and its dilution factor.

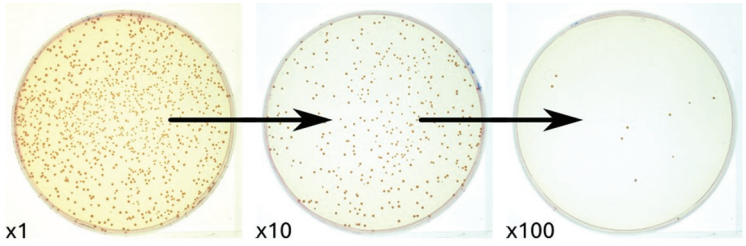


Fig. A solution of bacteria at an unknown concentration is often serially diluted in order to obtain at least one plate with a countable number of bacteria. In this figure, the "x10" plate is suitable for counting.

An advantage to this method is that different microbial species may give rise to colonies that are clearly different from each other, both microscopically and macroscopically. The colony morphology can be of great use in the identification of the microorganism present. A prior understanding of the microscopic anatomy of the organism can give a better understanding of how the observed CFU/mL relates to the number of viable cells per milliliter. Alternatively it is possible to decrease the average number of cells per CFU in some cases by vortexing the sample before conducting the dilution. However many microorganisms are delicate and would suffer a decrease in the proportion of cells that are viable when placed in a vortex.

Log Notation

Concentrations of colony-forming units can be expressed using logarithmic notation, where the value shown is the base 10 logarithm of the concentration.

USES

Colony-forming units are used to quantify results in many microbiological plating and counting methods, including:

- The Pour Plate method wherein the sample is suspended in a petri dish using molten agar cooled to approximately 40-45 °C (just above the point of solidification to minimize heat-induced cell death). After the nutrient agar solidifies the plate is incubated.
- The Spread Plate method wherein the sample (in a small volume) is spread across the surface of a nutrient agar plate and allowed to dry before incubation for counting.
- The Membrane Filter method wherein the sample is filtered through a membrane filter, then the filter placed on the surface of a nutrient agar plate (bacteria side up). During incubation nutrients leach up through the filter to support the growing cells. As the surface area of most filters is less than that of a standard petri dish, the linear range of the plate count will be less.

- The Miles and Misra Methods or drop-plate method wherein a very small aliquot (usually about 10 microliters) of sample from each dilution in series is dropped onto a petri dish. The drop dish must be read while the colonies are very small to prevent the loss of CFU as they grow together.

However, with the techniques that require the use of an agar plate, no fluid solution can be used because the purity of the specimen cannot be unidentified and it is not possible to count the cells one by one in the liquid.

TOOLS FOR COUNTING COLONIES

Counting colonies is traditionally performed manually using a pen and a click-counter. This is generally a straightforward task, but can become very laborious and time-consuming when many plates have to be enumerated. Alternatively semi-automatic (software) and automatic (hardware + software) solutions can be used.

Software for Counting CFUs

Colonies can be enumerated from pictures of plates using software tools. The experimenters would generally take a picture of each plate they need to count and then analyse all the pictures (this can be done with a simple digital camera or even a webcam). Since it takes less than 10 seconds to take a single picture, as opposed to several minutes to count CFU manually, this approach generally saves a lot of time. In addition, it is more objective and allows extraction of other variables such as the size and colour of the colonies.

- OpenCFU[1] is a free and open-source programme designed to optimise user friendliness, speed and robustness. It offers a wide range of filters and control as well as a modern user interface. OpenCFU is written in C++ and uses OpenCV for image analysis.
- NICE is a programme written in MATLAB providing an easy way to count colonies from images.
- ImageJ and CellProfiler: Some ImageJ macros and plugins and some CellProfiler pipelines can be used to count colonies. This often requires the user to change the code in order to achieve an efficient work-flow, but can prove useful and flexible. One main issue is the absence of specific GUI which can make the interaction with the processing algorithms tedious.

In addition to software based on traditional desktop computers, apps for both Android and iOS devices are available for semi-automated and automated colony counting. The integrated camera is used to take pictures of the agar plate and either an internal or an external algorithm is used to process the picture data and to estimate the number of colonies.

Automated Systems

Many of the automated systems are used to counteract human error as many of the research techniques done by humans counting individual cells have a high chance of error involved. Due to the fact that researchers regularly manually count the cells with the assistance of a transmitted light, this error prone technique can have a significant effect on the calculated concentration in the main liquid medium when the cells are in low numbers.

Completely automated systems are also available from some biotechnology manufacturers. They are generally expensive and not as flexible as standalone software since the hardware and software are designed to work together for a specific set-up. Alternatively, some automatic systems use the spiral plating paradigm. Some of the automated systems such as the systems from MATLAB allow the cells to be counted without having to stain them. This lets the colonies to be reused for other experiments without the risk of killing the microorganisms with stains. However, a disadvantage to these automated systems is that it is extremely difficult to differentiate between the microorganisms with dust or scratches on blood agar plates because both the dust and scratches can create a highly diverse combination of shapes and appearances.

Alternative Units

Instead of colony-forming units, the parameters Most Probable Number (MPN) and Modified Fishman Units (MFU) can be used. The Most Probable Number method counts viable cells and is useful when enumerating low concentrations of cells or enumerating microbes in products where particulates make plate counting impractical. Modified Fishman Units take into account bacteria which are viable, but non-culturable.

COMPANION DIAGNOSTIC

A companion diagnostic (CDx) is a diagnostic test used as a companion to a therapeutic drug to determine its applicability to a specific person. Companion diagnostics are co-developed with drugs to aid in selecting or excluding patient groups for treatment with that particular drug on the basis of their biological characteristics that determine responders and non-responders to the therapy. Companion diagnostics are developed based on companion biomarkers, biomarkers that prospectively help predict likely response or severe toxicity.

COMPLEX SYSTEMS BIOLOGY

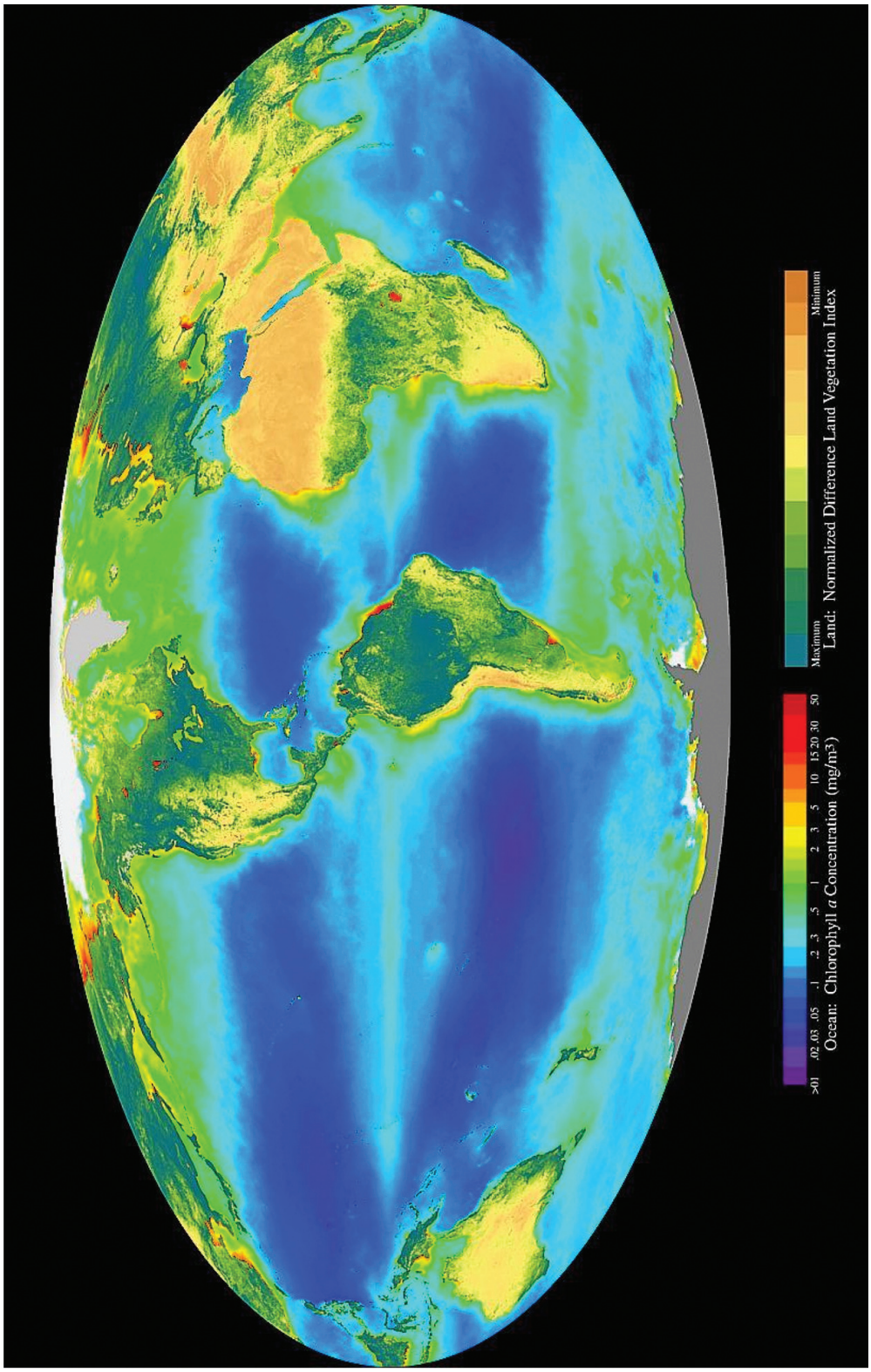
Complex systems biology (CSB) is a branch or subfield of mathematical and theoretical biology concerned with complexity of both structure and function in biological organisms, as well as the emergence and evolution of organisms and species, with emphasis being placed on the complex interactions of, and within, bionetworks, and on the fundamental relations and relational patterns that are essential to life. CSB is thus a field of theoretical sciences aimed at discovering and modeling the relational patterns essential to life that has only a partial overlap with complex systems theory, and also with the systems approach to biology called systems biology; this is because the latter is restricted primarily to simplified models of biological organization and organisms, as well as to only a general consideration of philosophical or semantic questions related to complexity in biology. Moreover, a wide range of abstract theoretical complex systems are studied as a field of applied mathematics, with or without relevance to biology, chemistry or physics.

COMPLEXITY OF ORGANISMS AND BIOSPHERE

A complete definition of complexity for individual organisms, species, ecosystems, biological evolution and the biosphere has eluded researchers, and still is an ongoing issue.

Most complex system models are often formulated in terms of concepts drawn from statistical physics, information theory and non-linear dynamics; however, such approaches are not focused on, or do not include, the conceptual part of complexity related to organization and topological attributes or algebraic topology, such as network connectivity of genomes, interactomes and biological organisms that are very important. Recently, the two complementary approaches based both on information theory, network topology/abstract graph theory concepts are being combined for example in the fields of neuroscience and human cognition. It is generally agreed that there is a hierarchy of complexity levels of organization that should be considered as distinct from that of the levels of reality in ontology. The hierarchy of complexity levels of organization in the biosphere is also recognized in modern classifications of taxonomic ranks, such as: biological domain and biosphere, biological kingdom, Phylum, biological class, order, family, genus and species. Because of their dynamic and composition variability, intrinsic "fuzziness", autopoietic attributes, ability to self-reproduce, and so on, organisms do not fit into the 'standard' definition of general systems, and they are therefore 'super-complex' in both their function and structure; organisms can be thus be defined in CSB only as 'meta-systems' of simpler dynamic systems Such a meta-system definition of organisms, species, 'ecosystems', and so on, is not equivalent to the definition of a *system of systems* as in Autopoietic Systems Theory,; it also differs from the definition proposed for example by K.D. Palmer in meta-system engineering, organisms being quite different from machines and automata

with fixed input-output transition functions, or a continuous dynamical system with fixed phase space, contrary to the Cartesian philosophical thinking; thus, organisms cannot be defined merely in terms of a quintuple *A* of *(states, startup state, input and output sets/alphabet, transition function)*, although 'non-deterministic automata', as well as 'fuzzy automata' have also been defined.



Tessellation or cellular automata provide however an intuitive, visual/computational insight into the lower levels of complexity, and have therefore become an increasingly popular, discrete model studied in computability theory, applied mathematics, physics, computer science, theoretical biology/systems biology, cancer simulations and microstructure modeling. Evolving cellular automata using genetic algorithms is also an emerging field attempting to bridge the gap between the tessellation automata and the higher level complexity approaches in CSB.

TOPICS IN COMPLEX SYSTEMS BIOLOGY

The following is only a partial list of topics covered in complex systems biology:

- Organisms and species relations and evolution
- Interactions among Species
- Evolution theories and population genetics
  - (a) Population genetics models
  - (b) Epigenetics
  - (c) Molecular evolution theories
- Quantum biocomputation
- Quantum genetics
- Relational biology
- Self-reproduction (also called self-replication in a more general context)
- Computational gene models
  - (a) DNA topology
  - (b) DNA sequencing theory
- Evolutionary developmental biology
- Autopoiesis
- Protein folding
- Telomerase conformations and functions *in vivo*
- Epigenetics
- Interactomics
- Cell signaling
- Signal transduction networks
- Complex neural nets
- Genetic networks
- Morphogenesis
- Digital morphogenesis
- Complex adaptive systems
- Topological models of morphogenesis
- Population dynamics of fisheries
- Epidemiology
- Theoretical ecology
- Immune system

DIAGNOSTIC ODDS RATIO

In medical testing with binary classification, the diagnostic odds ratio is a measure of the effectiveness of a diagnostic test. It is defined as the ratio of the odds of the test being positive if the subject has a disease relative to the odds of the test being positive if the subject does not have the disease.

The rationale for the diagnostic odds ratio is that it is a single indicator of test performance (like accuracy and Youden's J statistic) but which is independent of prevalence (unlike accuracy) and is presented as an odds ratio, which is familiar to medical practitioners.

INTERPRETATION

The diagnostic odds ratio ranges from zero to infinity, although for useful tests it is greater than one, and higher diagnostic odds ratios are indicative of better test performance. Diagnostic odds ratios less than one indicate that the test can be improved by simply inverting the outcome of the test – the test is in the wrong direction, while a diagnostic odds ratio of exactly one means that the test is equally likely to predict a positive outcome whatever the true condition – the test gives no information.

CRITICISMS

The diagnostic odds ratio is undefined when the number of false negatives *or* false positives is zero – if both false negatives *and* false positives are zero, then the test is perfect, but if only one is, this ratio does not give a usable measure. The typical response to such a scenario is to add 0.5 to all cells in the contingency table, although this should not be seen as a correction as it introduces a bias to results. It is suggested that the adjustment is made to all contingency tables, even if there are no cells with zero entries.

DILUTION ASSAY

The term dilution assay is generally used to designate a special type of bioassay in which one or more preparations (*e.g.* a drug) are administered to experimental units at different dose levels inducing a measurable biological response. The dose levels are prepared by dilution in a diluent that is inert in respect of the response. The experimental units can for example be cell-cultures, tissues, organs or living animals. The biological response may be quantal (*e.g.* positive/negative) or quantitative (*e.g.* growth). The goal is to relate the response to the dose, usually by interpolation techniques, and in many cases to express the potency/activity of the test preparation(s) relative to a standard of known potency/activity. Dilution assays can be direct or indirect. In a direct dilution assay the amount of dose needed to produce a specific (fixed) response is measured, so that the dose is a stochastic variable defining the tolerance distribution. Conversely, in an indirect dilution assay the dose levels are administered at fixed dose levels, so that the response is a stochastic variable.

SOFTWARE

The major statistical software packages do not cover dilution assays although a statistician should not have difficulties to write suitable scripts or macros to that end. Several special purpose software packages for dilution assays exist.

EPIDATA

EpiData is a group of applications used in combination for creating documented data structures and analysis of quantitative data. The EpiData Association, which created the software, was created in 1999 and is based in Denmark. EpiData was developed in Pascal and uses open standards such as HTML where possible.

EpiData is widely used by organizations and individuals to create and analyze large amounts of data. The World Health Organization (WHO) uses EpiData in its STEPS method of collecting epidemiological, medical, and public health data, for biostatistics, and for other quantitative-based projects. Epicentre, the research wing of Médecins Sans Frontières, uses EpiData to manage data from its international research studies and field epidemiology studies. *E.g.*: Piola P, Fogg C et al.: Supervised versus unsupervised intake of six-dose artemether-

lumefantrine for treatment of acute, uncomplicated Plasmodium falciparum malaria in Mbarara, Uganda: a randomised trial. Lancet. 2005 Apr 23-29;365(9469):1467-73 'PMID 15850630'. Other examples: 'PMID 16765397', 'PMID 15569777' or 'PMID 17160135'.

EpiData has two parts:

- Epidata Entry – used for simple or programmed data entry and data documentation. It handles simple forms or related systems
- EpiData Analysis – performs basic statistical analysis, graphs, and comprehensive data management, such as recoding data, label values and variables, and basic statistics. This application can create control charts, such as pareto charts or p-charts, and many other methods to visualize and describe statistical data.

The software is free; development is funded by governmental and non-governmental organizations like WHO.

EXPERIMENTAL EVENT RATE

In epidemiology and biostatistics, the experimental event rate (EER) is a measure of how often a particular statistical event (such as response to a drug, adverse event or death) occurs within the experimental group (non-control group) of an experiment.

This value is very useful in determining the therapeutic benefit or risk to patients in experimental groups, in comparison to patients in placebo or traditionally treated control groups.

Three statistical terms rely on EER for their calculation: absolute risk reduction, relative risk reduction and number needed to treat.

CONTROL EVENT RATE

The *control event rate (CER)* is identical to the experimental event rate except that is measured within the scientific control group of an experiment.

WORKED EXAMPLE

In a trial of hypothetical drug "X" where we are measuring event "Z", we have two groups.

Our control group (25 people) is given a placebo, and the experimental group (25 people) is given drug "X".

Event "Z" in control group: 4 in 25 people Control event rate: 4/25

Event "Z" in experimental group: 12 in 25 people Experimental event rate: 12/25

Table. Another worked example is as follows.

	Example 1: risk reduction			**Example 2: risk increase**		
	Experimental group (E)	**Control group (C)**	**Total**	**(E)**	**(C)**	**Total**
Events (E)	EE = 15	CE = 100	115	EE = 75	CE = 100	175
Non-events (N)	EN = 135	CN = 150	285	EN = 75	CN = 150	225
Total subjects (S)	ES = EE + EN = 150	CS = CE + CN = 250	400	ES = 150	CS = 250	400
Event rate (ER)	EER = EE / ES = 0.1, or 10%	CER = CE / CS = 0.4, or 40%		EER = 0.5 (50%)	CER = 0.4 (40%)	

GENSTAT

Genstat (General Statistics) is a statistical software package with data analysis capabilities, particularly in the field of agriculture.

Equation	Variable	Abbr.	Example 1	Example 2
EER – CER	< 0: absolute risk reduction	ARR	(–)0.3, or (–)30%	N/A
	> 0: absolute risk increase	ARI	N/A	0.1, or 10%
(EER – CER) / CER	< 0: relative risk reduction	RRR	(–)0.75, or (–)75%	N/A
	> 0: relative risk increase	RRI	N/A	0.25, or 25%
1 / (EER – CER)	< 0: number needed to treat	NNT	(–)3.33	N/A
	> 0: number needed to harm	NNH	N/A	10
EER / CER	relative risk	RR	0.25	1.25
(EE / EN) / (CE / CN)	odds ratio	OR	0.167	1.5
EER – CER	attributable risk	AR	(–)0.30, or (–)30%	0.1, or 10%
(RR – 1) / RR	attributable risk percent	ARP	N/A	20%
1 – RR (or 1 – OR)	preventive fraction	PF	0.75, or 75%	N/A

Since 1968, it has been developed by many scientific experts in Rothamsted Research, and has a user-friendly interface, professional modular design, excellent linear mixed models and graphic functions. Leading Genstat's continued development and distribution is VSN International (VSNi), which is owned by The Numerical Algorithms Group and Rothamsted Research. Genstat is used in a number of research areas, including plant science, forestry, animal science, and medicine, and is recognized by several world-class universities and enterprises.

APPLICATIONS

Genstat's statistical software can be applied to the following user areas:

- Agriculture (Animal and Plant)
- Biology, Genetics
- Ecology, Environment (Forestry and Soil)
- Food Science
- Medical and Pharmaceutical
- Finance
- Industry, Engineering
- Statistics and Mathematics

SOFTWARE PRODUCT

Statistical Features

- Manage data on Genstat's own spreadsheet (vector, scalar, table, matrix);
- Compatible with Excel spreadsheets (import/export);
- Illustrate data with graphics such as histograms, boxplots, scatter plots, line graphs, trellis plots, contour 3-dimensional surface plots, Kernel plots, species, Variogram, Regular grid, Irregular grid, circular plots and polar plots;
- Summarize and compare data with tabular reports, fitted distributions, and standard tests, such as t-tests, Chi-square tests, ANOVA, regression, and various nonparametric tests;
- Transform data using a general calculation facility with a wide range of mathematical and statistical functions;
- Model relationships between variables by linear or nonlinear regression, generalized linear models, generalized additive models, generalized linear mixed models or hierarchical generalized linear models, Logistics regression, Multinomial regression;

- Analyze experimental Design, ranging from One-Way ANOVA, Two-way ANOVA, Factorial Design, complex designs with several sources of error variation, using a balanced-ANOVA or a REML approach (including the modeling of correlation structures);
- Design investigations deciding on the sample size, or numbers of replicates, required to detect the anticipated treatment effects;
- Identify patterns in data by means of Multivariate techniques such as Canonical Variates Analysis, Discriminant Analysis, Factor Analysis, Cluster Analysis, Principal Components Analysis, principal coordinates analysis, MANOVA, correspondence analysis, partial least squares, classification trees and cluster analysis;
- Analyze results from Stratified Sampling or from Unstructured surveys, Simple Random Sampling, Cluster sampling;
- Analyze Six Sigma, plot Control charts, print Pareto tables and calculate capability statistics;
- Analyze Time Series, using Box-Jenkins Models or spectral analysis, Moving Average, ARIMA, Season Models;
- Analyze repeated measurements, by profile plot, analysis of variance, Multivariate, Generalized Estimating Equations, or using ante dependence structure, or by modeling the correlation over time;
- Analyze spatial patterns, using Variogram, Kriging, Automatic Analysis of Row-Column Design, Incomplete Block Design, or spatial point processes.

GROWTH CURVE (STATISTICS)

The growth curve model in statistics is a specific multivariate linear model, also known as GMANOVA (Generalized Multivariate ANalysis-Of-VAriance). It generalizes MANOVA by allowing post-matrices, as seen in the definition.

APPLICATIONS

GMANOVA is frequently used for the analysis of surveys, clinical trials, and agricultural data, as well as more recently in the context of Radar adaptive detection.

OTHER USES

In mathematical statistics, growth curves such as those used in biology are often modeled as being continuous stochastic processes, *e.g.* as being sample paths that almost surely solve stochastic differential equations. Growth curves have been also applied in forecasting market development.

HEALTH INDICATOR

Health indicators are quantifiable characteristics of a population which researchers use as supporting evidence for describing the health of a population. Typically, researchers will use a survey methodology to gather information about certain people, use statistics in an attempt to generalize the information collected to the entire population, then use the statistical analysis to make a statement about the health of the population.

Health indicators are often used by governments to guide health care policy.

EXAMPLE

A common example of a health indicator is life expectancy. A government might have a system for collecting information on each citizen's age at the time of death. This data about age at death can be used to support

statements about the national life expectancy, in which case life expectancy would be a “health indicator”. Life expectancy may be one of many “health indicators” which collectively researchers would use to describe the health of the population of the country.

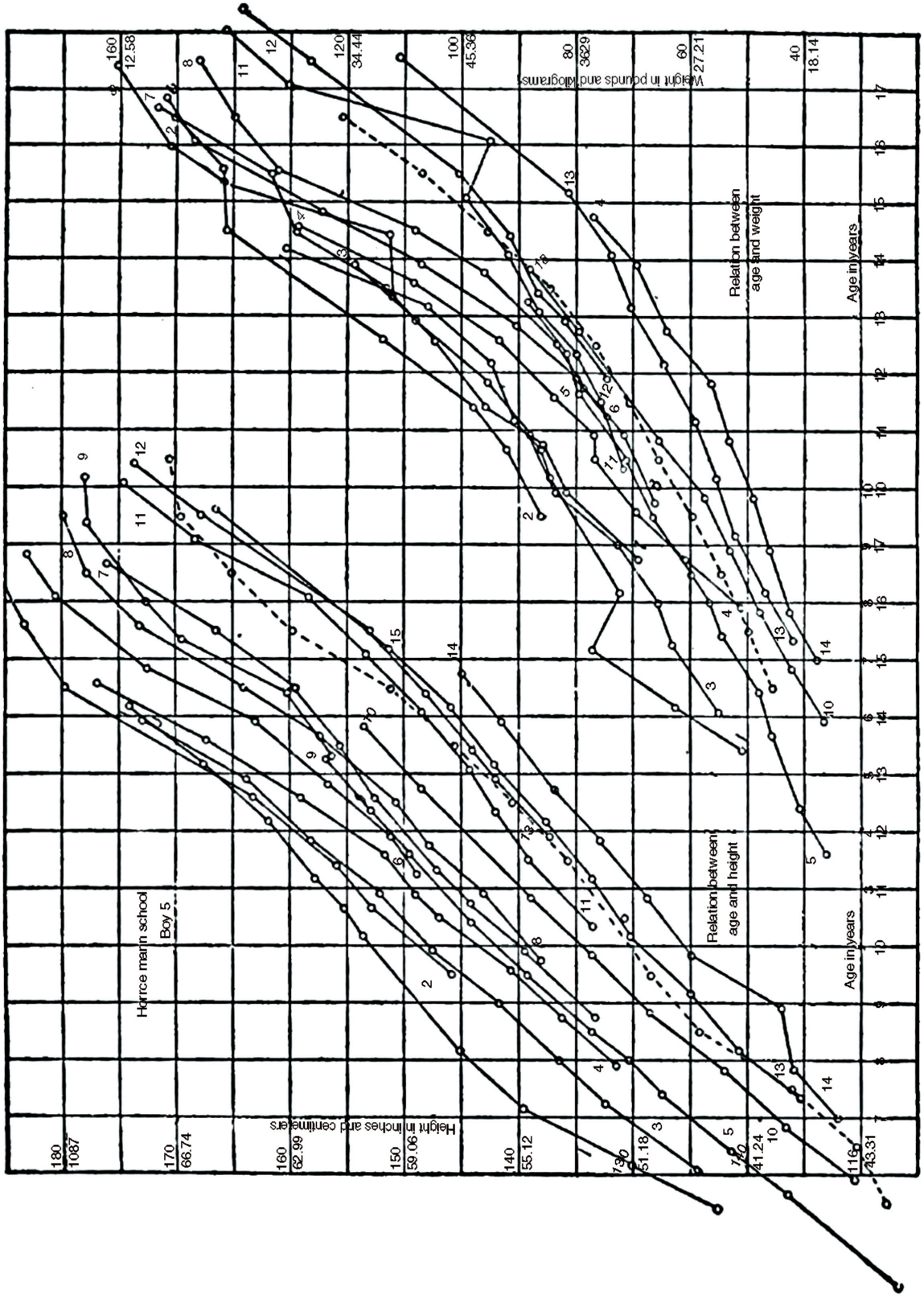


Fig. Table of height and weight for boys over time. The growth curve model (also known as GMANOVA) is used to analyze data such as this, where multiple observations are made on collections of individuals over time.

APPLICATIONS

Health indicators are commonly used to guide public health policy.

CHARACTERISTICS

A health indicator which will be used internationally to describe global health should have the following characteristics:

- It should be defined in such a way that it can be measured uniformly internationally.
- It must have statistical validity.
- The indicator must be data which can feasibly be collected.
- The analysis of the data must result in a recommendation on which people can make changes to improve health

LIST OF HEALTH INDICATORS

Health indicators are required in order to measure the health status of people and communities.

Mortality Indicators

- Crude death rate
- Life expectancy
- Infant mortality rate
- Maternal mortality rate
- Proportional mortality rate

Morbidity Indicators

- Prevalence
- Incidence
- Others

Health Status

Incidence counts of any of the following in a population may be health indicators:

- Low birth weight
- Obesity
- Arthritis
- Diabetes
- Asthma
- High blood pressure
- Cancer incidence
- Chronic pain
- Oral health
- Depression
- Hospital visits due to injury
- Reports of waterborne diseases or foodborne illness

Disability Indicators

- Disability adjusted life years (DALY)
- Others: Activities of daily living (ADL), Musculoskeletal disability (MSD) score etc.

Nutritional Indicators

- Proportion of low birth weight
- Prevalence of anaemia
- Proportion of overweight individuals
- Nutritional intake assessments

Social and Mental Health Indicators

- Alcohol related indicators
- Injury rates

Health System Indicators

- Health care delivery related
- Health policy indicators

Health Determinant

- Smoking habits
- alcohol consumption habits
- Physical exercise habits
- Breastfeeding

ORGANIZATIONS

Various organizations exist to identify, collect, measure, share, analyze, and publish on the topic of health indicators. Here are some example organizations doing this:

- Health Metrics Network
- Institute for Health Metrics and Evaluation

HEALTH SERVICES RESEARCH

Health services research (HSR), also known as health systems research or health policy and systems research (HPSR), is a multidisciplinary scientific field that examines how people get access to health care practitioners and health care services, how much care costs, and what happens to patients as a result of this care. Studies in HSR investigate how social factors, health policy, financing systems, organizational structures and processes, medical technology, and personal behaviors affect access to health care, the quality and cost of health care, and quantity and quality of life.

Compared with medical research, HSR is a relatively young science that developed through the bringing together of social science perspectives with the contributions of individuals and institutions engaged in delivering health services.

GOALS

The primary goals of health services research are to identify the most effective ways to organize, manage, finance, and deliver high quality care; reduce medical errors; and improve patient safety. HSR is more concerned with delivery and access to care, in contrast to medical research, which focuses on the development and evaluation of clinical treatments. Health services researchers come from a variety of specializations, including geography, nursing, economics, political science, epidemiology, public health, medicine, biostatistics, operations, management, engineering, pharmacy, psychology, usability and user experience design. While health services research is grounded in theory, its underlying aim is to perform research that can be applied by physicians, nurses, health managers and administrators, and other people who make decisions or deliver care in the health care system. For example, the application of epidemiological methods to the study of health services by managers is a type of health services research that can be described as Managerial epidemiology.

APPROACHES

Approaches to HSR include:

- *Implementation research*: research focusing on public policy analysis, or the concerns of programme managers regarding the effectiveness of specific health interventions;
- *Impact evaluation*: research with emphasis on effectiveness of health care practices and organisation of care, using a more narrow range of study methods such as systematic reviews of health system interventions.

BY COUNTRY

Many data and information sources are used to conduct health services research, such as population and health surveys, clinical administrative records, health care programme and financial administrative records, vital statistics records (births and deaths), and other special studies.

United States

Data Availability

Claims data on US Medicare and Medicaid beneficiaries are available for analysis. Data is divided into public data available to any entity and research data available only to qualified researchers. The US's Centers for Medicare and Medicaid Services (CMS) delegates some data export functions to a Research Data Assistance Center.

23 Claims data from various states that are not limited to any particular insurer are also available for analysis via AHRQ's HCUP project.

Centers

Colloquially, health services research departments are often referred to as "shops"; in contrast to basic science research "labs". Broadly, these shops are hosted by three general types of institutions—government, academic, or non-governmental think tanks or professional societies.

Government Sponsored:

- U.S. Department of Veterans Affairs Award in Health Services Research
- Institute of Medicine, U.S.-based policy research organization

University Sponsored:

- Center for Surgery and Public Health, U.S. -based research institute at the Brigham and Women's Hospital (Harvard University Affiliate)
- Regenstrief Institute
- Institute for Health care Policy and Innovation, U.S. -based research institute at the University of Michigan (Founded in 2011, IHPI includes smaller centers focused on specific health care topics)
- Leonard Davis Institute of Health Economics, U.S.-based center for HSR at the University of Pennsylvania

Think Tank or Professional Society Sponsored:

- Society of General Internal Medicine, U.S.-based professional organization in internal medicine research
- Commonwealth Fund, U.S.-based center for HSR
- Rand Corporation Health Division, U.S.-based center for HSR

INJURY PREVENTION

Injury prevention is an effort to prevent or reduce the severity of bodily injuries caused by external mechanisms, such as accidents, before they occur. Injury prevention is a component of safety and public health, and its goal is to improve the health of the population by preventing injuries and hence improving quality of life. Among laypersons, the term "accidental injury" is often used. However, "accidental" implies the causes of injuries are random in nature. Researchers use the term "unintentional injury" to refer to injuries that are nonvolitional but preventable. Within the field of public health, efforts are also made to prevent or reduce "intentional injury." Data from the U.S. Centers for Disease Control, for example, show unintentional injuries are the leading cause of death from early childhood until middle adulthood. During these years, unintentional injuries account for more deaths than the next nine leading causes of death combined. Injury prevention strategies cover a variety of approaches, many of which are classified as falling under the "3 E's" of injury prevention: education, engineering modifications, and enforcement/enactment. Some organizations, such as Safe Kids Worldwide, have expanded the list to six E's adding: evaluation, economic incentives and empowerment.

MEASURING EFFECTIVENESS

Researching is challenging, because the usual outcome of interest is deaths or injuries prevented, and it is nearly impossible to measure how many people *did not* get hurt who otherwise would have. Education efforts can be measured by changes in knowledge, attitudes, beliefs and behaviors, before and after the intervention, however tying these changes back into reductions in morbidity and mortality is often problematic.

Examining trends in morbidity and mortality in the population is usually not difficult and may provide some indication of the effectiveness of injury prevention interventions. However, this approach suffers from the potential of ecological fallacy, where the data shows an association between an intervention and a change in the outcome, but there is actually no causal relationship.

COMMON TYPES

Traffic and Automobile Safety

Traffic safety and automobile safety are a major component of injury prevention because it is the leading cause of death for children and young adults into their mid 30s. Injury prevention efforts began in the early 1960s when activist Ralph Nader, exposed the automobiles as being more dangerous than necessary with his

book Unsafe at Any Speed. This led to engineering changes in the way cars are designed to allow for more crush space between the vehicle and the occupant. The Centers for Disease Control and Prevention (CDC) also contributes much to automobile safety. The CDC Injury Prevention Champion, David Sleet, illustrated the importance of lowering the legal blood alcohol content limit to 0.08 percent for drivers; requiring disposable lighters to be child resistant; and using evidence to demonstrate the dangers of airbags to young children riding in the front seat of vehicles.

Engineering: vehicle crash worthiness, seat belts, airbags, locking seat belts for child seats.

Education: promote seat belt use, discourage impaired driving, promote child safety seats.

Enforcement and enactment: passage and enforcement of primary seat belt laws, speed limits, impaired driving enforcement.

Pedestrian Safety

Pedestrian safety is the focus of both epidemiological and psychological injury prevention research. Epidemiological studies typically focus on causes external to the individual such as traffic density, access to safe walking areas, socioeconomic status, injury rates, legislation for safety (*e.g.*, traffic fines), or even the shape of vehicles which affects the severity of injuries resulting from a collision. Epidemiological data show children aged 1–4 are at greatest risk for injury in driveway and sidewalks. Children aged 5–14 are at greatest risk while attempting to cross streets. The body of psychological research on pedestrian safety is currently much smaller than that in the epidemiological field, but is rapidly growing. Psychological pedestrian safety studies extend as far back as the mid-1980s when researchers began examining behavioral variables in children. Behavioral variables of interest include selection of crossing gaps in traffic, attention to traffic, the number of near hits or actual hits, or the routes children chose when crossing multiple streets such as while walking to school. Behavioral studies often collect such variables which imply risk of injury; *e.g.*, children engaging in risky behaviors may be assumed to be at greater risk if actually crossing a street alone. The most common technique used in behavioral pedestrian research is the pretend road, in which a child stands some distance from the curb and watches traffic on the real road. The child then walks to the edge of the street when a crossing opportunity is chosen. Research is gradually shifting to more ecologically valid virtual reality techniques. Leading scientists in psychological pedestrian safety research are Dr. Benjamin Barton, Dr. David Schwebel and Dr. James Thomson.

Other

The following is an abbreviated topic list of some common focus areas of injury prevention efforts:

- Bicycle safety
- Boat and water safety
- Child passenger safety
- Consumer product safety
- Firearm safety
- Fire and burn safety
- Home safety
- Impaired driving
- Pedestrian safety
- Poison control
- Toy safety
- Traffic safety
- Sports injury safety
- Occupational safety and health

INTERNATIONAL PSYCHOPHARMACOLOGY ALGORITHM PROJECT

The International Psychopharmacology Algorithm Project (IPAP) is a non-profit corporation whose purpose is to "enable, enhance, and propagate" use of algorithms for the treatment of some Axis I psychiatric disorders.

Kenneth O Jobson founded the Project. The Dean Foundation provides funding.

IPAP has organized and supported several international conferences on psychopharmacology algorithms. It has also supported the creation of several algorithms based on expert opinion. It is now in the process of creating "evidence-based algorithms," that is algorithms created by experts and annotated with the evidence that leads to these algorithms. A schizophrenia algorithm has been created and one on Post Traumatic Stress Disorder (PTSD) was released in July 2005. A general anxiety disorder (GAD) algorithm was released in 2006. Periodic updates of the algorithms are released as the basis of evidence changes. In addition, the algorithms are being translated into various non-English languages (Chinese, Japanese, Spanish, and Thai) as the availability of translators permits.

MATCHED MOLECULAR PAIR ANALYSIS

Matched molecular pair analysis (MMPA) is a method in cheminformatics that compares the properties of two molecules that differ only by a single chemical transformation, such as the substitution of a hydrogen atom by a chlorine one.

Such pairs of compounds are known as matched molecular pairs (MMP). Because the structural difference between the two molecules is small, any experimentally observed change in a physical or biological property between the matched molecular pair can more easily be interpreted. The term was first coined by Kenny and Sadowski in the book *Chemoinformatics in Drug Discovery*.

MMP can be defined as a pair of molecules that differ in only a minor single point change. Matched molecular pairs (MMPs) are widely used in medicinal chemistry to study changes in compound properties which includes biological activity, toxicity, environmental hazards and much more, which are associated with well-defined structural modifications. Single point changes in the molecule pairs are termed a chemical transformation or Molecular transformation. Each molecular pair is associated with a particular transformation. An example of transformation is the replacement of one functional group by another. More specifically, molecular transformation can be defined as the replacement of a molecular fragment having one, two or three attachment points with another fragment. Useful Molecular transformation in a specified context is termed as "Significant" transformations. For example, a transformation may systematically decrease or increase a desired property of chemical compounds. Transformations that affect a particular property/activity in a statistically significant sense are called as significant transformations. The transformation is considered significant, if it increases the property value "more often" than it decreases it or vice versa. Thus, the distribution of increasing and decreasing pairs should be significantly different from the binomial ("no effect") distribution with a particular p-value (usually 0.05).

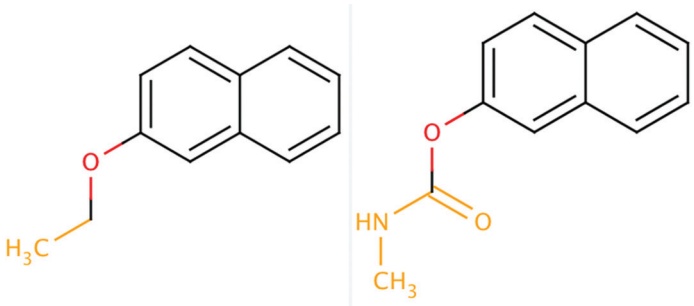


Fig. Examplary MMPs (differences highlighted in orange).

SIGNIFICANCE OF MMP BASED ANALYSIS

MMP based analysis is an attractive method for computational analysis because they can be algorithmically generated and they make it possible to associate defined structural modifications at the level of compound pairs with chemical property changes, including biological activity.

Interpretable QSAR Models

MMPA is quite useful in the field of quantitative structure–activity relationship (QSAR) modelling studies. One of the issues of QSAR models is they are difficult to interpret in a chemically meaningful manner. While it can be pretty easy to interpret simple linear regression models, the most powerful algorithms like neural networks, support vector machine are similar to "black boxes", which provide predictions that can't be easily interpreted. This problem undermines the applicability of QSAR model in helping the medicinal chemist to make the decision.

If the compound is predicted to be active against some microorganism, what are the driving factors of its activity? Or if it is predicted to be inactive, how its activity can be modulated? The black box nature of the QSAR model prevents it from addressing these crucial issues. The use of predicted MMPs allows to interpret models and identify which MMPs were learned by the model. The MMPs, which were not reproduced by the model, could correspond to experimental errors or deficiency of the model (inappropriate descriptors, too few data, etc.).

Analysis of MMPs (matched molecular pair) can be very useful for understanding the mechanism of action. A medicinal chemist might be interested particularly in "activity cliff". Activity cliff is a minor structural modification, which changes the target activity significantly.

Activity Cliff

Activity cliffs are minor structural modifications, with significant effect on molecular property. Activity cliffs usually have high SAR information content. Because small chemical changes in the set of similar compounds lead to large changes in activity. The assessment of activity cliffs requires careful consideration of similarity and potency difference criteria

TYPES OF MMP BASED ANALYSIS

Matched molecular pair (MMPA) analyses can be classified into two types: supervised and unsupervised MMPA.

Supervised MMPA

In supervised MMPA, the chemical transformations are predefined, then the corresponding matched pair compounds are found within the data set and the change in end point computed for each transformation.

Unsupervised MMPA

Also known as automated MMPAs. A machine learning algorithm is used to finds all possible matched pairs in a data set according to a set of predefined rules. This results in much larger numbers of matched pairs and unique transformations, which are typically filtered during the process to identify those transformations that correspond to statistically significant changes in the targeted property with a reasonable number of matched pairs.

MATCHED MOLECULAR SERIES

Here instead of looking at the pair of molecules which differ only at one point, a series of more than 2 molecules different at a single point is considered. The concept of matching molecular series was introduced by Wawer and Bajorath. It is argued that longer matched series is more likely to exhibit preferred molecular transformation while, matched pairs exhibit only a small preference.

LIMITATIONS

The application of the MMPA across large chemical databases for the optimization of ligand potency is problematic because same structural transformation may increase or decrease or doesn't affect the potency of different compounds in the dataset. Selection of practical significant transformation from a dataset of molecules is a challenging issue in the MMPA. Moreover, the effect of a particular molecular transformation can significantly depend on the Chemical context of transformations. Beside these, MMPA might pose some limitations in terms of computational resources, especially when dealing with databases of compounds with a large number of breakable bonds. Further, more atoms in the variable part of the molecule also leads to combinatorial explosion problems. To deal with this, the number of breakable bonds and number of atoms in the variable part can be used to pre-filter the database.

MEDCALC

MedCalc is a statistical software package designed for the biomedical sciences. It has an integrated spreadsheet for data input and can import files in several formats (Excel, SPSS, CSV,...).

MedCalc includes basic parametric and non-parametric statistical procedures and graphs such as descriptive statistics, ANOVA, Mann Whitney test, Wilcoxon test, χ test, correlation, linear as well as non-linear regression, logistic regression, etc.

Survival analysis includes Cox regression (Proportional hazards model) and Kaplan–Meier survival analysis.

Procedures for method evaluation and method comparison include ROC curve analysis, Bland–Altman plot, as well as Deming and Passing–Bablok regression.

The software also includes meta-analysis and sample size calculations.

The first DOS version of MedCalc was released in April 1993 and the first version for Windows was available in November 1996. On 7 March 2007, version 9.3 obtained the Certified for Windows Vista logo.

Version 15.2 introduced a user-interface in English, Chinese (simplified and traditional), French, German, Italian, Japanese, Korean, Polish, Portuguese (Brazilian), Russian and Spanish.

MEDIAN FOLLOW-UP

In statistics, median follow-up is the median time between a specified event and the time when data on outcomes are gathered. The concept is used in cancer survival analyses. Many cancer studies aim to assess the time between two events of interest, such as from treatment to remission, treatment to relapse, or diagnosis to death. This duration is generically called survival time, even if the end point is not death.

Time-to-event studies must have sufficiently long follow-up durations to capture enough events to reveal meaningful patterns in the data. A short follow-up duration is appropriate for studying very severe cancers with poor prognoses, whereas a long follow-up duration is better suited to studying less-severe disease, or participants with good prognoses.

Median follow-up time is included in about half the survival analyses published in cancer journals, but of those, only 31 per cent specify the method used to compute it.

MEDICAL STATISTICS

Medical statistics deals with applications of statistics to medicine and the health sciences, including epidemiology, public health, forensic medicine, and clinical research. Medical statistics has been a recognized branch of statistics in the United Kingdom for more than 40 years but the term has not come into general use in North America, where the wider term 'biostatistics' is more commonly used.

However, "biostatistics" more commonly connotes all applications of statistics to biology. Medical statistics is a subdiscipline of statistics. "It is the science of summarizing, collecting, presenting and interpreting data in medical practice, and using them to estimate the magnitude of associations and test hypotheses. It has a central role in medical investigations.

It not only provides a way of organizing information on a wider and more formal basis than relying on the exchange of anecdotes and personal experience, but also takes into account the intrinsic variation inherent in most biological processes."

PHARMACEUTICAL STATISTICS

Pharmaceutical statistics is the application of statistics to matters concerning the pharmaceutical industry. This can be from issues of design of experiments, to analysis of drug trials, to issues of commercialization of a medicine.

There are many professional bodies concerned with this field including:

- European Federation of Statisticians in the Pharmaceutical Industry (EFSPI)
- Statisticians In The Pharmaceutical Industry (PSI)

There are also journals including:

- *Statistics in Medicine*
- *Pharmaceutical Statistics*

BASIC CONCEPTS

For Describing Situations

- Incidence (epidemiology) vs. Prevalence vs. Cumulative incidence
- All types of medical test results can be true positive or false positive and true negative or false negative...these categories are hugely dependent on the prevalence of the disease being tested for; and, for example, a false positive test result is more likely when the prevalence of the disease being tested for is very low (such as a lab test for cold-weather influenza being done in June in South Carolina, USA. and giving a "positive" result).
- Transmission rate vs. force of infection
- Mortality rate vs. standardized mortality ratio vs. age-standardized mortality rate
- Pandemic vs. epidemic vs. endemic vs. syndemic
- Serial interval vs. incubation period
- Cancer cluster
- Sexual network
- Years of potential life lost
- Maternal mortality rate
- Perinatal mortality rate
- Low birth weight ratio

For Assessing the Effectiveness of an Intervention

- Absolute risk reduction
- Control event rate
- Experimental event rate
- Number needed to harm
- Number needed to treat
- Odds ratio
- Relative risk reduction
- Relative risk
- Relative survival
- Minimal clinically important difference

RELATED STATISTICAL THEORY

- Survival analysis
- Proportional hazards models
- Active control trials: clinical trials in which a kind of new treatment is compared with some other active agent rather than a placebo.
- ADLS(Activities of daily living scale): a scale designed to measure physical ability/disability that is used in investigations of a variety of chronic disabling conditions, such as arthritis. This scale is based on scoring responses to questions about self-care, grooming, etc.
- Actuarial statistics: the statistics used by actuaries to calculate liabilities, evaluate risks and plan the financial course of insurance, pensions, etc.

MINIMUM VIABLE POPULATION

Minimum viable population (MVP) is a lower bound on the population of a species, such that it can survive in the wild. This term is used in the fields of biology, ecology, and conservation biology. More specifically, MVP is the smallest possible size at which a biological population can exist without facing extinction from natural disasters or demographic, environmental, or genetic stochasticity. The term "population" rarely refers to an entire species. For example, the undomesticated dromedary camel is extinct in its natural wild habitat; however, there is a domestic population in captivity and an additional feral population in Australia. Two groups of house cats in separate houses which are not allowed outdoors are also technically distinct populations. Typically, however, MVP is used to refer solely to a wild population, such as the red wolf.

ESTIMATION

Minimum viable population is usually estimated as the population size necessary to ensure between 90 and 95 percent probability of survival between 100 and 1,000 years into the future. The MVP can be estimated using computer simulations for population viability analyses (PVA). PVA models populations using demographic and environmental information to project future population dynamics. The probability assigned to a PVA is arrived at after repeating the environmental simulation thousands of times. For example, for a theoretical simulation of a population of 50 giant pandas in which the simulated population goes completely extinct, 30 out of 100 stochastic simulations projected 100 years into the future are not viable. Causes of extinction in the simulation may include inbreeding depression, natural disaster, or climate change. Extinction occurring in 30 out of 100 runs would give a survival probability of 70 per cent. In contrast, in the same simulation with a starting population of 60 pandas, the panda population may only become extinct in four of the hundred runs,

resulting in a survival probability of 96 per cent. In this case the minimum viable population that satisfies the 90- to 95 per cent probability for survival is between 50 and 60 pandas. (These figures have been invented for the purpose of this example.)

EXTINCTION

MVP does not take external intervention into account. Thus, it is useful for conservation managers and environmentalists; a population may be increased above the MVP using a captive breeding programme, or by bringing other members of the species in from other reserves. There is naturally some debate on the accuracy of PVAs, since a wide variety of assumptions generally are required for future forecasting; however, the important consideration is not absolute accuracy, but promulgation of the concept that each species indeed has an MVP, which at least can be approximated for the sake of conservation biology and Biodiversity Action Plans. There is a marked trend for insularity, surviving genetic bottlenecks and r-strategy to allow far lower MVPs than average. Conversely, taxa easily affected by inbreeding depression – having high MVPs – are often decidedly K-strategists, with low population densities while occurring over a wide range. An MVP of 500 to 1,000 has often been given as an average for terrestrial vertebrates when inbreeding or genetic variability is ignored. When inbreeding effects are included, estimates of MVP for many species are in the thousands. Based on a meta-analysis of reported values in the literature for many species, Traill *et al.* reported a median MVP of 4,169 individuals.

POPULATION UNCERTAINTY

Population uncertainty may be divided into four sources:

- Demographic stochasticity
- Environmental stochasticity
- Natural catastrophes
- Genetic stochasticity

MOST PROBABLE NUMBER

The most probable number method, otherwise known as the method of Poisson zeroes, is a method of getting quantitative data on concentrations of discrete items from positive/negative (incidence) data. There are many discrete entities that are easily detected but difficult to count. Any sort of amplification reaction or catalysis reaction obliterates easy quantification but allows presence to be detected very sensitively. Common examples include microorganism growth, enzyme action, or catalytic chemistry. The MPN method involves taking the original solution or sample, and subdividing it by orders of magnitude (frequently 10× or 2×), and assessing presence/absence in multiple subdivisions. The degree of dilution at which absence begins to appear indicates that the items have been diluted so much that there are many subsamples in which none appear. A suite of replicates at any given concentration allow finer resolution, to use the number of positive and negative samples to estimate the original concentration within the appropriate order of magnitude.

In microbiology, the cultures are incubated and assessed by eye, bypassing tedious colony counting or expensive and tedious microscopic counts. Presumptive, Confirmative and Completed tests are a part of MPN In molecular biology, a common application involves DNA templates diluted into polymerase chain reactions (PCR). Reactions only proceed when a template is present, allowing for a form of quantitative PCR, to assess the original concentration of template molecules. Another application involves diluting enzyme stocks into solution containing a chromogenic substrate, or diluting antigens into solutions for ELISA (Enzyme-Linked ImmunoSorbent Assay) or some other antibody cascade detection reaction, to measure the original concentration of the enzyme or antigen.

The major weakness of MPN methods is the need for large numbers of replicates at the appropriate dilution to narrow the confidence intervals. However, it is a very important method for counts when the appropriate order of magnitude is unknown *a priori* and sampling is necessarily destructive.

OPENEPI

OpenEpi is a free, web-based, open source, operating system-independent series of programmes for use in epidemiology, biostatistics, public health, and medicine, providing a number of epidemiologic and statistical tools for summary data. OpenEpi was developed in JavaScript and HTML, and can be run in modern web browsers. The programme can be run from the OpenEpi web site or downloaded and run without a web connection. The source code and documentation is downloadable and freely available for use by other investigators. OpenEpi has been reviewed, both by media organizations and in research journals. The OpenEpi developers have had extensive experience in the development and testing of Epi Info, a programme developed by the Centers for Disease Control and Prevention (CDC) and widely used around the world for data entry and analysis. OpenEpi was developed to perform analyses found in the DOS version of Epi Info modules StatCalc and EpiTable, to improve upon the types of analyses provided by these modules, and to provide a number of tools and calculations not currently available in Epi Info. It is the first step towards an entirely web-based set of epidemiologic software tools. OpenEpi can be thought of as an important companion to Epi Info and to other programmes such as SAS, PSPP, SPSS, Stata, SYSTAT, Minitab, Epidata, and R.

Another functionally similar Windows-based programme is Winpepi. Both OpenEpi and Epi Info were developed with the goal of providing tools for low and moderate resource areas of the world. The initial development of OpenEpi was supported by a grant from the Bill and Melinda Gates Foundation to Emory University.

The types of calculations currently performed by OpenEpi include:

- Various confidence intervals for proportions, rates, standardized mortality ratio, mean, median, percentiles
- 2x2 crude and stratified tables for count and rate data
- Matched case-control analysis
- Test for trend with count data
- Independent t-test and one-way ANOVA
- Diagnostic and screening test analyses with receiver operating characteristic (ROC) curves
- Sample size for proportions, cross-sectional surveys, unmatched case-control, cohort, randomized controlled trials, and comparison of two means
- Power calculations for proportions (unmatched case-control, cross-sectional, cohort, randomized controlled trials) and for the comparison of two means
- Random number generator

For epidemiologists and other health researchers, OpenEpi performs a number of calculations based on tables not found in most epidemiologic and statistical packages. For example, for a single 2x2 table, in addition to the results presented in other programmes, OpenEpi provides estimates for:

- Etiologic or prevented fraction in the population and in exposed with confidence intervals, based on risk, odds, or rate data
- The cross-product and MLE odds ratio estimate
- Mid-p exact p-values and confidence limits for the odds ratio
- Calculations of rate ratios and rate differences with confidence intervals and statistical tests.

For stratified 2x2 tables with count data, OpenEpi provides:

- Mantel-Haenszel (MH) and precision-based estimates of the risk ratio and odds ratio
- Precision-based adjusted risk difference

- Tests for interaction for the risk ratio, odds ratio, and risk difference
- Four different confidence limit methods for the odds ratio.

Similar to Epi Info, in a stratified analysis, both crude and adjusted estimates are provided so that the assessment of confounding can be made. With rate data, OpenEpi provides adjusted rate ratio's and rate differences, and tests for interaction. Finally, with count data, OpenEpi also performs a test for trend, for both crude data and stratified data. In addition to being used to analyze data by health researchers, OpenEpi has been used as a training tool for teaching epidemiology to students at: Emory University, University of Massachusetts, University of Michigan, University of Minnesota, Morehouse College, Columbia University, University of Wisconsin, San Jose State University, University of Medicine and Dentistry of New Jersey, University of Washington, and elsewhere. This includes campus-based and distance learning courses. Because OpenEpi is easy to use, requires no programming experience, and can be run on the internet, students can use the programme and focus on the interpretation of results. Users can run the programme in English, French, Spanish, Portuguese or Italian. Comments and suggestions for improvements are welcomed and the developers respond to user queries. The developers encourage others to develop modules that could be added to OpenEpi and provide a developer's tool at the web site. Planned future development include improvements to existing modules, development of new modules, translation into other languages, and add the ability to cut and paste data and/or read data files.

POPULATION VIABILITY ANALYSIS

Population viability analysis (PVA) is a species-specific method of risk assessment frequently used in conservation biology. It is traditionally defined as the process that determines the probability that a population will go extinct within a given number of years. More recently, PVA has been described as a marriage of ecology and statistics that brings together species characteristics and environmental variability to forecast population health and extinction risk. Each PVA is individually developed for a target population or species, and consequently, each PVA is unique. The larger goal in mind when conducting a PVA is to ensure that the population of a species is self-sustaining over the long term.

USES

Population viability analysis (PVA) is used to estimate the likelihood of a population's extinction and indicate the urgency of recovery efforts, and identify key life stages or processes that should be the focus of recovery efforts. PVA is also used to compare proposed management options and assess existing recovery efforts. PVA is frequently used in endangered species management to develop a plan of action, rank the pros and cons of different management scenarios, and assess the potential impacts of habitat loss.

HISTORY

In the 1970s, Yellowstone National Park was the centre of a heated debate over different proposals to manage the park's problem grizzly bears (*Ursus arctos*). In 1978, Mark Shaffer proposed a model for the grizzlies that incorporated random variability, and calculated extinction probabilities and minimum viable population size. The first PVA is credited to Shaffer.

PVA gained popularity in the United States as federal agencies and ecologists required methods to evaluate the risk of extinction and possible outcomes of management decisions, particularly in accordance with the Endangered Species Act of 1973, and the National Forest Management Act of 1976. In 1986, Gilpin and Soulé broadened the PVA definition to include the interactive forces that affect the viability of a population, including genetics. The use of PVA increased dramatically in the late 1980s and early 1990s following advances in personal computers and software packages.

EXAMPLES

A PVA for the endangered Fender's blue butterfly (*Icaricia icarioides*) was recently performed with a goal of providing additional information to the United States Fish and Wildlife Service, which was developing a recovery plan for the species. The PVA concluded that the species was more at risk of extinction than previously thought and identified key sites where recovery efforts should be focused. The PVA also indicated that because the butterfly populations fluctuate widely from year to year, to prevent the populations from going extinct the minimum annual population growth rate must be kept much higher than at levels typically considered acceptable for other species. Following a recent outbreak of canine distemper virus, a PVA was performed for the critically endangered island fox (*Urocyon littoralis*) of Santa Catalina Island, California. The Santa Catalina island fox population is uniquely composed of two subpopulations that are separated by an isthmus, with the eastern subpopulation at greater risk of extinction than the western subpopulation. PVA was conducted with the goals of 1) evaluating the island fox's extinction risk, 2) estimating the island fox's sensitivity to catastrophic events, and 3) evaluating recent recovery efforts which include release of captive-bred foxes and transport of wild juvenile foxes from the west to the east side. Results of the PVA concluded that the island fox is still at significant risk of extinction, and is highly susceptible to catastrophes that occur more than once every 20 years. Furthermore, extinction risks and future population sizes on both sides of the island were significantly dependent on the number of foxes released and transported each year.

PVAs in combination with sensitivity analysis can also be used to identify which vital rates has the relative greatest effect on population growth and other measures of population viability. For example, a study by Manlik *et al.* (2016) forecast the viability of two bottlenose dolphin populations in Western Australia and identified reproduction as having the greatest influence on the forecast of these populations. One of the two populations was forecast to be stable, whereas the other population was forecast to decline, if it isolated from other populations and low reproductive rates persist. The difference in viability between the two studies was primarily due to differences in reproduction and not survival. The study also showed that temporal variation in reproduction had a greater effect on population growth than temporal variation in survival.

CONTROVERSY

Debates exist and remain unresolved over the appropriate uses of PVA in conservation biology and PVA's ability to accurately assess extinction risks. A large quantity of field data is desirable for PVA; some conservatively estimate that for a precise extinction probability assessment extending T years into the future, five-to-ten times T years of data are needed. Datasets of such magnitude are typically unavailable for rare species; it has been estimated that suitable data for PVA is available for only 2 per cent of threatened bird species. PVA for threatened and endangered species is particularly a problem as the predictive power of PVA plummets dramatically with minimal datasets. Ellner et al. (2002) argued that PVA has little value in such circumstances and is best replaced by other methods. Others argue that PVA remains the best tool available for estimations of extinction risk, especially with the use of sensitivity model runs.

Even with an adequate dataset, it is possible that a PVA can still have large errors in extinction rate predictions. It is impossible to incorporate all future possibilities into a PVA: habitats may change, catastrophes may occur, new diseases may be introduced. PVA utility can be enhanced by multiple model runs with varying sets of assumptions including the forecast future date. Some prefer to use PVA always in a relative analysis of benefits of alternative management schemes, such as comparing proposed resource management plans. Accuracy of PVAs has been tested in a few retrospective studies. For example, a study comparing PVA model forecasts with the actual fate of 21 well-studied taxa, showed that growth rate projections are accurate, if input variables are based on sound data, but highlighted the importance of understanding density-dependence (Brook *et al.*

2000). Also, McCarthey *et al.* (2003) showed that PVA predictions are relatively accurate, when they are based on long-term data. Still, the usefulness of PVA lies more in its capacity to identify and assess potential threats, than in making long-term, categorical predictions (Akçakaya and Sjögren-Gulve 2000).

FUTURE DIRECTIONS

Improvements to PVA likely to occur in the near future include: 1) creating a fixed definition of PVA and scientific standards of quality by which all PVA are judged and 2) incorporating recent genetic advances into PVA.

POSITIVE AND NEGATIVE PREDICTIVE VALUES

The positive and negative predictive values (PPV and NPV respectively) are the proportions of positive and negative results in statistics and diagnostic tests that are true positive and true negative results, respectively. The PPV and NPV describe the performance of a diagnostic test or other statistical measure. A high result can be interpreted as indicating the accuracy of such a statistic. The PPV and NPV are not intrinsic to the test; they depend also on the prevalence. The PPV can be derived using Bayes' theorem. Although sometimes used synonymously, a *positive predictive value* generally refers to what is established by control groups, while a post-test probability refers to a probability for an individual. Still, if the individual's pre-test probability of the target condition is the same as the prevalence in the control group used to establish the positive predictive value, the two are numerically equal. In information retrieval, the PPV statistic is often called the precision.

RELATIONSHIP

Although sometimes used synonymously, a *negative predictive value* generally refers to what is established by control groups, while a negative post-test probability rather refers to a probability for an individual. Still, if the individual's pre-test probability of the target condition is the same as the prevalence in the control group used to establish the negative predictive value, then the two are numerically equal.

Table. The following diagram illustrates how the *positive predictive value*, negative predictive value, sensitivity, and specificity are related.

		True condition				
	Total population	Condition positive	Condition negative	Prevalence = $\frac{\Sigma \text{ Condition positive}}{\Sigma \text{ Total population}}$	Accuracy (ACC) = $\frac{\Sigma \text{ True positive} + \Sigma \text{ True negative}}{\Sigma \text{ Total population}}$	
Predicted condition	Predicted condition positive	**True positive**, Power	**False positive**, Type I error	Positive predictive value (PPV), Precision = $\frac{\Sigma \text{ True positive}}{\Sigma \text{ Predicted condition positive}}$	False discovery rate (FDR) = $\frac{\Sigma \text{ False positive}}{\Sigma \text{ Predicted condition positive}}$	
	Predicted condition negative	**False negative**, Type II error	**True negative**	False omission rate (FOR) = $\frac{\Sigma \text{ False negative}}{\Sigma \text{ Predicted condition negative}}$	Negative predictive value (NPV) = $\frac{\Sigma \text{ True negative}}{\Sigma \text{ Predicted condition negative}}$	
		True positive rate (TPR), Recall, Sensitivity, probability of detection = $\frac{\Sigma \text{ True positive}}{\Sigma \text{ Condition positive}}$	False positive rate (FPR), Fall-out, probability of false alarm = $\frac{\Sigma \text{ False positive}}{\Sigma \text{ Condition negative}}$	Positive likelihood ratio (LR+) = $\frac{\text{TPR}}{\text{FPR}}$	Diagnostic odds ratio (DOR) = $\frac{\text{LR+}}{\text{LR−}}$	F_1 score = $\frac{2}{\frac{1}{\text{Recall}} + \frac{1}{\text{Precision}}}$
		False negative rate (FNR), Miss rate = $\frac{\Sigma \text{ False negative}}{\Sigma \text{ Condition positive}}$	Specificity (SPC), Selectivity, True negative rate (TNR) = $\frac{\Sigma \text{ True negative}}{\Sigma \text{ Condition negative}}$	Negative likelihood ratio (LR−) = $\frac{\text{FNR}}{\text{TNR}}$		

Note that the positive and negative predictive values can only be estimated using data from a cross-sectional study or other population-based study in which valid prevalence estimates may be obtained. In contrast, the sensitivity and specificity can be estimated from case-control studies.

WORKED EXAMPLE

Table. Suppose the fecal occult blood (FOB) screen test is used in 2030 people to look for bowel cancer.

		Patients with bowel cancer (as confirmed on endoscopy)		
		Condition positive	**Condition negative**	
Fecaloccult blood screen test outcome	Test outcome positive	True positive (TP) = 20	False positive (FP) = 180	Positive predictive value = TP / (TP + FP) = 20 / (20 + 180) = 10%
	Test outcome negative	False negative (FN) = 10	True negative (TN) = 1820	Negative predictive value = TN / (FN + TN) = 1820 / (10 + 1820) ≈ 99.5%
		Sensitivity = TP / (TP + FN) = 20 / (20 + 10) ≈ 67%	Specificity = TN / (FP + TN) = 1820 / (180 + 1820) = 91%	

The small positive predictive value (PPV = 10%) indicates that many of the positive results from this testing procedure are false positives. Thus it will be necessary to follow up any positive result with a more reliable test to obtain a more accurate assessment as to whether cancer is present. Nevertheless, such a test may be useful if it is inexpensive and convenient. The strength of the FOB screen test is instead in its negative predictive value — which, if negative for an individual, gives us a high confidence that its negative result is true.

PROBLEMS

Other Individual Factors

Note that the PPV is not intrinsic to the test—it depends also on the prevalence. Due to the large effect of prevalence upon predictive values, a standardized approach has been proposed, where the PPV is normalized to a prevalence of 50 per cent. PPV is directly proportional to the prevalence of the disease or condition. In the above example, if the group of people tested had included a higher proportion of people with bowel cancer, then the PPV would probably come out higher and the NPV lower. If everybody in the group had bowel cancer, the PPV would be 100 per cent and the NPV 0 per cent. To overcome this problem, NPV and PPV should only be used if the ratio of the number of patients in the disease group and the number of patients in the healthy control group used to establish the NPV and PPV is equivalent to the prevalence of the diseases in the studied population, or, in case two disease groups are compared, if the ratio of the number of patients in disease group 1 and the number of patients in disease group 2 is equivalent to the ratio of the prevalences of the two diseases studied. Otherwise, positive and negative likelihood ratios are more accurate than NPV and PPV, because likelihood ratios do not depend on prevalence.

When an individual being tested has a different pre-test probability of having a condition than the control groups used to establish the PPV and NPV, the PPV and NPV are generally distinguished from the positive and negative post-test probabilities, with the PPV and NPV referring to the ones established by the control groups, and the post-test probabilities referring to the ones for the tested individual (as estimated, for example, by likelihood ratios). Preferably, in such cases, a large group of equivalent individuals should be studied, in order to establish separate positive and negative predictive values for use of the test in such individuals.

Different Target Conditions

PPV is used to indicate the probability that in case of a positive test, that the patient really has the specified disease. However, there may be more than one cause for a disease and any single potential cause may not always result in the overt disease seen in a patient. There is potential to mix up related target conditions of PPV and NPV, such as interpreting the PPV or NPV of a test as having a disease, when that PPV or NPV value actually refers only to a predisposition of having that disease. An example is the microbiological throat swab used in patients with a sore throat. Usually publications stating PPV of a throat swab are reporting on the probability that this bacterium is present in the throat, rather than that the patient is ill from the bacteria found. If presence of this bacterium always resulted in a sore throat, then the PPV would be very useful. However the bacteria may colonise individuals in a harmless way and never result in infection or disease. Sore throats occurring in these individuals are caused by other agents such as a virus. In this situation the gold standard used in the evaluation study represents only the presence of bacteria (that might be harmless) but not a causal bacterial sore throat illness. It can be proven that this problem will affect positive predictive value far more than negative predictive value. To evaluate diagnostic tests where the gold standard looks only at potential causes of disease, one may use an extension of the predictive value termed the Etiologic Predictive Value.

RATE RATIO

A rate ratio (sometimes called an incidence density ratio) in epidemiology, is a relative difference measure used to compare the incidence rates of events occurring at any given point in time. A common application for this measure in analytic epidemiologic studies is in the search for a causal association between a certain risk factor and an outcome.

$$\text{Rate Ratio} = \frac{\text{Incidence Rate}1}{\text{Incidence Rate}2}$$

Where incidence rate is the occurrence of an event over person-time, for example person-years.

$$\text{Incidence Rate} = \frac{\text{events}}{\text{person time}}$$

Note: the same time intervals must be used for both incidence rates.

RECEIVER OPERATING CHARACTERISTIC

A receiver operating characteristic curve, *i.e.*, ROC curve, is a graphical plot that illustrates the diagnostic ability of a binary classifier system as its discrimination threshold is varied. The ROC curve is created by plotting the true positive rate (TPR) against the false positive rate (FPR) at various threshold settings. The true-positive rate is also known as sensitivity, recall or *probability of detection* in machine learning. The false-positive rate is also known as the fall-out or *probability of false alarm* and can be calculated as (1 - specificity). It can also be thought of as a plot of the Power as a function of the Type I Error of the decision rule (when the performance is calculated from just a sample of the population, it can be thought of as estimators of these quantities). The ROC

curve is thus the sensitivity as a function of fall-out. In general, if the probability distributions for both detection and false alarm are known, the ROC curve can be generated by plotting the cumulative distribution function (area under the probability distribution from $-\infty$ to the discrimination threshold) of the detection probability in the y-axis versus the cumulative distribution function of the false-alarm probability on the x-axis.

ROC analysis provides tools to select possibly optimal models and to discard suboptimal ones independently from (and prior to specifying) the cost context or the class distribution. ROC analysis is related in a direct and natural way to cost/benefit analysis of diagnostic decision making. The ROC curve was first developed by electrical engineers and radar engineers during World War II for detecting enemy objects in battlefields and was soon introduced to psychology to account for perceptual detection of stimuli. ROC analysis since then has been used in medicine, radiology, biometrics, forecasting of natural hazards, meteorology, model performance assessment, and other areas for many decades and is increasingly used in machine learning and data mining research. The ROC is also known as a relative operating characteristic curve, because it is a comparison of two operating characteristics (TPR and FPR) as the criterion changes.

BASIC CONCEPT

A classification model (classifier or diagnosis) is a mapping of instances between certain classes/groups. The classifier or diagnosis result can be a real value (continuous output), in which case the classifier boundary between classes must be determined by a threshold value (for instance, to determine whether a person has hypertension based on a blood pressure measure). Or it can be a discrete class label, indicating one of the classes. Let us consider a two-class prediction problem (binary classification), in which the outcomes are labeled either as positive (p) or negative (n). There are four possible outcomes from a binary classifier. If the outcome from a prediction is p and the actual value is also p, then it is called a *true positive* (TP); however if the actual value is n then it is said to be a *false positive* (FP). Conversely, a *true negative* (TN) has occurred when both the prediction outcome and the actual value are n, and *false negative* (FN) is when the prediction outcome is n while the actual value is p. To get an appropriate example in a real-world problem, consider a diagnostic test that seeks to determine whether a person has a certain disease. A false positive in this case occurs when the person tests positive, but does not actually have the disease. A false negative, on the other hand, occurs when the person tests negative, suggesting they are healthy, when they actually do have the disease.

ROC SPACE

The contingency table can derive several evaluation "metrics". To draw an ROC curve, only the true positive rate (TPR) and false positive rate (FPR) are needed (as functions of some classifier parameter). The TPR defines how many correct positive results occur among all positive samples available during the test. FPR, on the other hand, defines how many incorrect positive results occur among all negative samples available during the test. An ROC space is defined by FPR and TPR as x and y axes, respectively, which depicts relative trade-offs between true positive (benefits) and false positive (costs). Since TPR is equivalent to sensitivity and FPR is equal to 1 " specificity, the ROC graph is sometimes called the sensitivity vs (1 " specificity) plot. Each prediction result or instance of a confusion matrix represents one point in the ROC space.

The best possible prediction method would yield a point in the upper left corner or coordinate (0,1) of the ROC space, representing 100 per cent sensitivity (no false negatives) and 100 per cent specificity (no false positives). The (0,1) point is also called a *perfect classification*. A random guess would give a point along a diagonal line (the so-called *line of no-discrimination*) from the left bottom to the top right corners (regardless of the positive and negative base rates). An intuitive example of random guessing is a decision by flipping coins. As the size of the sample increases, a random classifier's ROC point tends towards the diagonal line. In the case of a balanced coin, it will tend to the point (0.5, 0.5).

The diagonal divides the ROC space. Points above the diagonal represent good classification results (better than random); points below the line represent bad results (worse than random). Note that the output of a consistently bad predictor could simply be inverted to obtain a good predictor.

Table. Let us look into four prediction results from 100 positive and 100 negative instances.

A	B	C	C′
TP= 63 FP= 28 91 FN= 37 TN= 72 109 100 100 200	TP= 77 FP= 77 154 FN= 23 TN= 23 46 100 100 200	TP= 24 FP= 88 112 FN= 76 TN= 12 88 100 100 200	TP= 76 FP= 12 88 FN= 24 TN= 88 112 100 100 200
TPR = 0.63	TPR = 0.77	TPR = 0.24	TPR = 0.76
FPR = 0.28	FPR = 0.77	FPR = 0.88	FPR = 0.12
PPV = 0.69	PPV = 0.50	PPV = 0.21	PPV = 0.86
F1 = 0.66	F1 = 0.61	F1 = 0.23	F1 = 0.81
ACC = 0.68	ACC = 0.50	ACC = 0.18	ACC = 0.82

Plots of the four results above in the ROC space are given in the figure. The result of method A clearly shows the best predictive power among A, B, and C. The result of B lies on the random guess line (the diagonal line), and it can be seen in the table that the accuracy of B is 50 per cent. However, when C is mirrored across the center point (0.5,0.5), the resulting method C2 is even better than A. This mirrored method simply reverses the predictions of whatever method or test produced the C contingency table. Although the original C method has negative predictive power, simply reversing its decisions leads to a new predictive method C2 which has positive predictive power. When the C method predicts p or n, the C2 method would predict n or p, respectively. In this manner, the C2 test would perform the best. The closer a result from a contingency table is to the upper left corner, the better it predicts, but the distance from the random guess line in either direction is the best indicator of how much predictive power a method has. If the result is below the line (*i.e.* the method is worse than a random guess), all of the method's predictions must be reversed in order to utilize its power, thereby moving the result above the random guess line.

FURTHER INTERPRETATIONS

Sometimes, the ROC is used to generate a summary statistic. Common versions are:

- The intercept of the ROC curve with the line at 45 degrees orthogonal to the no-discrimination line - the balance point where Sensitivity = Specificity
- The intercept of the ROC curve with the tangent at 45 degrees parallel to the no-discrimination line that is closest to the error-free point (0,1) - also called Youden's J statistic and generalized as Informedness
- The area between the ROC curve and the no-discrimination line multiplied by two - Gini Coefficient
- The area between the full ROC curve and the triangular ROC curve including only (0,0), (1,1) and one selected operating point (tpr,fpr) - Consistency
- The area under the ROC curve, or "AUC" ("Area Under Curve"), or A' (pronounced "a-prime"), or "c-statistic".
- The sensitivity index *d'* (pronounced "d-prime"), the distance between the mean of the distribution of activity in the system under noise-alone conditions and its distribution under signal-alone conditions,

divided by their standard deviation, under the assumption that both these distributions are normal with the same standard deviation. Under these assumptions, the shape of the ROC is entirely determined by *d'*.

However, any attempt to summarize the ROC curve into a single number loses information about the pattern of tradeoffs of the particular discriminator algorithm.

Other Measures

The Total Operating Characteristic (TOC) also characterizes diagnostic ability while revealing more information than the ROC. For each threshold, ROC reveals two ratios, TP/(TP + FN) and FP/(FP + TN). In other words, ROC reveals hits/(hits + misses)and false alarms/(false alarms + correct rejections). On the other hand, TOC shows the total information in the contingency table for each threshold. The TOC method reveals all of the information that the ROC method provides, plus additional important information that ROC does not reveal, *i.e.* the size of every entry in the contingency table for each threshold. TOC also provides the popular AUC of the ROC.

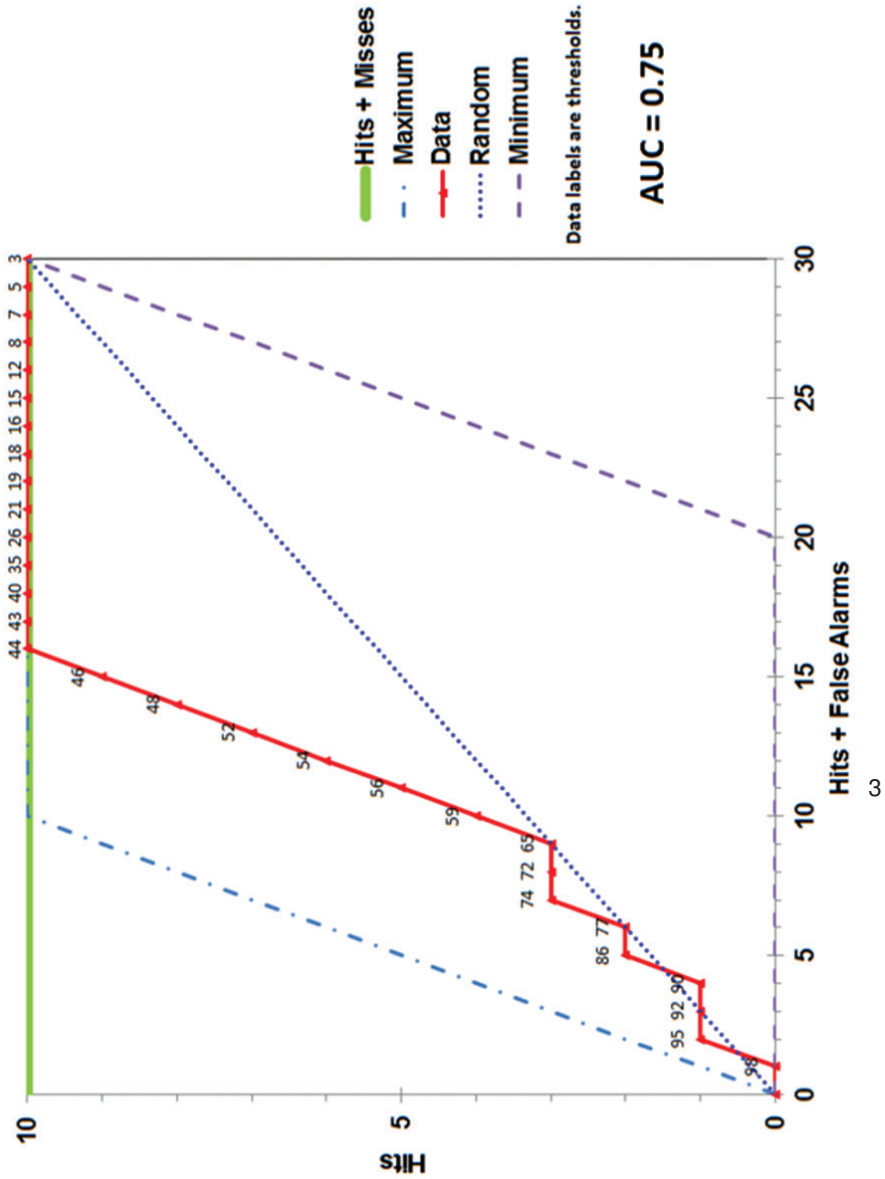


Fig. TOC Curve.

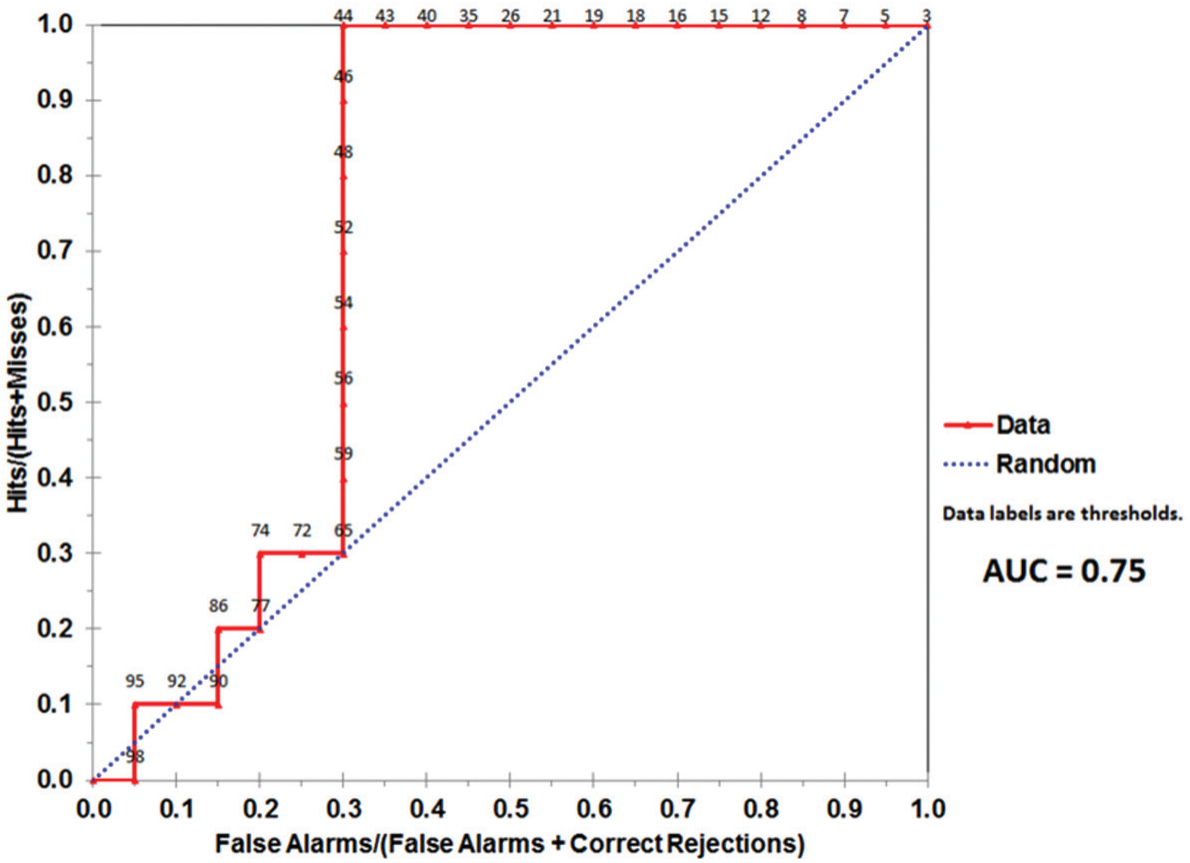


Fig. ROC Curve.

These figures are the TOC and ROC curves using the same data and thresholds. Consider the point that corresponds to a threshold of 74. The TOC curve shows the number of hits, which is 3, and hence the number of misses, which is 7. Additionally, the TOC curve shows that the number of false alarms is 4 and the number of correct rejections is 16. At any given point in the ROC curve, it is possible to glean values for the ratios of false alarms/(false alarms + correct rejections) and hits/(hits + misses). For example, at threshold 74, it is evident that the x coordinate is 0.3 and the y coordinate is 0.2. However, these two values are insufficient to construct all entries of the underlying two-by-two contingency table.

DETECTION ERROR TRADEOFF GRAPH

An alternative to the ROC curve is the detection error tradeoff (DET) graph, which plots the false negative rate (missed detections) vs. the false positive rate (false alarms) on non-linearly transformed x- and y-axes. The transformation function is the quantile function of the normal distribution, *i.e.*, the inverse of the cumulative normal distribution.

It is, in fact, the same transformation as zROC, below, except that the complement of the hit rate, the miss rate or false negative rate, is used. This alternative spends more graph area on the region of interest. Most of the ROC area is of little interest; one primarily cares about the region tight against the y-axis and the top left corner – which, because of using miss rate instead of its complement, the hit rate, is the lower left corner in a DET plot. Furthermore, DET graphs have the useful property of linearity and a linear threshold behaviour for normal distributions. The DET plot is used extensively in the automatic speaker recognition community, where the name DET was first used. The analysis of the ROC performance in graphs with this warping of the axes was used by psychologists in perception studies halfway through the 20th century, where this was dubbed "double probability paper".

Z-SCORE

If a standard score is applied to the ROC curve, the curve will be transformed into a straight line. This z-score is based on a normal distribution with a mean of zero and a standard deviation of one. In memory strength theory, one must assume that the zROC is not only linear, but has a slope of 1.0. The normal distributions of targets (studied objects that the subjects need to recall) and lures (non- studied objects that the subjects attempt to recall) is the factor causing the zROC to be linear. The linearity of the zROC curve depends on the standard deviations of the target and lure strength distributions. If the standard deviations are equal, the slope will be 1.0. If the standard deviation of the target strength distribution is larger than the standard deviation of the lure strength distribution, then the slope will be smaller than 1.0. In most studies, it has been found that the zROC curve slopes constantly fall below 1, usually between 0.5 and 0.9. Many experiments yielded a zROC slope of 0.8. A slope of 0.8 implies that the variability of the target strength distribution is 25 per cent larger than the variability of the lure strength distribution. Another variable used is *d'* (d prime) (discussed above in "Other measures"), which can easily be expressed in terms of z-values. Although *d*' is a commonly used parameter, it must be recognized that it is only relevant when strictly adhering to the very strong assumptions of strength theory made above.

The z-score of an ROC curve is always linear, as assumed, except in special situations. The Yonelinas familiarity-recollection model is a two-dimensional account of recognition memory. Instead of the subject simply answering yes or no to a specific input, the subject gives the input a feeling of familiarity, which operates like the original ROC curve. What changes, though, is a parameter for Recollection (R). Recollection is assumed to be all-or-none, and it trumps familiarity. If there were no recollection component, zROC would have a predicted slope of 1.

However, when adding the recollection component, the zROC curve will be concave up, with a decreased slope. This difference in shape and slope result from an added element of variability due to some items being recollected. Patients with anterograde amnesia are unable to recollect, so their Yonelinas zROC curve would have a slope close to 1.0.

HISTORY

The ROC curve was first used during World War II for the analysis of radar signals before it was employed in signal detection theory. Following the attack on Pearl Harbour in 1941, the United States army began new research to increase the prediction of correctly detected Japanese aircraft from their radar signals. For these purposes they measured the ability of a radar receiver operator to make these important distinctions, which was called the Receiver Operating Characteristic. In the 1950s, ROC curves were employed in psychophysics to assess human (and occasionally non-human animal) detection of weak signals.

In medicine, ROC analysis has been extensively used in the evaluation of diagnostic tests. ROC curves are also used extensively in epidemiology and medical research and are frequently mentioned in conjunction with evidence-based medicine. In radiology, ROC analysis is a common technique to evaluate new radiology techniques. In the social sciences, ROC analysis is often called the ROC Accuracy Ratio, a common technique for judging the accuracy of default probability models. ROC curves are widely used in laboratory medicine to assess the diagnostic accuracy of a test, to choose the optimal cut-off of a test and to compare diagnostic accuracy of several tests.

ROC curves also proved useful for the evaluation of machine learning techniques. The first application of ROC in machine learning was by Spackman who demonstrated the value of ROC curves in comparing and evaluating different classification algorithms.

ROC curves are also used in verification of forecasts in meteorology.

RECURSIVE PARTITIONING

Recursive partitioning is a statistical method for multivariable analysis. Recursive partitioning creates a decision tree that strives to correctly classify members of the population by splitting it into sub-populations based on several dichotomous independent variables. The process is termed recursive because each sub-population may in turn be split an indefinite number of times until the splitting process terminates after a particular stopping criterion is reached.

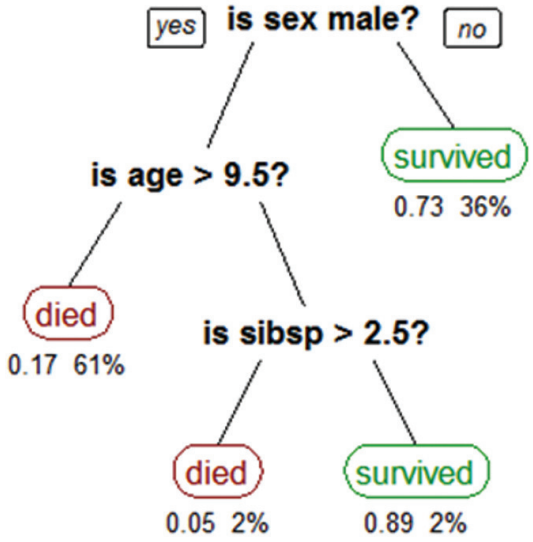


Fig. A recursive partitioning tree showing survival of passengers on the Titanic ("sibsp" is the number of spouses or siblings aboard). The figures under the leaves show the probability of survival and the percentage of observations in the leaf. Summarizing: Your chances of survival were good if you were (i) a female or (ii) a young boy without several family members.

Recursive partitioning methods have been developed since the 1980s. Well known methods of recursive partitioning include Ross Quinlan's ID3 algorithm and its successors, C4.5 and C5.0 and Classification and Regression Trees. Ensemble learning methods such as Random Forests help to overcome a common criticism of these methods - their vulnerability to overfitting of the data - by employing different algorithms and combining their output in some way.

This article focuses on recursive partitioning for medical diagnostic tests, but the technique has far wider applications. As compared to regression analysis, which creates a formula that health care providers can use to calculate the probability that a patient has a disease, recursive partition creates a rule such as 'If a patient has finding x, y, or z they probably have disease q'. A variation is 'Cox linear recursive partitioning'.

ADVANTAGES AND DISADVANTAGES

Compared to other multivariable methods, recursive partitioning has advantages and disadvantages.

- Advantages are:
  - (a) Generates clinically more intuitive models that do not require the user to perform calculations.
  - (b) Allows varying prioritizing of misclassifications in order to create a decision rule that has more sensitivity or specificity.
  - (c) May be more accurate.
- Disadvantages are:
  - (a) Does not work well for continuous variables
  - (b) May overfit data.

EXAMPLES

Examples are available of using recursive partitioning in research of diagnostic tests. Goldman used recursive partitioning to prioritize sensitivity in the diagnosis of myocardial infarction among patients with chest pain in the emergency room.

RELATIVE INDEX OF INEQUALITY

The relative index of inequality (RII) is a regression-based index which summarizes the magnitude of socio-economic status (SES) as a source of inequalities in health. RII is useful because it takes into account the size of the population and the relative disadvantage experienced by different groups. The disease outcome is regressed on the proportion of the population that has a higher position in the hierarchy.

The RII is particularly valuable when comparing risk factors (independent variables) that are on very different scales (e.g. low SES, low IQ, cigarette smoking). The RII is calculated in the following way:

- Rank cases on each of the variables
- For tied ranks and for categorical variables, assign the mean rank
- Divide the ranks by the sample size, creating a value ranging from 0 to 1

INTERPRETATION OF RII

The interpretation of RII is similar to the relative risk. It summarizes the relative risk for the most advantaged group (at the top of the hierarchy) compared to the least advantaged group (at the bottom of the hierarchy). This interpretation assumes that the variables have been scored so that higher scores are consistent with increased risk. For example, an RII of 1.88 (95 per cent confidence intervals 1.27 to 2.77), an indicator of low SES, on the risk of long term illness, implies that those in the most deprived group are 1.88 times more likely to experience illness than those in the least deprived group.

LIMITATIONS OF RII

One disadvantage of the RII is that it may capitalize on skewed data, inflating the apparent relative risk. A second limitation is that a large RII may arise for two reasons. First, it may represent a large effect of SES on disease. Second, it may reflect large differences between those with the most SES and those with the least (*i.e.* large inequalities in SES itself).

SEED-BASED D MAPPING

Seed-based d mapping (formerly Signed differential mapping) or SDM is a statistical technique created by Joaquim Radua for meta-analyzing studies on differences in brain activity or structure which used neuroimaging techniques such as fMRI, VBM, DTI or PET. It may also refer to a specific piece of software created by the SDM Project to carry out such meta-analyses.

THE SEED-BASED D MAPPING APPROACH

Overview of the Method

SDM adopted and combined various positive features from previous methods, such as ALE or MKDA, and introduced a series of improvements and novel features. One of the new features, introduced to avoid positive and negative findings in the same voxel as seen in previous methods, was the representation of both positive

differences and negative differences in the same map, thus obtaining a signed differential map (SDM). Another relevant feature, introduced in version 2.11, was the use of effect sizes (leading to effect-size SDM or 'ES-SDM'), which allows combination of reported peak coordinates with statistical parametric maps, thus allowing more exhaustive and accurate meta-analyses. The method has three steps. First, coordinates of cluster peaks (*e.g.* the voxels where the differences between patients and healthy controls were highest), and statistical maps if available, are selected according to SDM inclusion criteria. Second, coordinates are used to recreate statistical maps, and effect-sizes maps and their variances are derived from t-statistics (or equivalently from p-values or z-scores). Finally, individual study maps are meta-analyzed using different tests to complement the main outcome with sensitivity and heterogeneity analyses.

Inclusion Criteria

It is not uncommon in neuroimaging studies that some regions (*e.g.* a priori regions of interest) are more liberally thresholded than the rest of the brain. However, a meta-analysis of studies with such intra-study regional differences in thresholds would be biased towards these regions, as they are more likely to be reported just because authors apply more liberal thresholds in them. In order to overcome this issue SDM introduced a criterion in the selection of the coordinates: while different studies may employ different thresholds, you should ensure that the same threshold throughout the whole brain was used within each included study.

Pre-processing of Studies

After conversion of statistical parametric maps and peak coordinates to Talairach space, an SDM map is created for each study within a specific gray or white matter template. Pre-processing of statistical parametric maps is straightforward, while pre-processing of reported peak coordinates requires recreating the clusters of difference by means of an un-normalized Gaussian Kernel, so that voxels closer to the peak coordinate have higher values. A rather large full-width at half-maximum (FWHM) of 20mm is used to account for different sources of spatial error, *e.g.* coregistration mismatch in the studies, the size of the cluster or the location of the peak within the cluster. Within a study, values obtained by close Gaussian kernels are summed, though values are combined by square-distance-weighted averaging.

Statistical Comparisons

SDM provides several different statistical analyses in order to complement the main outcome with sensitivity and heterogeneity analyses.

- The main statistical analysis is the mean analysis, which consists in calculating the mean of the voxel values in the different studies. This mean is weighted by the inverse of the variance and accounts for inter-study heterogeneity (QH maps).
- Subgroup analyses are mean analyses applied to groups of studies to allow the study of heterogeneity.
- Linear model analyses (*e.g.* meta-regression) are a generalization of the mean analysis to allow comparisons between groups and the study of possible confounds. A low variability of the regressor is critical in meta-regressions, so they are recommended to be understood as exploratory and to be more conservatively thresholded.
- Jack-knife analysis consists in repeating a test as many times as studies have been included, discarding one different study each time, *i.e.* removing one study and repeating the analyses, then putting that study back and removing another study and repeating the analysis, and so on. The idea is that if a significant brain region remains significant in all or most of the combinations of studies it can be concluded that this finding is highly replicable.

The statistical significance of the analyses is checked by standard randomization tests. It is recommended to use uncorrected p-values = 0.005, as this significance has been found in this method to be approximately equivalent to a corrected p-value = 0.05. A false discovery rate (FDR) = 0.05 has been found in this method to be too conservative. Values in a Talairach label or coordinate can also be extracted for further processing or graphical presentation.

SDM SOFTWARE

SDM is software written by the SDM project to aid the meta-analysis of voxel-based neuroimaging data. It is distributed as freeware including a graphical interface and a menu/command-line console. It can also be integrated as an SPM extension.

SENSITIVITY AND SPECIFICITY

Sensitivity and specificity are statistical measures of the performance of a binary classification test, also known in statistics as a classification function:

- Sensitivity (also called the true positive rate, the recall, or probability of detection in some fields) measures the proportion of actual positives that are correctly identified as such (*e.g.*, the percentage of sick people who are correctly identified as having the condition).
- Specificity (also called the true negative rate) measures the proportion of actual negatives that are correctly identified as such (*e.g.*, the percentage of healthy people who are correctly identified as not having the condition).

Equivalently, in medical tests sensitivity is the extent to which actual positives are not overlooked (so false negatives are few), and specificity is the extent to which actual negatives are classified as such (so false positives are few).

Thus a highly sensitive test rarely overlooks an actual positive (for example, showing "nothing bad" despite something bad existing); a highly specific test rarely registers a positive classification for anything that is not the target of testing (for example, finding one bacterial species and mistaking it for another closely related one that is the true target); and a test that is highly sensitive *and* highly specific does both, so it "rarely overlooks a thing that it is looking for" *and* it "rarely mistakes anything else for that thing." Because most medical tests do not have sensitivity and specificity values above 99 per cent, "rarely" does *not* equate to certainty. But for practical reasons, tests with sensitivity and specificity values above 90 per cent have high credibility, albeit usually no certainty, in differential diagnosis.

Sensitivity therefore quantifies the avoiding of false negatives, and specificity does the same for false positives. For any test, there is usually a trade-off between the measures – for instance, in airport security since testing of passengers is for potential threats to safety, scanners may be set to trigger alarms on low-risk items like belt buckles and keys (low specificity), in order to increase the probability of identifying dangerous objects and minimize the risk of missing objects that do pose a threat (high sensitivity). This trade-off can be represented graphically using a receiver operating characteristic curve. A perfect predictor would be described as 100 per cent sensitive, meaning all sick individuals are correctly identified as sick, and 100 per cent specific, meaning no healthy individuals are incorrectly identified as sick. In reality, however, any non-deterministic predictor will possess a minimum error bound known as the Bayes error rate.

DEFINITIONS

In the terminology *true/false positive/negative*, *true* or *false* refers to the assigned classification being correct or incorrect, while *positive* or *negative* refers to assignment to the positive or the negative category.

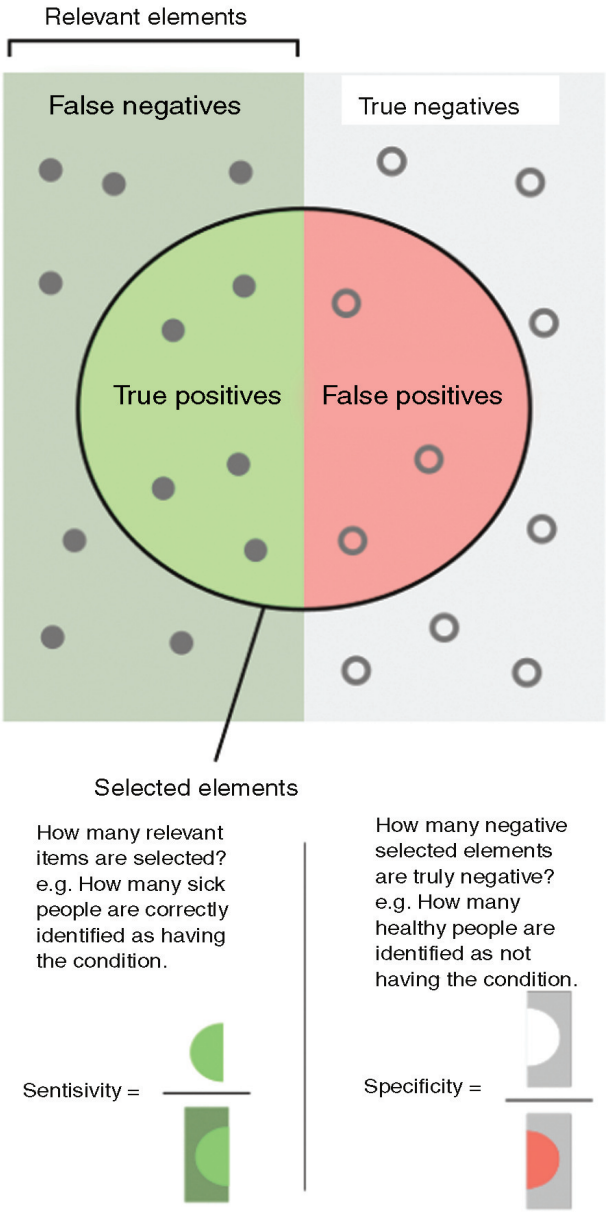


Fig. Sensitivity and specificity.

Application to Screening Study

Imagine a study evaluating a new test that screens people for a disease. Each person taking the test either has or does not have the disease. The test outcome can be positive (classifying the person as having the disease) or negative (classifying the person as not having the disease). The test results for each subject may or may not match the subject's actual status. In that setting:

- *True positive:* Sick people correctly identified as sick
- *False positive:* Healthy people incorrectly identified as sick
- *True negative:* Healthy people correctly identified as healthy
- *False negative:* Sick people incorrectly identified as healthy

In general, Positive = identified and negative = rejected. Therefore:

- True positive = correctly identified
- False positive = incorrectly identified

- True negative = correctly rejected
- False negative = incorrectly rejected

MEDICAL EXAMPLES

In medical diagnosis, test sensitivity is the ability of a test to correctly identify those with the disease (true positive rate), whereas test specificity is the ability of the test to correctly identify those without the disease (true negative rate). If 100 patients known to have a disease were tested, and 43 test positive, then the test has 43 per cent sensitivity. If 100 with no disease are tested and 96 return a negative result, then the test has 96 per cent specificity. Sensitivity and specificity are prevalence-independent test characteristics, as their values are intrinsic to the test and do not depend on the disease prevalence in the population of interest. Positive and negative predictive values, but not sensitivity or specificity, are values influenced by the prevalence of disease in the population that is being tested. These concepts are illustrated graphically in this applet Bayesian clinical diagnostic model which show the positive and negative predictive values as a function of the prevalence, the sensitivity and specificity.

Misconceptions

It is often claimed that a highly specific test is effective at ruling in a disease when positive, while a highly sensitive test is deemed effective at ruling out a disease when negative. This has led to the widely used mnemonics SPIN and SNOUT, according to which a highly SPecific test, when Positive, rules IN disease (SP-P-IN), and a highly 'SeNsitive' test, when Negative rules OUT disease (SN-N-OUT). Both rules of thumb are, however, inferentially misleading, as the diagnostic power of any test is determined by both its sensitivity *and* its specificity.

The tradeoff between Specificity and Sensitivity is explored in ROC analysis as a trade off between TPR and FPR (that is Recall and Fallout). Giving them equal weight optimizes Informedness = Specificity+Sensitivity-1 = TPR-FPR, the magnitude of which gives the probability of an informed decision between the two classes (>0 represents appropriate use of information, 0 represents chance-level performance, <0 represents perverse use of information).

ESTIMATION OF ERRORS IN QUOTED SENSITIVITY OR SPECIFICITY

Sensitivity and specificity values alone may be highly misleading. The 'worst-case' sensitivity or specificity must be calculated in order to avoid reliance on experiments with few results. For example, a particular test may easily show 100 per cent sensitivity if tested against the gold standard four times, but a single additional test against the gold standard that gave a poor result would imply a sensitivity of only 80 per cent. A common way to do this is to state the binomial proportion confidence interval, often calculated using a Wilson score interval. Confidence intervals for sensitivity and specificity can be calculated, giving the range of values within which the correct value lies at a given confidence level (*e.g.*, 95 per cent).

TERMINOLOGY IN INFORMATION RETRIEVAL

In information retrieval, the positive predictive value is called precision, and sensitivity is called recall. Unlike the Specificity vs Sensitivity tradeoff, these measures are both independent of the number of true negatives, which is generally unknown and much larger than the actual numbers of relevant and retrieved documents. This assumption of very large numbers of true negatives versus positives is rare in other applications. The F-score can be used as a single measure of performance of the test for the positive class. The F-score is the harmonic mean of precision and recall:

$$F = 2 \times \frac{\text{precision} \times \text{recall}}{\text{precision} + \text{recall}}$$

In the traditional language of statistical hypothesis testing, the sensitivity of a test is called the statistical power of the test, although the word *power* in that context has a more general usage that is not applicable in the present context. A sensitive test will have fewer Type II errors.

SHIFTING BASELINE

A shifting baseline (also known as sliding baseline) is a type of change to how a system is measured, usually against previous reference points (baselines), which themselves may represent significant changes from an even earlier state of the system. The concept arose in landscape architect Ian McHarg's 1969 manifesto *Design With Nature* in which the modern landscape is compared to that on which ancient people once lived. The concept was then considered by the fisheries scientist Daniel Pauly in his paper "Anecdotes and the shifting baseline syndrome of fisheries". Pauly developed the concept in reference to fisheries management where fisheries scientists sometimes fail to identify the correct "baseline" population size (*e.g.* how abundant a fish species population was *before* human exploitation) and thus work with a shifted baseline. He describes the way that radically depleted fisheries were evaluated by experts who used the state of the fishery at the start of their careers as the baseline, rather than the fishery in its untouched state. Areas that swarmed with a particular species hundreds of years ago, may have experienced long term decline, but it is the level of decades previously that is considered the appropriate reference point for current populations. In this way large declines in ecosystems or species over long periods of time were, and are, masked. There is a loss of perception of change that occurs when each generation redefines what is "natural". Most modern fisheries stock assessments do not ignore historical fishing and account for it by either including the historical catch or use other techniques to reconstruct the depletion level of the population at the start of the period for which adequate data is available. Anecdotes about historical populations levels can be highly unreliable and result in severe mismanagement of the fishery.

The concept was further refined and applied to the ecology of kelp forests by Paul Dayton and others from the Scripps Institution of Oceanography. They used a slightly different version of the term in their paper, "Sliding baselines, ghosts, and reduced expectations in kelp forest communities". Both terms refer to a shift over time in the expectation of what a healthy ecosystem baseline looks like.

BROADENED MEANING

In 2002, filmmaker and former marine biologist Randy Olson broadened the definition of shifting baselines with an op-ed in the *Los Angeles Times*. He explained the relevance of the concept to all aspects of change and the failure to notice change in the world today. He and coral reef ecologist Jeremy Jackson (of Scripps Institution of Oceanography) co-founded *The Shifting Baselines Ocean Media Project* in 2003 to help promote a wider understanding and use of the concept in conservation policy. The Shifting Baselines Ocean Media Project grew from its three founding partners (Scripps Institution of Oceanography, The Ocean Conservancy, and Surfrider Foundation) to over twenty conservation groups and science organizations. The project has produced dozens of short films, public service announcements, and Flash videos along with photography, video, and stand-up comedy contests, all intended to promote the term to a broader audience. The Shifting Baselines Blog, "the cure for planetary amnesia" is run by the Shifting Baselines Ocean Media Project on the Seed (magazine) Science Blogs.

SPECTRUM BIAS

In biostatistics, spectrum bias refers to the phenomenon that the performance of a diagnostic test may vary in different clinical settings because each setting has a different mix of patients. Because the performance may be

dependent on the mix of patients, performance at one clinic may not be predictive of performance at another clinic. These differences are interpreted as a kind of *bias*. Mathematically, the spectrum bias is a sampling bias and not a traditional statistical bias; this has led some authors to refer to the phenomenon as *spectrum effects*, whilst others maintain it is a bias if the true performance of the test differs from that which is 'expected'. Usually the performance of a diagnostic test is measured in terms of its sensitivity and specificity and it is changes in these that are considered when referring to spectrum bias. However, other performance measures such as the likelihood ratios may also be affected by spectrum bias. Generally spectrum bias is considered to have three causes. The first is due to a change in the case-mix of those patients with the target disorder (disease) and this affects the sensitivity. The second is due to a change in the case-mix of those without the target disorder (disease-free) and this affects the specificity. The third is due to a change in the prevalence, and this affects both the sensitivity and specificity. This final cause is not widely appreciated, but there is mounting empirical evidence as well as theoretical arguments which suggest that it does indeed affect a test's performance. Examples where the sensitivity and specificity change between different sub-groups of patients may be found with the carcinoembryonic antigen test and urinary dipstick tests.

Diagnostic test performances reported by some studies may be artificially overestimated if it is a case-control design where a healthy population ('fittest of the fit') is compared with a population with advanced disease ('sickest of the sick'); that is two extreme populations are compared, rather than typical healthy and diseased populations. If properly analyzed, recognition of heterogeneity of subgroups can lead to insights about the test's performance in varying populations.

STANDARDIZED MORTALITY RATIO

The standardized mortality ratio or SMR, is a quantity, expressed as either a ratio or percentage quantifying the increase or decrease in mortality of a study cohort with respect to the general population.

STANDARDIZED MORTALITY RATIO

The standardized mortality ratio is the ratio of observed deaths in the study group to expected deaths in the general population. This ratio can be expressed as a percentage simply by multiplying by 100. The SMR may be quoted as either a ratio or a percentage. If the SMR is quoted as a ratio and is equal to 1.0, then this means the number of observed deaths equals that of expected cases. If higher than 1.0, then there is a higher number of deaths than is expected. SMR constitutes an indirect form of standardization. It has an advantage over the direct method of standardization since age-adjustment is permitted in situations where age stratification may not be available for the cohort being studied or where strata-specific data are subject to excessive random variability.

Definition

The requirements for calculating SMR for a cohort are:

- The number of persons in each age group in the population being studied
- The age specific death rates of the general population in the same age groups of the study population
- The observed deaths in the study population

Expected deaths would then be calculated simply by multiplying the death rates of the general population by the total number of participants in the study group at the corresponding age group and summing up all the values for each age group to arrive at the number of expected deaths. The study groups are weighted based on their particular distribution (for example, age), as opposed to the general populations's distribution. This is a fundamental distinction between an indirect method of standardization like SMR from direct standardization techniques.

The SMR may well be quoted with an indication of the uncertainty associated with its estimation, such as a confidence interval (CI) or p value, which allows it to be interpreted in terms of statistical significance.

Example

An example might be a cohort study into cumulative exposure to arsenic from drinking water, whereby the mortality rates due to a number of cancers in a highly exposed group (which drinks water with a mean arsenic concentration of, say 10 mg) is compared with those in the general population. An SMR for bladder cancer of 1.70 in the exposed group would mean that there is 70 per cent more cases of death due to bladder cancer in the cohort than in the reference population (in this case the national population, which is generally considered not to exhibit cumulative exposure to high arsenic levels).

STANDARDIZED MORTALITY RATE

Standardized mortality rate tells how many persons, per thousand of the population, will die in a given year and what the causes of death will be.

Such statistics have many uses:

- Life insurance companies periodically update their premiums based on the mortality rate, adjusted for age.
- Medical researchers can track disease-related deaths and shift focus and funding to address increasing or decreasing risks.
- Organizations, both non- and for-profit, can utilize such statistics to justify their missions.
- Regarding occupational uses:

Mortality tables are also often used when numbers of deaths for each age-specific stratum are not available. It is also used to study mortality rate in an occupationally exposed population: Do people who work in a certain industry, such as mining or construction, have a higher mortality than people of the same age in the general population? Is an additional risk associated with that occupation? To answer the question of whether a population of miners has a higher mortality than we would expect in a similar population that is not engaged in mining, the age-specific rates for such a known population, such as all men of the same age, are applied to each age group in the population of interest. This will yield the number of deaths expected in each age group in the population of interest, if this population had had the mortality experience of the known population. Thus, for each age group, the number of deaths expected is calculated, and these numbers are totaled. The numbers of deaths that were actually observed in that population are also calculated and totaled. The ratio of the total number of deaths actually observed to the total number of deaths expected, if the population of interest had had the mortality experience of the known population, is then calculated. This ratio is called the standardized mortality ratio (SMR). The SMR is defined as follows: SMR = (Observed no. of deaths per year)/(Expected no. of deaths per year).

STANDARDIZED RATE

Standardized rates are a statistical measure of any rates in a population. These are adjusted rates that take into account the vital differences between populations that may affect their birthrates or death rates.

EXAMPLES

The most common are birth, death and unemployment rates. For example, in a community made up of primarily young couples, the birthrate might appear to be high when compared to that of other populations. However, by calculating the standardized birthrates that is by comparing the same age group in other populations), a more realistic picture of childbearing capacity will be developed.

FORMULA

The formula for standardized rates is as follows:

Σ(crude rate for age group × standard population for age group)/ Σstandard population

STATISTICAL EPIDEMIOLOGY

Statistical epidemiology is an emerging branch of the disciplines of epidemiology and biostatistics that aims to:

- Bring more statistical rigour to bear in the field of epidemiology
- Recognise the importance of applied statistics, especially with respect to the context in which statistical methods are appropriate and inappropriate
- Aid and improve our interpretation of observations

The science of epidemiology has had enormous growth, particularly with charity and government funding. Many researchers have been trained to conduct studies, requiring multiple skills ranging from liaising with clinical staff to the statistical analysis of complex data, such as using Bayesian methods. The role of a Statistical Epidemiologist is to bring the most appropriate methods available to bear on observational study from medical research, requiring a broad appreciation of the underpinning methods and their context of applicability and interpretation. The earliest mention of this phrase was in an article by EB Wilson, taking a critical look at the way in which statistical methods were developing and being applied in the science of epidemiology.

ACADEMIC RECOGNITION

There are two Professors of Statistical Epidemiology in the United Kingdom (University of Leeds and Imperial College, London) and a Statistical Epidemiology group (Oxford University).

STATISTICAL PARAMETRIC MAPPING

Statistical parametric mapping or SPM is a statistical technique created by Karl Friston for examining differences in brain activity recorded during functional neuroimaging experiments using neuroimaging technologies such as fMRI or PET. It may also refer to a specific piece of software created by the *Wellcome Department of Imaging Neuroscience* (part of University College London) to carry out such analyses.

APPROACH

Unit of measurement (l2h)

Functional neuroimaging, one type of 'brain scanning', involves the measurement of brain activity. The specific technique used to measure brain activity depends on the imaging technology being used. Regardless of which technology is used, the scanner produces a 'map' of the area being scanned that is represented as voxels. Each voxel typically represents the activity of a particular coordinate in three-dimensional space. The exact size of a voxel will vary depending on the technology used, although fMRI voxels typically represent a volume of 27 mm (a cube with 3mm length sides).

Experimental Design

Researchers are often interested in examining brain activity linked to a specific psychological process or processes. An experimental approach to this problem might involve asking the question 'which areas of the brain are significantly more active when a person is doing task A compared to task B?'. Although each task might be designed to be identical, except for the aspect of behaviour under investigation, the brain is still likely

to show changes in activity between tasks due to factors other than task differences (as the brain is involved with co-ordinating a whole range of parallel functions unrelated to the experimental task). Furthermore, the signal may contain noise from the imaging process itself. To accommodate these random effects, and to highlight the areas of activity linked specifically to the process under investigation, statistics are used to look for the most significant difference above and beyond background brain activity. This involves a multi-stage process to prepare the data, and to subsequently analyse it using a statistical method known as the general linear model.

Image Pre-processing

Images from the brain scanner may be pre-processed before any statistical comparison takes place to remove noise or correct for sampling errors. A study will usually scan a subject several times. To account for the motion of the head between scans, the images will usually be adjusted so each of the voxels in the images corresponds (approximately) to the same site in the brain.

Functional neuroimaging studies usually involve several participants, who will have slightly differently shaped brains. All are likely to have the same gross anatomy, but there will be minor differences in overall brain size, individual variation in topography of the gyri and sulci of the cerebral cortex, and morphological differences in deep structures such as the corpus callosum. To aid comparisons, the 3D image of each brain is transformed so that superficial structures line up, a process known as *spatial normalization*. Such normalization typically involves not only translation and rotation, but also scaling and nonlinear warping of the brain surface to match a standard template. Standard brain maps such as the Talairach-Tournoux or templates from the Montréal Neurological Institute (MNI) are often used to allow researchers from across the world to compare their results. Images are often smoothed (similar to the 'blur' effect used in some image-editing software) by which voxels are averaged with their neighbours, typically using a Gaussian filter or by wavelet transformation, to make the data less noisy.

Statistical Comparison

Parametric statistical models are assumed at each voxel, using the general linear model to describe the variability in the data in terms of experimental and confounding effects, and residual variability. Hypotheses expressed in terms of the model parameters are assessed at each voxel with univariate statistics. Analyses may also be conducted to examine differences over a time series (*i.e.* correlations between a task variable and brain activity in a certain area) using linear convolution models of how the measured signal is caused by underlying changes in neural activity. Because many statistical tests are being conducted, adjustments have to be made to control for Type I errors (false positives) potentially caused by the comparison of levels of activity at a large number of voxels. In this case, a Type I error would result in falsely detecting background brain activity as activity related to the task. Adjustments are made, based on the number of resels in the image and the theory of continuous random fields in order to set a new criterion for statistical significance that adjusts for the problem of multiple comparisons.

Graphical Representations

Differences in measured brain activity can be represented in a number of ways. Most simply, they can be presented as a table, displaying coordinates that show the most significant differences in activity between tasks. However, differences in brain activity are more often shown as patches of colour on an MRI brain 'slice', with the colours representing the location of voxels that have shown statistically significant differences between conditions. The gradient of colour is mapped to statistical values, such as t-values or z-scores. This creates an intuitive and visually appealing means of delineating the relative statistical strength of a given area of activation.

Differences in activity may also be represented as a 'glass brain', a representation of three outline views of the brain as if it were transparent. Only the patches of activation are visible as areas of shading. This is useful as a quick means of summarizing the total area of significant change in a given statistical comparison.

SPM SOFTWARE

SPM is software written by the Wellcome Department of Imaging Neuroscience at University College London to aid in the analysis of functional neuroimaging data. It is written using MATLAB and is distributed as free software.

STATPLUS

StatPlus is a software product for basic univariate and multivariate statistical analysis (MANOVA, GLM, Latin squares), as well as time series analysis, nonparametric statistics, survival analysis and statistical charts including control charts. It was originally developed for use in the biomedical sciences. The original version is now known as BioStat. It is mostly used in biomedicine and natural sciences. The software has a version for the Mac OS X known as StatPlus:mac. The software may also be used as an add-on to the Microsoft Excel.

WINPEPI

WinPepi is a freeware package of statistical programmes for epidemiologists, comprising seven programmes with over 120 modules. WinPepi is not a complete compendium of statistical routines for epidemiologists but it provides a very wide range of procedures, including those most commonly used and many that are not easy to find elsewhere. This has repeatedly led reviewers to use a "Swiss army knife" analogy. Each programme has a comprehensive fully referenced manual. WinPepi had its origins in 1983 in a book of programmes for hand-held calculators,. In 1993, this was developed into a set of DOS-based computer programmes by Paul M. Gahlinger with the assistance of one of the original authors of calculator programmes, Prof. JH Abramson that came to be called Pepi (an acronym for "Programmes for EPIdemiologists") and evolved, after its fourth version in 2001, into WinPepi (Pepi-for-Windows). New expanded versions are issued at frequent intervals. The programmes are notable for their user-friendliness. A portal links to programmes and manuals. Menus, buttons, on-screen instructions, help screens, pop-up hints, and built-in error traps are also provided. The programmes can also be operated from a USB flash drive.

WinPepi does not provide data management facilities. With some exceptions, it requires the entry (at the keyboard or by pasting from a spreadsheet or text file) of data that have already been counted or summarized.

YOUDEN'S J STATISTIC

Youden's J statistic (also called Youden's index) is a single statistic that captures the performance of a dichotomous diagnostic test. Informedness is its generalization to the multiclass case and estimates the probability of an informed decision.

DEFINITION

Youden's *J* statistic is

$$J = \text{sensitivity} + \text{specificity} - 1$$

with the two right-hand quantities being sensitivity and specificity. Thus the expanded formula is:

$$J = \frac{\text{true positives}}{\text{true positives} + \text{false negatives}} + \frac{\text{true negatives}}{\text{ture negatives} + \text{false positives}} - 1$$

The index was suggested by W.J. Youden in 1950 as a way of summarising the performance of a diagnostic test. Its value ranges from 0 through 1 (inclusive) [source: Youden's original article in 'Cancer'], and has a zero value when a diagnostic test gives the same proportion of positive results for groups with and without the disease, i.e the test is useless. A value of 1 indicates that there are no false positives or false negatives, *i.e.* the test is perfect. The index gives equal weight to false positive and false negative values, so all tests with the same value of the index give the same proportion of total misclassified results. Youden's index is often used in conjunction with receiver operating characteristic (ROC) analysis. The index is defined for all points of an ROC curve, and the maximum value of the index may be used as a criterion for selecting the optimum cut-off point when a diagnostic test gives a numeric rather than a dichotomous result. The index is represented graphically as the height above the chance line, and it is also equivalent to the area under the curve subtended by a single operating point. Youden's index is also known as deltap' and generalizes from the dichotomous to the multiclass case as Informedness. The use of a single index is "not generally to be recommended", but Informedness or Youden's index is the probability of an informed decision (as opposed to a random guess) and takes into account all predictions.

An unrelated but commonly used combination of basic statistics from Information Retrieval is the F-score, being a (possibly weighted) harmonic mean of recall and precision where recall = sensitivity = true positive rate, but specificity and precision are totally different measures. F-score, like recall and precision, only considers the so-called positive predictions, with recall being the probability of predicting just the positive class, precision being the probability of a positive prediction being correct, and F-score equating these probabilities under the effective assumption that the positive labels and the positive predictions should have the same distribution and prevalence, similar to the assumption underlying of Fleiss' kappa. Youden's J, Informedness, Recall, Precision and F-score are intrinsically undirectional, aiming to assess the deductive effectiveness of predictions in the direction proposed by a rule, theory or classifier. Markedness (deltap) is Youden's J used to assess the reverse or abductive direction, and matches well human learning of associations; rules and, superstitions as we model possible causation; while correlation and kappa evaluate bidirectionally.

Matthews correlation coefficient is the geometric mean of the regression coefficient of the problem and its dual, where the component regression coefficients of the Matthews correlation coefficient are Markedness (inverse of Youden's J or deltap) and Informedness (Youden's J or deltap'). Kappa statistics such as Fleiss' kappa and Cohen's kappa are methods for calculating inter-rater reliability based on different assumptions about the marginal or prior distributions, and are increasingly used as *chance corrected* alternatives to accuracy in other contexts. Fleiss' kappa, like F-score, assumes that both variables are drawn from the same distribution and thus have the same expected prevalence, while Cohen's kappa assumes that the variables are drawn from distinct distributions and referenced to a model of expectation that assumes prevalences are independent. When the true prevalences for the two positive variables are equal as assumed in Fleiss kappa and F-score, that is the number of positive predictions matches the number of positive classes in the dichotomous (two class) case, the different kappa and correlation measure collapse to identity with Youden's J, and recall, precision and F-score are similarly identical with accuracy.

3

Measures

OCCURRENCE

Incidence

Incidence in epidemiology is a measure of the probability of occurrence of a given medical condition in a population within a specified period of time. Although sometimes loosely expressed simply as the number of new cases during some time period, it is better expressed as a proportion or a rate with a denominator. Incidence proportion (also known as cumulative incidence) is the number of new cases within a specified time period divided by the size of the population initially at risk. For example, if a population initially contains 1,000 non-diseased persons and 28 develop a condition over two years of observation, the incidence proportion is 28 cases per 1,000 persons per two years, *i.e.* 2.8 per cent per two years.

Incidence Rate

The incidence rate is the number of new cases per population at risk in a given time period. When the denominator is the sum of the person-time of the at risk population, it is also known as the incidence density rate or person-time incidence rate. In the same example as above, the incidence rate is 14 cases per 1000 person-years, because the incidence proportion (28 per 1,000) is divided by the number of years (two). Using person-time rather than just time handles situations where the amount of observation time differs between people, or when the population at risk varies with time. Use of this measure implies the assumption that the incidence rate is constant over different periods of time, such that for an incidence rate of 14 per 1000 persons-years, 14 cases would be expected for 1000 persons observed for 1 year or 50 persons observed for 20 years. When this assumption is substantially violated, such as in describing survival after diagnosis of metastatic cancer, it may be more useful to present incidence data in a plot of cumulative incidence, over time, taking into account loss to follow-up, using a Kaplan-Meier Plot.

Consider the following example. Say you are looking at a sample population of 225 people, and want to determine the incidence rate of developing HIV over a 10-year period:

- At the beginning of the study (t=0) you find 25 cases of existing HIV. These people are not counted as they cannot develop HIV a second time.
- A follow-up at 5 years (t=5 years) finds 20 new cases of HIV.
- A second follow-up at the end of the study (t=10 years) finds 30 new cases.

If you were to measure prevalence you would simply take the total number of cases (25 + 20 + 30 = 75) and divide by your sample population (225). So prevalence would be 75/225 = 0.33 or 33 per cent (by the end of the study). This tells you how widespread HIV is in your sample population, but little about the actual risk of

developing HIV for any person over a coming year. To measure incidence you must take into account how many years each person contributed to the study, and when they developed HIV. When it is not known exactly when a person develops the disease in question, epidemiologists frequently use the actuarial method, and assume it was developed at a half-way point between follow-ups. In this calculation:

- At 5 yrs you found 20 new cases, so you assume they developed HIV at 2.5 years, thus contributing (20 * 2.5) =50 person-years of disease-free life.
- At 10 years you found 30 new cases. These people did not have HIV at 5 years, but did at 10, so you assume they were infected at 7.5 years, thus contributing (30 * 7.5)= 225 person-years of disease-free life. That is a total of (225 + 50)= 275 person years so far.
- You also want to account for the 150 people who never had or developed HIV over the 10-year period, (150 * 10) contributing 1500 person-years of disease-free life.

That is a total of (1500 + 275) = 1775 person-years of life. Now take the 50 new cases of HIV, and divide by 1775 to get 0.028, or 28 cases of HIV per 1000 population, per year. In other words, if you were to follow 1000 people for one year, you would see 28 new cases of HIV.

This is a much more accurate measure of risk than prevalence.

Incidence vs. Prevalence

Incidence should not be confused with prevalence, which is the proportion of cases in the population at a given time rather than rate of occurrence of new cases. Thus, incidence conveys information about the risk of contracting the disease, whereas prevalence indicates how widespread the disease is. Prevalence is the proportion of the total number of cases to the total population and is more a measure of the burden of the disease on society with no regard to time at risk or when subjects may have been exposed to a possible risk factor. Prevalence can also be measured with respect to a specific subgroup of a population. Incidence is usually more useful than prevalence in understanding the disease etiology: for example, if the incidence rate of a disease in a population increases, then there is a risk factor that promotes the incidence. For example, consider a disease that takes a long time to cure and was widespread in 2002 but dissipated in 2003. This disease will have both high incidence and high prevalence in 2002, but in 2003 it will have a low incidence yet will continue to have a high prevalence (because it takes a long time to cure, so the fraction of individuals that are affected remains high). In contrast, a disease that has a short duration may have a low prevalence and a high incidence. When the incidence is approximately constant for the duration of the disease, prevalence is approximately the product of disease incidence and average disease duration, so *prevalence* = *incidence* × *duration*. The importance of this equation is in the relation between prevalence and incidence; for example, when the incidence increases, then the prevalence must also increase. Note that this relation does not hold for age-specific prevalence and incidence, where the relation becomes more complicated.

CUMULATIVE INCIDENCE

Cumulative incidence or incidence proportion is a measure of frequency, as in epidemiology, where it is a measure of disease frequency during a period of time. Where the period of time considered is an entire lifetime, the incidence proportion is called lifetime risk.

Cumulative incidence is defined as the probability that a particular event, such as occurrence of a particular disease, has occurred before a given time. It is equivalent to the incidence, calculated using a period of time during which all of the individuals in the population are considered to be at risk for the outcome. It is sometimes also referred to as the incidence proportion.

Cumulative incidence is calculated by the number of new cases during a period divided by the number of subjects at risk in the population at the beginning of the study.

It may also be calculated by the incidence rate multiplied by duration:

$$\mathrm{CI}(t) = 1 - e^{-1R(t)\cdot D}.$$

PREVALENCE

Prevalence in epidemiology is the proportion of a particular population found to be affected by a medical condition (typically a disease or a risk factor such as smoking or seat-belt use). It is arrived at by comparing the number of people found to have the condition with the total number of people studied, and is usually expressed as a fraction, as a percentage, or as the number of cases per 10,000 or 100,000 people. Point prevalence is the proportion of a population that has the condition at a specific point in time. Period prevalence is the proportion of a population that has the condition at some time during a given period (*e.g.*, 12 month prevalence), and includes people who already have the condition at the start of the study period as well as those who acquire it during that period. Lifetime prevalence (LTP) is the proportion of a population that at some point in their life (up to the time of assessment) have experienced the condition. Prevalence estimates are used by epidemiologists, health care providers, government agencies, toxicologists, and insurers.

Prevalence is contrasted with incidence, which is a measure of *new* cases arising in a population over a given period (month, year, etc.). The difference between prevalence and incidence can be summarized thus: prevalence answers "How many people have this disease right now?" or "How many people have had this disease during this time period?" and incidence answers "How many people per year newly acquire this disease?"

Examples and Utility

In science, *prevalence* describes a proportion (typically expressed as a percentage). For example, the prevalence of obesity among American adults in 2001 was estimated by the U. S. Centers for Disease Control (CDC) at approximately 20.9 per cent. Prevalence is a term which means being widespread and it is distinct from incidence. Prevalence is a measurement of *all* individuals affected by the disease at a particular time, whereas incidence is a measurement of the number of *new* individuals who contract a disease durin;lm,;lm parameter when talking about long lasting diseases, such as HIV, but incidence is more useful when talking about diseases of short duration, such as chickenpox.

Uses

Lifetime Prevalence

Lifetime prevalence (LTP) is the proportion of individuals in a population that at some point in their life (up to the time of assessment) have experienced a "case", *e.g.*, a disease; a traumatic event; or a behaviour, such as committing a crime. Often, a 12-month prevalence (or some other type of "period prevalence") is provided in conjunction with lifetime prevalence. *Point prevalence* is the prevalence of disorder at a specific point in time (a month or less). *Lifetime morbid risk* is "the proportion of a population that might become afflicted with a given disease at any point in their lifetime."

Period Prevalence

Period prevalence is the proportion of the population with a given disease or condition over a specific period of time. It could describe how many people in a population had a cold over the cold season in 2006, for example. It is expressed as a percentage of the population and can be described by the following formula:

Period prevalence (ratio) = Number of cases that existed in a given period ÷ Number of people in the population during this period

The relationship between incidence (rate), point prevalence (ratio) and period prevalence (ratio) is easily explained via an analogy with photography. Point prevalence is akin to a flashlit photograph: what is happening at this instant frozen in time. Period prevalence is analogous to a long exposure (seconds, rather than an instant) photograph: the number of events recorded in the photo whilst the camera shutter was open. In a movie each frame records an instant (point prevalence); by looking from frame to frame one notices new events (incident events) and can relate the number of such events to a period (number of frames).

Point Prevalence

Point prevalence is a measure of the proportion of people in a population who have a disease or condition at a particular time, such as a particular date. It is like a snap shot of the disease in time. It can be used for statistics on the occurrence of chronic diseases. This is in contrast to period prevalence which is a measure of the proportion of people in a population who have a disease or condition over a specific period of time, say a season, or a year. Point prevalence can be described by the formula: Prevalence = Number of existing cases on a specific date ÷ Number of people in the population on this date.

Limitations

It can be said that a very small error applied over a very large number of individuals (that is, those who are *not affected* by the condition in the general population during their lifetime; for example, over 95 per cent) produces a relevant, non-negligible number of subjects who are incorrectly classified as having the condition or any other condition which is the object of a survey study: these subjects are the so-called false positives; such reasoning applies to the 'false positive' but not the 'false negative' problem where we have an error applied over a relatively very small number of individuals to begin with (that is, those who are *affected* by the condition in the general population; for example, less than 5 per cent). Hence, a very high percentage of subjects who seem to have a history of a disorder at interview are false positives for such a medical condition and apparently never suffered a fully clinical syndrome. A different but related problem in evaluating the public health significance of psychiatric conditions has been highlighted by Robert Spitzer of Columbia University: fulfillment of diagnostic criteria and the resulting diagnosis do not necessarily imply need for treatment.

A well-known statistical problem arises when ascertaining rates for disorders and conditions with a relatively low population prevalence or base rate. Even assuming that lay interview diagnoses are highly accurate in terms of sensitivity and specificity and their corresponding area under the ROC curve (that is, AUC, or area under the receiver operating characteristic curve), a condition with a relatively low prevalence or base-rate is bound to yield high false positive rates, which exceed false negative rates; in such a circumstance a limited positive predictive value, PPV, yields high false positive rates even in presence of a specificity which is very close to 100 per cent.

NUMBER NEEDED TO TREAT

The number needed to treat (NNT) is an epidemiological measure used in communicating the effectiveness of a health-care intervention, typically a treatment with medication. The NNT is the average number of patients who need to be treated to prevent one additional bad outcome (*e.g.* the number of patients that need to be treated for one of them to benefit compared with a control in a clinical trial). It is defined as the inverse of the absolute risk reduction, and computed as $1/(I_u - I_e)$, where I_e is the incidence in the treated (exposed) group, and I_u is the incidence in the control (unexposed) group.

A type of effect size, the NNT was described in 1988 by McMaster University's Laupacis, Sackett and Roberts. The ideal NNT is 1, where everyone improves with treatment and no one improves with control. The higher the NNT, the less effective is the treatment.

NNT is similar to number needed to harm (NNH), where NNT usually refers to a therapeutic intervention and NNH to a detrimental effect or risk factor.

Relevance

The NNT is an important measure in pharmacoeconomics. If a clinical endpoint is devastating enough (*e.g.*death, heart attack), drugs with a high NNT may still be indicated in particular situations. If the endpoint is minor, health insurers may decline to reimburse drugs with a high NNT. NNT is significant to consider when comparing possible side effects of a medication against its benefits. For medications with a high NNT, even a small incidence of adverse effects may outweigh the benefits. Even though NNT is an important measure in a clinical trial, it is infrequently included in medical journal articles reporting the results of clinical trials. There are several important problems with the NNT, involving bias and lack of reliable confidence intervals, as well as difficulties in excluding the possibility of no difference between two treatments or groups.

NNT values are time-specific. For example, if a study ran for 5 years and another ran for 1 year, the NNT values would not be directly comparable.

Illustration of Different Values

There are a number of factors that can affect the NNT. Let's say we have a disease, and a pill to treat the disease, that should work over the course of a week.

- I_e is the probability of still having the disease after taking the pill (*i.e.* complement of the probability of being cured after taking the pill). This is the treated (exposed) group.
- I_u is the probability of still having the disease after not taking the pill (*i.e.* complement of the probability of the disease going away by itself). This is the control (unexposed) group, who probably got a placebo pill instead of the real pill.

Description	I_u	I_u	NNT	Interpretation
Perfect drug	0.0	1.0	1.0	Everybody is cured with the pill; nobody without
Very good drug	0.1	0.9	1.25	Ten take the pill; 8 cured by the pill, 1 cured by itself, 1 still sick.
Satisfactory drug	0.3	0.7	2.5	Ten take the pill; 4 cured by the pill, 3 cured by itself, 3 still sick.
High placebo effect	0.4	0.5	10	Ten take the pill; 6 cured but 5 of those would be cured anyway.
Low cure rate	0.8	0.9	10	Ten take the pill, one is cured by the pill, one cured by itself, 8 still have the disease.
Goes away by itself	0.1	0.2	10	Ten take the pill and 9 are cured; but 8 would have been cured anyway.
Counter-productive	0.9	0.8	−10	Ten take the pill, two would have been cured without it, but with the pill, only one is cured, so NNH=10.

Real Life Example

ASCOT-LLA manufacturer-sponsored study addressed the benefit of atorvastatin 10 mg (a cholesterol-lowering drug) in patients with hypertension (high blood pressure) but no previous cardiovascular disease (primary prevention). The trial ran for 3.3 years, and during this period the relative risk of a "primary event" (heart attack) was reduced by 36 per cent (relative risk reduction, RRR). The *absolute* risk reduction (ARR), however, was much smaller, because the study group did not have a very high rate of cardiovascular events over the study period: 2.67 per cent in the control group, compared to 1.65 per cent in the treatment group. Taking atorvastatin for 3.3 years, therefore, would lead to an ARR of only 1.02 per cent (2.67 per cent minus 1.65 per cent). The number needed to treat to prevent one cardiovascular event would then be 98.04 for 3.3 years.

Numerical Example

Example of risk reduction

Experimental group (E)	Control group (C)	Total	
Events (E)	EE = 15	CE = 100	115
Non-events (N)	EN = 135	CN = 150	285
Total subjects (S)	ES = EE + EN = 150	CS = CE + CN = 250	400
Event rate (ER)	EER = EE / ES = 0.1, or 10%	CER = CE / CS = 0.4, or 40%	

Equation	Variable	Abbr.	Value
CER - EER	Absolute risk reduction	ARR	0.3, or 30%
(CER - EER) / CER	Relative risk reduction	RRR	0.75, or 75%
1 / (CER – EER)	Number needed to treat	NNT	3.33
EER / CER	Risk ratio	RR	0.25
(EE / EN) / (CE / CN)	Odds ratio	OR	0.167
(CER - EER) / CER	Preventable fraction among the unexposed	PFu	0.75

SIMPSON'S PARADOX

Simpson's paradox, or the Yule–Simpson effect, is a phenomenon in probability and statistics, in which a trend appears in several different groups of data but disappears or reverses when these groups are combined. It is sometimes given the descriptive title reversal paradox or amalgamation paradox. This result is often encountered in social-science and medical-science statistics and is particularly problematic when frequency data is unduly given causal interpretations. The paradoxical elements disappear when causal relations are brought into consideration. It has been used to try to inform the non-specialist or public audience about the kind of misleading results mis-applied statistics can generate. Martin Gardner wrote a popular account of Simpson's paradox in his March 1976 Mathematical Games column in *Scientific American*.

Edward H. Simpson first described this phenomenon in a technical paper in 1951, but the statisticians Karl Pearson et al., in 1899, and Udny Yule, in 1903, had mentioned similar effects earlier. The name *Simpson's paradox* was introduced by Colin R. Blyth in 1972.

Description

Suppose two people, Lisa and Bart, each edit articles for two weeks. In the first week, Lisa fails to improve the only article she edited, and Bart improves 1 of the 4 articles he edited. In the second week, Lisa improves 3 of 4 articles she edited, while Bart improves the only article he edited.

Period Editor	Week 1	Week 2	Total
Lisa	0/1	3/4	3/5
Bart	1/4	1/1	2/5

Both times Bart improved a higher percentage of articles than Lisa, but the actual number of articles each edited (the bottom number of their ratios, also known as the *sample size*) were not the same for both of them either week. When the totals for the two weeks are added together, Bart and Lisa's work can be judged from an equal sample size; *i.e.*, the total number of articles edited by each. Looked at in this more accurate manner, Lisa's ratio is higher and, therefore, so is her percentage. Also when the two tests are combined using a weighted average, overall, Lisa has improved a much higher percentage than Bart because the quality modifier had a significantly higher percentage. Therefore, like many paradoxes, it only appears to be a paradox because of incorrect assumptions, incomplete or misguided information, or a lack of understanding a particular concept. However, the presupposition has been made that all the articles are of equal weight by some measure.

Period Editor	Week 1 quantity	Week 2 quantity	Total quantity and weighted quality
Lisa	0%	75%	60%
Bart	25%	100%	40%

This imagined paradox is caused when the percentage is provided but not the ratio. In this example, if only the 25 per cent in the first week for Bart was provided but not the ratio (1:4), it would distort the information and so cause the imagined paradox. Even though Bart's percentage is higher for the first and second week, when two weeks of articles is combined, overall Lisa had improved a greater proportion, 60 per cent of the 5 total articles. Lisa's proportional total of articles improved exceeds Bart's total.

Examples

UC Berkeley Gender Bias

One of the best-known examples of Simpson's paradox is a study of gender bias among graduate school admissions to University of California, Berkeley. The admission figures for the fall of 1973 showed that men applying were more likely than women to be admitted, and the difference was so large that it was unlikely to be due to chance.

	Men		Women	
	Applicants	Admitted	Applicants	Admitted
Total	8442	44%	4321	35%

But when examining the individual departments, it appeared that six out of 85 departments were significantly biased against men, whereas only four were significantly biased against women. In fact, the pooled and corrected data showed a "small but statistically significant bias in favour of women." The data from the six largest departments are listed below, the top two departments by number of applicants for each gender italicised.

Department	Men		Women	
	Applicants	Admitted	Applicants	Admitted
A	825	62%	108	82%
B	560	63%	25	68%
C	325	37%	593	34%
D	417	33%	375	35%
E	191	28%	393	24%
F	373	6%	341	7%

The research paper by Bickel et al. concluded that women tended to apply to competitive departments with low rates of admission even among qualified applicants (such as in the English Department), whereas men tended to apply to less-competitive departments with high rates of admission among the qualified applicants (such as in engineering and chemistry).

Kidney Stone Treatment

This is a real-life example from a medical study comparing the success rates of two treatments for kidney stones.

The table below shows the success rates and numbers of treatments for treatments involving both small and large kidney stones, where Treatment A includes all open surgical procedures and Treatment B is percutaneous nephrolithotomy (which involves only a small puncture). The numbers in parentheses indicate the number of success cases over the total size of the group.

Treatment Stone size	Treatment A	Treatment B
Small stones	Group 1 93% (81/87)	Group 2 87% (234/270)
Large stones	Group 3 73% (192/263)	Group 4 69% (55/80)
Both	78% (273/350)	83% (289/350)

The paradoxical conclusion is that treatment A is more effective when used on small stones, and also when used on large stones, yet treatment B is more effective when considering both sizes at the same time. In this example, the "lurking" variable (or confounding variable) is the severity of the case (represented by the doctors' treatment decision trend of favoring B for less severe cases), which was not previously known to be important until its effects were included.

Which treatment is considered better is determined by an inequality between two ratios (successes/total). The reversal of the inequality between the ratios, which creates Simpson's paradox, happens because two effects occur together:

- The sizes of the groups, which are combined when the lurking variable is ignored, are very different. Doctors tend to give the severe cases (large stones) the better treatment (A), and the milder cases (small stones) the inferior treatment (B). Therefore, the totals are dominated by groups 3 and 2, and not by the two much smaller groups 1 and 4.
- The lurking variable has a large effect on the ratios; *i.e.*, the success rate is more strongly influenced by the severity of the case than by the choice of treatment. Therefore, the group of patients with large stones using treatment A (group 3) does worse than the group with small stones (groups 1 and 2), even if the latter used the inferior treatment B (group 2).

Based on these effects, the paradoxical result is seen to arise by suppression of the causal effect of the severity of the case on successful treatment. The paradoxical result can be rephrased more accurately as follows: When the less effective treatment (B) is applied more frequently to less severe cases, it can appear to be a more effective treatment.

Batting Averages

A common example of Simpson's Paradox involves the batting averages of players in professional baseball. It is possible for one player to have a higher batting average than another player each year for a number of years, but to have a lower batting average across all of those years. This phenomenon can occur when there are large differences in the number of at-bats between the years. (The same situation applies to calculating batting averages for the first half of the baseball season, and during the second half, and then combining all of the data for the season's batting average.)

Table. A real-life example is provided by Ken Ross and involves the batting average of two baseball players, Derek Jeter and David Justice, during the years 1995 and 1996.

Year Batter	1995		1996		Combined	
Derek Jeter	12/48	.250	183/582	.314	195/630	.310
David Justice	104/411	.253	45/140	.321	149/551	.270

In both 1995 and 1996, Justice had a higher batting average (in bold type) than Jeter did. However, when the two baseball seasons are combined, Jeter shows a higher batting average than Justice. According to Ross, this phenomenon would be observed about once per year among the possible pairs of interesting baseball players.

Correlation between Variables

Simpson's paradox can also arise in correlations, in which two variables appear to have (say) a positive correlation towards one another, when in fact they have a negative correlation, the reversal having been brought about by a "lurking" confounder. Berman et al. give an example from economics, where a dataset suggests overall demand is positively correlated with price (that is, higher prices lead to *more* demand), in contradiction of expectation. Analysis reveals time to be the confounding variable: plotting both price and demand against time reveals the expected negative correlation over various periods, which then reverses to become positive if the influence of time is ignored by simply plotting demand against price.

Implications for Decision Making

he practical significance of Simpson's paradox surfaces in decision making situations where it poses the following dilemma: Which data should we consult in choosing an action, the aggregated or the partitioned? In the Kidney Stone example above, it is clear that if one is diagnosed with "Small Stones" or "Large Stones" the data for the respective subpopulation should be consulted and Treatment A would be preferred to Treatment B. But what if a patient is not diagnosed, and the size of the stone is not known; would it be appropriate to consult the aggregated data and administer Treatment B? This would stand contrary to common sense; a treatment that is preferred both under one condition and under its negation should also be preferred when the condition is unknown. On the other hand, if the partitioned data is to be preferred a priori, what prevents one from partitioning the data into arbitrary sub-categories (say based on eye colour or post-treatment pain) artificially constructed to yield wrong choices of treatments? Pearl shows that, indeed, in many cases it is the aggregated, not the partitioned data that gives the correct choice of action. Worse yet, given the same table, one should sometimes follow the partitioned and sometimes the aggregated data, depending on the story behind the data, with each story dictating its own choice.

Pearl considers this to be the real paradox behind Simpson's reversal. As to why and how a story, not data, should dictate choices, the answer is that it is the story which encodes the causal relationships among the variables. Once we explicate these relationships and represent them formally, we can test which partition gives the correct treatment preference. For example, if we represent causal relationships in a graph called "causal diagram", we can test whether nodes that represent the proposed partition intercept spurious paths in the diagram. This test, called "back-door," reduces Simpson's paradox to an exercise in graph theory.

Psychology

Psychological interest in Simpson's paradox seeks to explain why people deem sign reversal to be impossible at first, offended by the idea that an action preferred both under one condition and under its negation should be rejected when the condition is unknown. The question is where people get this strong intuition from, and how it is encoded in the mind.

Simpson's paradox demonstrates that this intuition cannot be derived from either classical logic or probability calculus alone, and thus led philosophers to speculate that it is supported by an innate causal logic that guides people in reasoning about actions and their consequences. Savage's sure-thing principle is an example of what such logic may entail. A qualified version of Savage's sure thing principle can indeed be derived from Pearl's *do*-calculus and reads: "An action A that increases the probability of an event B in each subpopulation C_i of C must also increase the probability of B in the population as a whole, provided that the action does not change the distribution of the subpopulations." This suggests that knowledge about actions and consequences is stored in a form resembling Causal Bayesian Networks.

Probability

A paper by Pavlides and Perlman presents a proof, due to Hadjicostas, that in a random 2 × 2 × 2 table with uniform distribution, Simpson's paradox will occur with a probability of exactly $/_{60}$. A study by Kock suggests that the probability that Simpson's paradox would occur at random in path models (*i.e.*, models generated by path analysis) with two predictors and one criterion variable is approximately 12.8 percent; slightly higher than 1 occurrence per 8 path models.

ASSOCIATION

Odds ratio

The odds ratio (OR) is a statistic defined as the ratio of the odds of A in the presence of B and the odds of A without the presence of B. This statistic attempts to quantify the strength of the association between A and B. If the OR is greater than 1, then A is considered to be associated with B in the sense that, compared to the absence of B, the presence of B raises the odds of A. Note that this does not establish that B causes A. Often the odds ratio is used to compare the occurrence of some outcome (A) in the presence of some exposure (B), with the occurrence of the outcome (A) in the absence of a particular exposure (absence of B). Two similar statistics that are often used to quantify associations are the risk ratio (RR) and the absolute risk reduction (ARR). Often, the parameter of greatest interest is actually the RR, which is the ratio of the probabilities analogous to the odds used in the OR. However, available data frequently do not allow for the computation of the RR or the ARR but do allow for the computation of the OR, as in case-control studies, as explained below. On the other hand, if one of the properties (A or B) is sufficiently rare (in epidemiology this is called the rare disease assumption), then the OR is approximately equal to the corresponding RR. The OR plays an important role in logistic regression.

Use in Quantitative Research

Due to the widespread use of logistic regression, the odds ratio is widely used in many fields of medical and social science research. The odds ratio is commonly used in survey research, in epidemiology, and to express the results of some clinical trials, such as in case-control studies. It is often abbreviated "OR" in reports. When data from multiple surveys is combined, it will often be expressed as "pooled OR".

Invertibility and Invariance

The odds ratio has another unique property of being directly mathematically invertible whether analyzing the OR as either disease survival or disease onset incidence – where the OR for survival is direct reciprocal of 1/OR for risk.

This is known as the 'invariance of the odds ratio'. In contrast, the relative risk does not possess this mathematical invertible property when studying disease survival vs. onset incidence. This phenomenon of OR invertibility vs. RR non-invertibility is best illustrated with an example: Suppose in a clinical trial, one has an adverse event risk of 4/100 in drug group, and 2/100 in placebo... yielding a RR=2 and OR=2.04166 for drug-vs-placebo adverse risk. However, if analysis was inverted and adverse events were instead analyzed as event-free survival, then the drug group would have a rate of 96/100, and placebo group would have a rate of 98/100—yielding a drug-vs-placebo a RR=0.9796 for survival, but an OR=0.48979.

This is again what is called the 'invariance of the odds ratio', and why a RR for survival is not the same as a RR for risk, while the OR has this symmetrical property when analyzing either survival or adverse risk. The danger to clinical interpretation for the OR comes when the adverse event rate is not rare, thereby exaggerating

differences when the OR rare-disease assumption is not met. On the other hand, when the disease is rare, using a RR for survival (*e.g.* the RR=0.9796 from above example) can clinically hide and conceal an important doubling of adverse risk associated with a drug or exposure.

Alternative Estimators of the Odds Ratio

The sample odds ratio $n_{11}n_{00}/\ n_{10}n_{01}$ is easy to calculate, and for moderate and large samples performs well as an estimator of the population odds ratio. When one or more of the cells in the contingency table can have a small value, the sample odds ratio can be biased and exhibit high variance. A number of alternative estimators of the odds ratio have been proposed to address this issue. One alternative estimator is the conditional maximum likelihood estimator, which conditions on the row and column margins when forming the likelihood to maximize (as in Fisher's exact test). Another alternative estimator is the Mantel-Haenszel estimator.

Numerical Examples

The following four contingency tables contain observed cell counts, along with the corresponding sample odds ratio (*OR*) and sample log odds ratio (*LOR*):

	OR = 1, LOR = 0		OR = 1, LOR = 0		OR = 4, LOR = 1.39		OR = 0.25, LOR = −1.39	
	Y = 1	Y = 0	Y = 1	Y = 0	Y = 1	Y = 0	Y = 1	Y = 0
X = 1	10	10	100	100	20	10	10	20
X = 0	5	5	50	50	10	20	20	10

The following joint probability distributions contain the population cell probabilities, along with the corresponding population odds ratio (*OR*) and population log odds ratio (*LOR*):

	OR = 1, LOR = 0		OR = 1, LOR = 0		OR = 16, LOR = 2.77		OR = 0.67, LOR = −0.41	
	Y = 1	Y = 0	Y = 1	Y = 0	Y = 1	Y = 0	Y = 1	Y = 0
X = 1	0.2	0.2	0.4	0.4	0.4	0.1	0.1	0.3
X = 0	0.3	0.3	0.1	0.1	0.1	0.4	0.2	0.4

HAZARD RATIO

In survival analysis, the hazard ratio (HR) is the ratio of the hazard rates corresponding to the conditions described by two levels of an explanatory variable. For example, in a drug study, the treated population may die at twice the rate per unit time as the control population. The hazard ratio would be 2, indicating higher hazard of death from the treatment. Or in another study, men receiving the same treatment may suffer a certain complication ten times more frequently per unit time than women, giving a hazard ratio of 10.

Hazard ratios differ from relative risks and odds ratios in that RRs and ORs are cumulative over an entire study, using a defined endpoint, while HRs represent instantaneous risk over the study time period, or some subset thereof. Hazard ratios suffer somewhat less from selection bias with respect to the endpoints chosen and can indicate risks that happen before the endpoint.

Interpretation

In its simplest form, the hazard ratio can be interpreted as the chance of an event occurring in the treatment arm divided by the chance of the event occurring in the control arm, or vice versa, of a study. The resolution of these endpoints are usually depicted using Kaplan–Meier survival curves. These curves relate the proportion of each group where the endpoint has not been reached. The endpoint could be any dependent variable associated with the covariate (independent variable), *e.g.* death, remission of disease or contraction of disease. The curve

represents the odds of an endpoint having occurred at each point in time (the hazard). The hazard ratio is simply the relationship between the instantaneous hazards in the two groups and represents, in a single number, the magnitude of distance between the Kaplan–Meier plots. Hazard ratios do not reflect a time unit of the study. The difference between hazard-based and time-based measures is akin to the difference between the odds of winning a race and the margin of victory. When a study reports one hazard ratio per time period, it is assumed that difference between groups was proportional. Hazard ratios become meaningless when this assumption of proportionality is not met. If the proportional hazard assumption holds, a hazard ratio of one means equivalence in the hazard rate of the two groups, whereas a hazard ratio other than one indicates difference in hazard rates between groups. The researcher indicates the probability of this sample difference being due to chance by reporting the probability associated with some test statistic. For instance, the β from the Cox-model or the log-rank test might then be used to assess the significance of any differences observed in these survival curves. Conventionally, probabilities lower than 0.05 are considered significant and researchers provide a 95 per cent confidence interval for the hazard ratio, *e.g.* derived from the standard deviation of the Cox-model regression coefficient, *i.e.* β. Statistically significant hazard ratios cannot include unity (one) in their confidence intervals.

The Proportional Hazards Assumption

The proportional hazards assumption for hazard ratio estimation is strong and often unreasonable. Complications, adverse effects and late effects are all possible causes of change in the hazard rate over time. For instance, a surgical procedure may have high early risk, but excellent long term outcomes. If the hazard ratio between groups remain constant, this is not a problem for interpretation. However, interpretation of hazard ratios become impossible when selection bias exists between group. For instance, a particularly risky surgery might result in the survival of a systematically more robust group who would have fared better under any of the competing treatment conditions, making it look as if the risky procedure was better. Follow-up time is also important. A cancer treatment associated with better remission rates, might on follow-up be associated with higher relapse rates. The researchers' decision about when to follow up is arbitrary and may lead to very different reported hazard ratios.

The Hazard Ratio, Treatment Effect and Time-based Endpoints

Treatment effect depends on the underlying disease related to survival function, not just the hazard ratio. Since the hazard ratio does not give us direct time-to-event information, researchers have to report median endpoint times and calculate the median endpoint time ratio by dividing the control group median value by the treatment group median value. While the median endpoint ratio is a relative speed measure, the hazard ratio is not. The relationship between treatment effect and the hazard ratio is given as e^{β}. A statistically important, but practically insignificant effect can produce a large hazard ratio, *e.g.* a treatment increasing the number of one-year survivors in a population from one in 10,000 to one in 1,000 has a hazard ratio of 10. It is unlikely that such a treatment would have had much impact on the median endpoint time ratio, which likely would have been close to unity, *i.e.* mortality was largely the same regardless of group membership and clinically insignificant. By contrast, a treatment group in which 50 per cent of infections are resolved after one week (versus 25 per cent in the control) yields a hazard ratio of two. If it takes ten weeks for all cases in the treatment group and half of cases in the control group to resolve, the ten-week hazard ratio remains at two, but the median

POPULATION IMPACT

Population Impact Measures (PIMs) are biostatistical measures of risk and benefit used in epidemiological and public health research. They are used to describe the impact of health risks and benefits in a population, to

inform health policy. Frequently used measures of risk and benefit identified by Jerkel, Katz and Elmore, describe measures of risk difference (attributable risk), rate difference (often expressed as the odds ratio or relative risk), Population Attributable Risk (PAR), and the relative risk reduction, which can be recalculated into a measure of *absolute benefit*, called the Number needed to treat. Population Impact Measures are an extension of these statistics, as they are measures of absolute risk at the population level, which are calculations of number of people in the population who are at risk to be harmed, or who will benefit from Public Health interventions. They are measures of absolute risk and benefit, producing numbers of people who will benefit from an intervention or be at risk from a risk factor within a particular local or national population. They provide local context to previous measures, allowing policy-makers to identify and prioritise the potential benefits of interventions on their own population. They are simple to compute, and contain the elements to which policy-makers would have to pay attention in the commissioning or improvement of services. They may have special relevance for local policy-making. They depend on the ability to obtain and use local data, and by being explicit about the data required may have the added benefit of encouraging the collection of such data.

THE MEASURES

To describe the impact of preventive and treatment interventions, the Number of Events Prevented in a Population (NEPP) is defined as *"the number of events prevented by the intervention in your population over a defined time period"*. NEPP extends the well-known measure Number needed to treat (NNT) beyond the individual patient to the population. To describe the impact of a risk factor on causing ill health and disease the Population Impact Number of Eliminating a Risk factor (PIN-ER-t) is defined as *"the potential number of disease events prevented in a population over the next t years by eliminating a risk factor"*. The PIN-ER-t extends the well-known Population Attributable Risk (PAR) to a particular population and relates it to disease incidence, converting the PAR from a measure of relative to absolute risk.

The components for the calculations are as follows: Population denominator (size of the population); Proportion of the population with the disease; Proportion of the population exposed to the risk factor or the incremental proportion of the diseased population eligible for the proposed intervention (the latter requires the actual or estimated proportion who are currently receiving the interventions 'subtracted' from best practice goal from guidelines or targets, adjusted for likely compliance with the intervention); Baseline risk – the probability of the outcome of interest in this or similar populations; and Relative Risk of outcome given exposure to a risk factor or Relative Risk Reduction associated with the intervention.

NUMBER OF EVENTS PREVENTED IN YOUR POPULATION; NEPP

The formula is: NEPP=N*Pd*Pe*ru*RRR where: N = population size, Pd = prevalence of the disease, Pe = proportion eligible for treatment, ru = risk of the event of interest in the untreated group or baseline risk over appropriate time period (can be multiplied by life expectancy to produce life-years), RRR = relative risk reduction associated with treatment. In order to reflect the incremental effect of changing from current to 'best' practice, and to adjust for levels of compliance, the proportion eligible for treatment, Pe, is (Pb-Pt)*Pc, where Pt is the proportion currently treated, Pb is the proportion that would be treated if best practice was adopted, and Pc is the proportion of the population who are compliant with the intervention.

POPULATION IMPACT NUMBER OF ELIMINATING A RISK FACTOR; PIN-ER-T

The formula is: PIN-ER-t = N*Ip*PAR Where: N is the number of people in the population; Ip the baseline risk of the outcome of interest in the population as a whole; t is the time period over which the outcome is

measured. The PAR/F, Population Attributable Risk (or Fraction), is calculated for two or multiple strata. The basic formula to compute the PAR for dichotomous variables is PAR = Pe*(RR-1)/1+ Pe*(RR-1). Where: Pe is the prevalence of the population within each income stratum as the exposure, and RR is the prevalence of risk factors in each stratum relative to the highest income fifth. This is modified where there are multiple strata to: PAR = [Pe1(RR1-1)+Pe2(RR2-1)+Pe3(RR3-1)…]/[1+Pe1(RR1-1)+Pe2(RR2-1)+ Pe3(RR3-1)...]. You can calculate the PIN-ER-t and its confidence intervals at this web site from the Population Health Decision Support and Simulation site.

OTHER

Clinical endpoint

In a clinical research trial, a clinical endpoint generally refers to occurrence of a disease, symptom, sign or laboratory abnormality that constitutes one of the target outcomes of the trial, but may also refer to any such disease or sign that strongly motivates the withdrawal of that individual or entity from the trial, then often termed *humane (clinical) endpoint*. The primary endpoint of a clinical trial is the endpoint for which subjects are randomized and for which the trial is powered. Secondary endpoints are endpoints that are analyzed *post hoc*, for which the trial may not be powered nor randomized.

Scope

In a general sense, a clinical endpoint is included in the entities of interest in a trial. The results of a clinical trial generally indicate the number of people enrolled who reached the pre-determined clinical endpoint during the study interval compared with the overall number of people who were enrolled. Once a patient reaches the endpoint, he or she is generally excluded from further experimental intervention (the origin of the term *endpoint*). For example, a clinical trial investigating the ability of a medication to prevent heart attack might use *chest pain* as a clinical endpoint. Any patient enrolled in the trial who develops chest pain over the course of the trial, then, would be counted as having reached that clinical endpoint. The results would ultimately reflect the fraction of patients who reached the endpoint of having developed chest pain, compared with the overall number of people enrolled. When an experiment involves a control group, the proportion of individuals who reach the clinical endpoint after an intervention is compared with the proportion of individuals in the control group who reached the same clinical endpoint, reflecting the ability of the intervention to prevent the endpoint in question.

A clinical trial will usually define or specify a *primary endpoint* as a measure that will be considered success of the therapy being trialled (*e.g.* in justifying a marketing approval). The primary endpoint might be a statistically significant improvement in *overall survival* (OS). A trial might also define one or more *secondary endpoints* such as *progression-free-survival* (PFS) that will be measured and are expected to be met. A trial might also define *exploratory endpoints* that are less likely to be met.

Examples

Clinical endpoints can be obtained from different modalities, such as, behavioural or cognitive scores, or biomarkers from Electroencephalography (qEEG), MRI, PET, or biochemical biomarkers. In clinical cancer research, common endpoints include discovery of local recurrence, discovery of regional metastasis, discovery of distant metastasis, onset of symptoms, hospitalization, increase or decrease in pain medication requirement, onset of toxicity, requirement of salvage chemotherapy, requirement of salvage surgery, requirement of salvage radiotherapy, death from any cause or death from disease. A cancer study may be powered for overall survival, usually indicating time until death from any cause, or disease specific survival, where the endpoint is death

from disease or death from toxicity. These are expressed as a period of time (survival duration) *e.g.*, in months. Frequently the median is used so that the trial endpoint can be calculated once 50 per cent of subjects have reached the endpoint, whereas calculation of an arithmetical mean can only be done after all subjects have reached the endpoint.

Disease free Survival

The disease free survival is usually used to analyze the results of the treatment for the localized disease which renders the patient apparently disease free, such as surgery or surgery plus adjuvant therapy. In the disease-free survival, the event is relapse rather than death.

The people who relapse are still surviving but they are no longer disease-free. Just as in the survival curves not all patients die, in "disease-free survival curves" not all patients relapse and the curve may have a final plateau representing the patients who didn't relapse after the study's maximum follow-up. Because the patients survive for at least some time after the relapse, the curve for the actual survival would look better than disease free survival curve.

Progression free Survival

The Progression Free Survival is usually used in analysing the results of the treatment for the advanced disease. The event for the progression free survival is that the disease gets worse or progresses, or the patient dies from any cause. *Time to Progression* is a similar endpoint that ignores patients who die before the disease progresses.

Response Duration

The response duration is occasionally used to analyze the results of the treatment for the advanced disease. The event is progression of the disease (relapse). This endpoint involves selecting a subgroup of the patients. It measures the length of the response in those patients who responded. The patients who don't respond aren't included.

Overall survival

Overall survival is based on death from any cause, not just the condition being treated, thus it picks up death from side effects of the treatment, and effects on survival after relapse.

Humane Endpoint

A humane endpoint can be defined as the point at which pain and/or distress is terminated, minimized or reduced for an entity in a trial (such as an experimental animal), by taking action such as killing the animal humanely, terminating a painful procedure, or giving treatment to relieve pain and/or distress. The occurrence of an individual in a trial having reached may necessitate withdrawal from the trial before the target outcome of interest has been fully reached.

Surrogate Endpoint

A surrogate endpoint (or *marker*) is a measure of effect of a certain treatment that may correlate with a *real* clinical endpoint but doesn't necessarily have a guaranteed relationship. The National Institutes of Health (USA) define surrogate endpoint as "a biomarker intended to substitute for a clinical endpoint".

Combined Endpoint

Some studies will examine the incidence of a *combined endpoint*, which can merge a variety of outcomes into one group. For example, the heart attack study above may report the incidence of the *combined endpoint* of chest pain, myocardial infarction, or death. An example of a cancer study powered for a combined endpoint is disease-free survival (DFS); trial participants experiencing either death or discovery of any recurrence would constitute the endpoint. Overall Treatment Utility is an example of a multidimensional composite endpoint in cancer clinical trials. Regarding humane endpoints, a combined endpoint may constitute a threshold where there is enough cumulative degree of disease, symptoms, signs or laboratory abnormalities to motivate an intervention.

Response Rates

Each trial may define what is considered a complete response (CR) or partial response (PR) to the therapy or intervention. Hence the trials report the 'complete response rate' and the overall response rate which includes CR and PR.

Consistency

Various studies on a particular topic often do not address the same outcomes, making it difficult to draw clinically useful conclusions when a group of studies is looked at as a whole. The Core Outcomes in Women's Health (CROWN) Initiative is one effort to standardize outcomes.

VIRULENCE

Virulence is a pathogen's or microbe's ability to infect or damage a host. In the context of gene for gene systems, often in plants, virulence refers to a pathogen's ability to infect a resistant host. In most other contexts, especially in animal systems, virulence refers to the degree of damage caused by a microbe to its host. The pathogenicity of an organism - its ability to cause disease - is determined by its virulence factors. The noun *virulence* derives from the adjective *virulent*. *Virulent* can describe either disease severity or a pathogen's infectivity. The word *virulent* derives from the Latin word *virulentus*, meaning "a poisoned wound" or "full of poison." In ecology, virulence is the host's parasite-induced loss of fitness. Virulence can be understood in terms of proximate causes—those specific traits of the pathogen that help make the host ill—and ultimate causes—the evolutionary pressures that lead to virulent traits occurring in a pathogen strain.

Virulent Bacteria

The ability of bacteria to cause disease is described in terms of the number of infecting bacteria, the route of entry into the body, the effects of host defence mechanisms, and intrinsic characteristics of the bacteria called virulence factors. Many virulence factors are so-called effector proteins that are injected into the host cells by special secretion machines such as the type 3 secretion system. Host-mediated pathogenesis is often important because the host can respond aggressively to infection with the result that host defence mechanisms do damage to host tissues while the infection is being countered. The virulence factors of bacteria are typically proteins or other molecules that are synthesized by enzymes. These proteins are coded for by genes in chromosomal DNA, bacteriophage DNA or plasmids. Certain bacteria employ mobile genetic elements and horizontal gene transfer. Therefore, strategies to combat certain bacterial infections by targeting these specific virulence factors and mobile genetic elements have been proposed. Bacteria use quorum sensing to synchronise release of the molecules. These are all proximate causes of morbidity in the host.

Methods by which Bacteria Cause Disease

- Adhesion. Many bacteria must first bind to host cell surfaces. Many bacterial and host molecules that are involved in the adhesion of bacteria to host cells have been identified. Often, the host cell receptors for bacteria are essential proteins for other functions. Due to presence of mucous lining and of anti-microbial substances around some host cells, it is difficult for certain pathogens to establish direct contact-adhesion.
- Colonization. Some virulent bacteria produce special proteins that allow them to colonize parts of the host body. *Helicobacter pylori* is able to survive in the acidic environment of the human stomach by producing the enzyme urease. Colonization of the stomach lining by this bacterium can lead to Gastric ulcer and cancer. The virulence of various strains of *Helicobacter pylori* tends to correlate with the level of production of urease.
- Invasion. Some virulent bacteria produce proteins that either disrupt host cell membranes or stimulate their own endocytosis or macro-pinocytosis into host cells. These virulence factors allow the bacteria to enter host cells and facilitate entry into the body across epithelial tissue layers at the body surface.
- Immune response inhibitors. Many bacteria produce virulence factors that inhibit the host's immune system defenses. For example, a common bacterial strategy is to produce proteins that bind host antibodies. The polysaccharide capsule of *Streptococcus pneumoniae* inhibits phagocytosis of the bacterium by host immune cells.
- Toxins. Many virulence factors are proteins made by bacteria that poison host cells and cause tissue damage. For example, there are many food poisoning toxins produced by bacteria that can contaminate human foods. Some of these can remain in "spoiled" food even after cooking and cause illness when the contaminated food is consumed. Some bacterial toxins are chemically altered and inactivated by the heat of cooking.

Virulent Viruses

Virus virulence factors allow it replicate, modify host defenses, allow it to spread within the host, and are toxic to the host. They determine whether infection occurs and how severe the resulting viral disease symptoms are. Viruses often require receptor proteins on host cells to which they specifically bind. Typically, these host cell proteins are endocytosed and the bound virus then enters the host cell. Virulent viruses such as HIV, which causes AIDS, have mechanisms for evading host defenses. HIV infects T-Helper Cells, which leads to a reduction of the adaptive immune response of the host and eventually leads to an immunocompromised state. Death results from opportunistic infections secondary to disruption of the immune system caused by AIDS. Some viral virulence factors confer ability to replicate during the defensive inflammation responses of the host such as during virus-induced fever. Many viruses can exist inside a host for long periods during which little damage is done. Extremely virulent strains can eventually evolve by mutation and natural selection within the virus population inside a host. The term "neurovirulent" is used for viruses such as rabies and herpes simplex which can invade the nervous system and cause disease there. Extensively studied model organisms of virulent viruses include virus T4 and other T-even bacteriophages which infect Escherichia coli and a number of related bacteria. The lytic life cycle of virulent bacteriophages is contrasted by the temperate lifecycle of Temperate bacteriophages.

Evolution

According to evolutionary medicine, optimal virulence increases with horizontal transmission (between non-relatives) and decreases with vertical transmission (from parent to child). This is because the fitness of the host is bound to the fitness in vertical transmission but is not so bound in horizontal transmission.

INFECTIVITY

In epidemiology, infectivity is the ability of a pathogen to establish an infection. More specifically, infectivity is a pathogen's capacity for horizontal transmission that is, how frequently it spreads among hosts that are not in a parent-child relationship. The measure of infectivity in a population is called incidence. Infectivity has been shown to positively correlate with virulence. This means that as a pathogen's ability to infect a greater number of hosts increases, so does the level of harm it brings to the host. A pathogen's infectivity is subtly but importantly different from its transmissibility, which refers to a pathogen's capacity to pass from parent to child.

MORTALITY RATE

Mortality rate, or death rate, is a measure of the number of deaths (in general, or due to a specific cause) in a particular population, scaled to the size of that population, per unit of time. Mortality rate is typically expressed in units of deaths per 1,000 individuals per year; thus, a mortality rate of 9.5 (out of 1,000) in a population of 1,000 would mean 9.5 deaths per year in that entire population, or 0.95 per cent out of the total. It is distinct from "morbidity", which is either the prevalence or incidence of a disease, and also from the incidence rate (the number of newly appearing cases of the disease per unit of time).

In the generic form, mortality rates are calculated as:

$d/p * 10^n$

where d represents the deaths occurring within a given time period and p represents the size of the population in which the deaths occur.

Related Measures of Mortality

Other specific measures of mortality include:

Measures of Mortality
Crude death rate – the total number of deaths per year per 1,000 people. As of 2017 the crude death rate for the whole world is 8.33 per 1,000 (up from 7.8 per 1,000 in 2016) according to the current CIA World Factbook.
Perinatal mortality rate – the sum of neonatal deaths and fetal deaths (stillbirths) per 1,000 births.
Maternal mortality ratio – the number of maternal deaths per 100,000 live births in same time period.
Maternal mortality rate – the number of maternal deaths per 1,000 women of reproductive age in the population (generally defined as 15–44 years of age).
Infant mortality rate – the number of deaths of children less than 1 year old per 1,000 live births.
Child mortality rate: the number of deaths of children less than 5 years old per 1,000 live births.
Standardized mortality ratio (SMR) – a proportional comparison to the numbers of deaths that would have been expected if the population had been of a standard composition in terms of age, gender, etc.
Age-specific mortality rate (ASMR) – the total number of deaths per year per 1,000 people of a given age (e.g. age 62 last birthday).
Cause-specific mortality rate – the mortality rate for a specified cause of death.
Cumulative death rate: a measure of the (growing) proportion of a group that die over a specified period (often as estimated by techniques that account for missing data by statistical censoring).
Case fatality rate (CFR) – the proportion of cases of a particular medical condition that lead to death.
Sex-specific mortality rate - Total number of deaths in a population of a specific sex within a given time interval

Use in Epidemiology

In most cases, there are few ways, if at all possible to obtain exact mortality rates, so epidemiologists use estimation to predict correct mortality rates. Mortality rates are usually difficult to predict due to language

barriers, health infrastructure related issues, conflict, and other reasons. Maternal mortality has additional challenges, especially as they pertain to stillbirths, abortions, and multiple births. In some countries, during the 1920s a stillbirth was defined as "a birth of at least twenty weeks' gestation in which the child shows no evidence of life after complete birth". In most countries, however, a stillbirth was defined as "the birth of a fetus, after 28 weeks of pregnancy, in which pulmonary respiration does not occur".

Census Data and Vital Statistics

Ideally, all mortality estimation would be done using vital statistics and census data. Census data will give detailed information about the population at risk of death. The vital statistics provide information about live births and deaths in the population. Often, either census data and vital statistics data is not available. This is especially true in developing countries, countries that are in conflict, areas where natural disasters have caused mass displacement, and other areas where there is a humanitarian crisis.

Household Surveys

Household surveys or interviews are another way in which mortality rates are often assessed. There are several methods to estimate mortality in different segments of the population. One such example is the sisterhood method. This technique involves researchers estimating maternal mortality by contacting women in populations of interest and asking whether or not they have a sister, if the sister is of child-rearing age (usually 15) and conducting an interview or written questions about possible deaths among sisters. The sisterhood method, however, does not work in cases where sisters may have died before the sister being interviewed was born.

Orphanhood surveys estimate mortality by questioning children are asked about the mortality of their parents. It has often been criticized as an adult mortality rate that is very biased for several reasons. The adoption effect is one such instance in which orphans often do not realize that they are adopted. Additionally, interviewers may not realize that an adoptive or foster parent is not the child's biological parent. There is also the issue of parents being reported on by multiple children while some adults have no children, thus are not counted in mortality estimates. Widowhood surveys estimate adult mortality by responding to questions about the deceased husband or wife. One limitation of the widowhood survey surrounds the issues of divorce, where people may be more likely to report that they are widowed in places where there is the great social stigma around being a divorcee. Another limitation is that multiple marriages introduce biased estimates, so individuals are often asked about first marriage. Biases will be significant if the association of death between spouses, such as those in countries with large AIDS epidemics.

Sampling

Sampling refers to the selection of a subset of the population of interest to efficiently gain information about the entire population. Samples should be representative of the population of interest. Cluster sampling is an approach to non-probability sampling; this is an approach in which each member of the population is assigned to a group (cluster), and then clusters are randomly selected, and all members of selected clusters are included in the sample. Often combined with stratification techniques (in which case it is called multistage sampling), cluster sampling is the approach most often used by epidemiologists. In areas of forced migration, there is more significant sampling error. Thus cluster sampling is not the ideal choice.

4

Reproducibility, Laboratory and Experimental Methods

Reproducibility is the closeness of the agreement between the results of measurements of the same measur and carried out with same methodology described in the corresponding scientific evidence (*e.g.* a publication in a peer-reviewed journal). Reproducibilty can also be applied under changed conditions of measurement for the same measur and to check, that the results are not an artefact of the measurment procedures. A related concept is replicability, meaning the ability to independently achieve non-identical conclusions that are at least similar, when differences in sampling, research procedures and data analysis methods may exist. Reproducibility and replicability together are among the main beliefs of 'the scientific method'—with the concrete expressions of the ideal of such a method varying considerably across research disciplines and fields of study. The reproduced measurement may be based on the raw data and computer programmes provided by researchers. The values obtained from distinct experimental trials are said to be *commensurate* if they are obtained according to the same reproducible experimental description and procedure. The basic idea can be seen in Aristotle's dictum that there is no scientific knowledge of the individual, where the word used for *individual* in Greek had the connotation of the *idiosyncratic*, or wholly isolated occurrence. Thus all knowledge, all science, necessarily involves the formation of general concepts and the invocation of their corresponding symbols in language (cf. Turner). Aristotle2s conception about the knowledge of the individual being considered unscientific is due to lack of the field of statistics in his time, so he could not appeal to statistical averaging by the individual.

A particular experimentally obtained value is said to be reproducible if there is a high degree of agreement between measurements or observations conducted on replicate specimens in different locations by different people—that is, if the experimental value is found to have a high precision. However, in science, a very well reproduced result is one that can be confirmed using as many *different* experimental setups as possible and as many lines of evidence as possible (consilience).

HISTORY

The first to stress the importance of reproducibility in science was the Irish chemist Robert Boyle, in England in the 17th century. Boyle's air pump was designed to generate and study vacuum, which at the time was a very controversial concept. Indeed, distinguished philosophers such as René Descartes and Thomas Hobbes denied the very possibility of vacuum existence. Historians of science *e.g.* Steven Shapin and Simon Schaffer, in their 1985 book *Leviathan and the Air-Pump*, describe the debate between Boyle and Hobbes, ostensibly over the nature of vacuum, as fundamentally an argument about how useful knowledge should be gained. Boyle, a pioneer of the experimental method, maintained that the foundations of knowledge should be constituted by experimentally produced facts, which can be made believable to a scientific community by their reproducibility.

By repeating the same experiment over and over again, Boyle argued, the certainty of fact will emerge. The air pump, which in the 17th century was a complicated and expensive apparatus to build, also led to one of the first documented disputes over the reproducibility of a particular scientific phenomenon. In the 1660s, the Dutch scientist Christiaan Huygens built his own air pump in Amsterdam, the first one outside the direct management of Boyle and his assistant at the time Robert Hooke. Huygens reported an effect he termed "anomalous suspension", in which water appeared to levitate in a glass jar inside his air pump (in fact suspended over an air bubble), but Boyle and Hooke could not replicate this phenomenon in their own pumps. As Shapin and Schaffer describe, "it became clear that unless the phenomenon could be produced in England with one of the two pumps available, then no one in England would accept the claims Huygens had made, or his competence in working the pump". Huygens was finally invited to England in 1663, and under his personal guidance Hooke was able to replicate anomalous suspension of water. Following this Huygens was elected a Foreign Member of the Royal Society. However, Shapin and Schaffer also note that "the accomplishment of replication was dependent on contingent acts of judgement. One cannot write down a formula saying when replication was or was not achieved".

The philosopher of science Karl Popper noted briefly in his famous 1934 book *The Logic of Scientific Discovery* that "non-reproducible single occurrences are of no significance to science". The Statistician Ronald Fisher wrote in his 1935 book *The Design of Experiments*, which set the foundations for the modern scientific practice of hypothesis testing and statistical significance, that "we may say that a phenomenon is experimentally demonstrable when we know how to conduct an experiment which will rarely fail to give us statistically significant results". Such assertions express a common dogma in modern science that reproducibility is a necessary condition (although not necessarily sufficient) for establishing a scientific fact, and in practice for establishing scientific authority in any field of knowledge. However, as noted above by Shapin and Schaffer, this dogma is not well-formulated quantitatively, such as statistical significance for instance, and therefore it is not explicitly established how many times must a fact be replicated to be considered reproducible.

REPRODUCIBLE DATA

Reproducibility is one component of the precision of a measurement or test method. The other component is repeatability which is the degree of agreement of tests or measurements on replicate specimens by the same observer in the same laboratory. Both repeatability and reproducibility are usually reported as a standard deviation. A reproducibility limit is the value below which the difference between two test results obtained under reproducibility conditions may be expected to occur with a probability of approximately 0.95 (95 per cent).

Reproducibility is determined from controlled interlaboratory test programmes or a measurement systems analysis. Although they are often confused, there is an important distinction between replicates and an independent repetition of an experiment. Replicates are performed within an experiment. They are not and cannot provide independent evidence of reproducibility. Rather they serve as an internal "check" on an experiment and should not be shown as part of the experimental results within a scientific publication. It is the independent repetition of an experiment that serves to underpin its reproducibility.

REPRODUCIBLE RESEARCH

The term *reproducible research* refers to the idea that the ultimate product of academic research is the paper along with the laboratory notebooks and full computational environment used to produce the results in the paper such as the code, data, etc. that can be used to reproduce the results and create new work based on the research. Typical examples of reproducible research comprise compendia of data, code and text files, often organised around an R Markdown source document or a Jupyter notebook. Psychology has seen a renewal of internal concerns about irreproducible results. Researchers showed in a 2006 study that, of 141 authors of a publication from the American Psychology Association (APA) empirical articles, 103 (73 per cent) did not

respond with their data over a 6-month period. In a follow up study published in 2015, it was found that 246 out of 394 contacted authors of papers in APA journals did not share their data upon request (62 per cent). In a 2012 paper, it was suggested that researchers should publish data along with their works, and a dataset was released alongside as a demonstration, in 2017 it was suggested in an article published in *Scientific Data* that this may not be sufficient and that the whole analysis context should be disclosed. In 2015, Psychology became the first discipline to conduct and publish an open, registered empirical study of reproducibility called the Reproducibility Project. 270 researchers from around the world collaborated to replicate 100 empirical studies from three top Psychology journals. Fewer than half of the attempted replications were successful. There have been initiatives to improve reporting and hence reproducibility in the medical literature for many years, which began with the CONSORT initiative, which is now part of a wider initiative, the EQUATOR Network. This group has recently turned its attention to how better reporting might reduce waste in research, especially biomedical research.

Reproducible research is key to new discoveries in pharmacology. A Phase I discovery will be followed by Phase II reproductions as a drug develops towards commercial production. In recent decades Phase II success has fallen from 28 per cent to 18 per cent. A 2011 study found that 65 per cent of medical studies were inconsistent when re-tested, and only 6 per cent were completely reproducible. In 2012, a study by Begley and Ellis was published in *Nature* that reviewed a decade of research. That study found that 47 out of 53 medical research papers focused on cancer research were irreproducible. The irreproducible studies had a number of features in common, including that studies were not performed by investigators blinded to the experimental versus the control arms, there was a failure to repeat experiments, a lack of positive and negative controls, failure to show all the data, inappropriate use of statistical tests and use of reagents that were not appropriately validated. John P. A. Ioannidis writes, "While currently there is unilateral emphasis on 'first' discoveries, there should be as much emphasis on replication of discoveries." The *Nature* study was itself reproduced in the journal *PLOS ONE*, which confirmed that a majority of cancer researchers surveyed had been unable to reproduce a result. In 2016, *Nature* conducted a survey of 1,576 researchers who took a brief online questionnaire on reproducibility in research. According to the survey, more than 70 per cent of researchers have tried and failed to reproduce another scientist's experiments, and more than half have failed to reproduce their own experiments. "Although 52 per cent of those surveyed agree there is a significant 'crisis' of reproducibility, less than 31 per cent think failure to reproduce published results means the result is probably wrong, and most say they still trust the published literature."

NOTEWORTHY IRREPRODUCIBLE RESULTS

Hideyo Noguchi became famous for correctly identifying the bacterial agent of syphilis, but also claimed that he could culture this agent in his laboratory. Nobody else has been able to produce this latter result.

In March 1989, University of Utah chemists Stanley Pons and Martin Fleischmann reported the production of excess heat that could only be explained by a nuclear process ("cold fusion"). The report was astounding given the simplicity of the equipment: it was essentially an electrolysis cell containing heavy water and a palladium cathode which rapidly absorbed the deuterium produced during electrolysis. The news media reported on the experiments widely, and it was a front-page item on many newspapers around the world. Over the next several months others tried to replicate the experiment, but were unsuccessful. Nikola Tesla claimed as early as 1899 to have used a high frequency current to light gas-filled lamps from over 25 miles (40 km) away without using wires. In 1904 he built Wardenclyffe Tower on Long Island to demonstrate means to send and receive power without connecting wires. The facility was never fully operational and was not completed due to economic problems, so no attempt to reproduce his first result was ever carried out.

Other examples which contrary evidence has refuted the original claim:

- Stimulus-triggered acquisition of pluripotency, revealed to be the result of fraud

- GFAJ-1, a bacterium that could purportedly incorporate arsenic into its DNA in place of phosphorus
- MMR vaccine controversy – a study in *The Lancet* claiming the MMR vaccine caused autism was revealed to be fraudulent
- Schön scandal – semiconductor "breakthroughs" revealed to be fraudulent
- Power posing – a social psychology phenomenon that went viral after being the subject of a very popular TED talk, but was unable to be replicated in dozens of studies

STOCHASTIC PROCESSES

The reproducibility requirement cannot be applied to individual samples of phenomena which have a partially or totally non-deterministic nature. However, it still applies to the probabilistic description of such phenomena, with error tolerance given by probability theory.

LABORATORY

A laboratory is a facility that provides controlled conditions in which scientific or technological research, experiments, and measurement may be performed. Laboratories used for scientific research take many forms because of the differing requirements of specialists in the various fields of science and engineering. A physics laboratory might contain a particle accelerator or vacuum chamber, while a metallurgy laboratory could have apparatus for casting or refining metals or for testing their strength. A chemist or biologist might use a wet laboratory, while a psychologist's laboratory might be a room with one-way mirrors and hidden cameras in which to observe behaviour. In some laboratories, such as those commonly used by computer scientists, computers (sometimes supercomputers) are used for either simulations or the analysis of data. Scientists in other fields will use still other types of laboratories. Engineers use laboratories as well to design, build, and test technological devices. Scientific laboratories can be found as research room and learning spaces in schools and universities, industry, government, or military facilities, and even aboard ships and spacecraft.

Despite the underlying notion of the lab as a confined space for experts, the term "laboratory" is also increasingly applied to workshop spaces such as Living Labs, Fab Labs, or Hackerspaces, in which people meet to work on societal problems or make prototypes, working collaboratively or sharing resources. This development is inspired by new, participatory approaches to science and innovation and relies on user-centred design methods and concepts like Open innovation or User innovation,. One distinctive feature of work in Open Labs is phenomena of translation, driven by the different backgrounds and levels of expertise of the people involved.

HISTORY

Early instances of "laboratories" recorded in English involved alchemy and the preparation of medicines. The emergence of Big Science during World War II increased the size of laboratories and scientific equipment, introducing particle accelerators and similar devices.

The Early Laboratories

The earliest laboratory according to the present evidence is a home laboratory of Pythagoras of Samos, the well-known Greek philosopher and scientist. This laboratory was created when Pythagoras conducted an experiment about tones of sound and vibration of string.

In the painting of Louis Pasteur by Albert Edelfelt in 1885, Louis Pasteur is shown comparing a note in his left hand with a bottle filled with a solid in his right hand, and not wearing any personal protective equipment. Researching in teams started in the 19th century, and many new kinds of equipment were developed in the 20th century.

A 16th century underground alchemical laboratory was accidentally discovered in the year 2002. Rudolf II, Holy Roman Emperor was believed to be the owner. The laboratory is called Speculum Alchemiae and is preserved as a museum in Prague.

Fig. Chemistry laboratory of the 18th century, of the sort used by Antoine Lavoisier and his contemporaries.

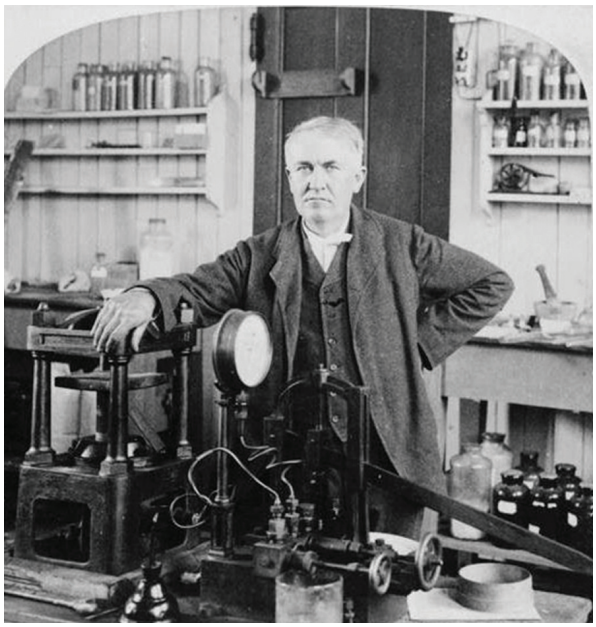


Fig. Thomas Edison in his laboratory, 1901.

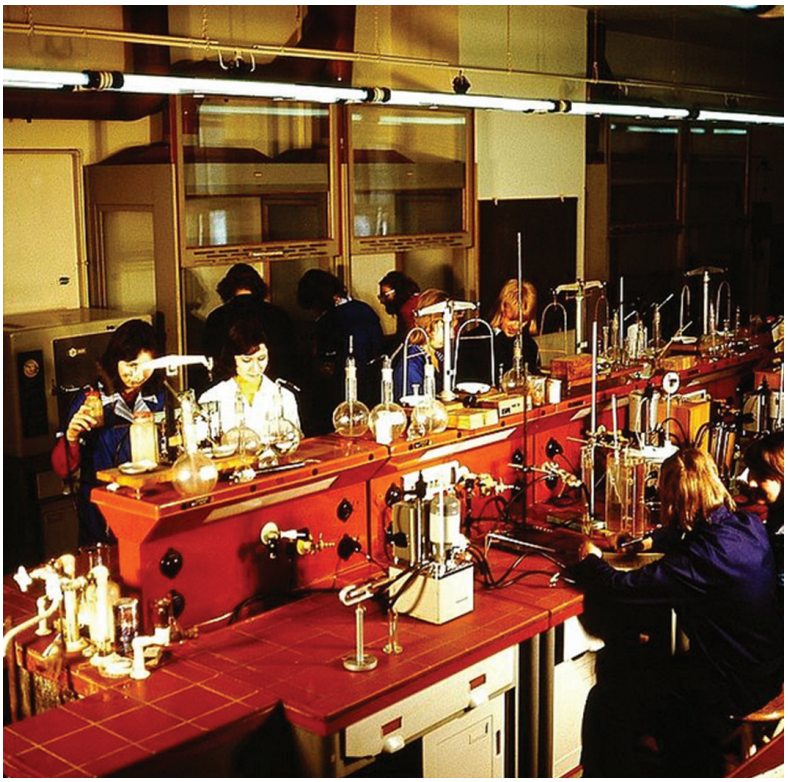


Fig. A laboratory in the 1970s.

Fig. Chemical laboratory in Mahidol University International College since 2009.

Fig. Early 2000s style of counter in Chemical Laboratory, Mahidol University International College, Thailand.

TECHNIQUES

Laboratory techniques are the set of procedures used on natural sciences such as chemistry, biology, physics to conduct an experiment, all of them follow the scientific method; while some of them involve the use of complex laboratory equipment from laboratory glassware to electrical devices, and others require more specific or expensive supplies.

EQUIPMENT AND SUPPLIES

Laboratory equipment refers to the various tools and equipment used by scientists working in a laboratory: The classical equipment includes tools such as Bunsen burners and microscopes as well as specialty equipment such as operant conditioning chambers, spectrophotometers and calorimeters.

Chemical laboratories:

- Laboratory glassware such as the beaker or reagent bottle
- Analytical devices as HPLC or spectrophotometers

Molecular biology laboratories + Life science laboratories:

- Autoclave
- Microscope
- Centrifuges
- Shakers and mixers
- Pipette
- Thermal cyclers (PCR)
- Photometer
- Refrigerators and Freezers
- Universal testing machine
- ULT Freezers
- Incubators
- Bioreactor
- Biological safety cabinets
- Sequencing instruments
- Fume hoods
- Environmental chamber
- Humidifier
- Weighing scale
- Reagents (supply)
- Pipettes tips (supply)
- Polymer (supply) consumables for small volumes (µL and mL scale), mainly sterile

Laboratory equipment is generally used to either perform an experiment or to take measurements and gather data. Larger or more sophisticated equipment is generally called a scientific instrument.

SPECIALIZED TYPES

The title of laboratory is also used for certain other facilities where the processes or equipment used are similar to those in scientific laboratories. These notably include:

- Film laboratory or Darkroom
- Clandestine lab for the production of illegal drugs
- Computer lab

- Crime lab used to process crime scene evidence
- Language laboratory
- Medical laboratory (involves handling of chemical compounds)
- Public health laboratory
- Industrial laboratory

SAFETY

Many laboratories contain significant risks, and the prevention of laboratory accidents requires great care and constant vigilance. Examples of risk factors include high voltages, high and low pressures and temperatures, corrosive and toxic chemicals, and biohazards including infective organisms and their toxins. Measures to protect against laboratory accidents include safety training and enforcement of laboratory safety policies, safety review of experimental designs, the use of personal protective equipment, and the use of the buddy system for particularly risky operations.

In many countries, laboratory work is subject by health and safety legislation. In some cases, laboratory activities can also present environmental health risks, for example, the accidental or deliberate discharge of toxic or infective material from the laboratory into the environment.

Chemical Hazards

Hazardous chemicals present physical and/or health threats to workers in clinical, industrial, and academic laboratories. Laboratory chemicals include cancer-causing agents (carcinogens), toxins (*e.g.*, those affecting the liver, kidney, and nervous system), irritants, corrosives, sensitizers, as well as agents that act on the blood system or damage the lungs, skin, eyes, or mucous membranes.

Biological Hazards

Biological Agents and Biological Toxins

Many laboratory workers encounter daily exposure to biological hazards. These hazards are present in various sources throughout the laboratory such as blood and body fluids, culture specimens, body tissue and cadavers, and laboratory animals, as well as other workers.

These are federally regulated biological agents (*e.g.*, viruses, bacteria, fungi, and prions) and toxins that have the potential to pose a severe threat to public health and safety, to animal or plant health, or to animal or plant products.

- Anthrax - Anthrax is an acute infectious disease caused by a spore-forming bacterium called *Bacillus anthracis*.
- Avian Flu - Avian influenza is caused by *Influenza A viruses*.
- Botulism - Cases of botulism are usually associated with consumption of preserved foods.
- Foodborne Disease - Foodborne illnesses are caused by viruses, bacteria, parasites, toxins, metals, and prions (microscopic protein particles). Symptoms range from mild gastroenteritis to life-threatening neurologic, hepatic and renal syndromes.
- Hantavirus - Hantaviruses are transmitted to humans from the dried droppings, urine, or saliva of mice and rats.
- Legionnaires' Disease - Legionnaires' disease is a bacterial disease commonly associated with water-based aerosols.

- Molds and Fungi - Molds and fungi produce and release millions of spores small enough to be air, water, or insect-borne which may have negative effects on human health including, allergic reactions, asthma, and other respiratory problems.
- Plague - The World Health Organization reports 1,000 to 3,000 cases of plague every year. A bioterrorist release of plague could result in a rapid spread of the pneumonic form of the disease, which could have devastating consequences.
- Ricin - Ricin is one of the most toxic and easily produced plant toxins. It has been used in the past as a bioterrorist weapon and remains a serious threat.
- Smallpox - Smallpox is a highly contagious disease unique to humans. It is estimated that no more than 20 percent of the population has any immunity from previous vaccination.
- Tularemia - Tularemia is also known as "rabbit fever" or "deer fly fever" and is extremely infectious. Relatively few bacteria are required to cause the disease, which is why it is an attractive weapon for use in bioterrorism.

Physical Hazards and others

Besides exposure to chemicals and biological agents, laboratory workers can also be exposed to a number of physical hazards. Some of the common physical hazards that they may encounter include the following: ergonomic, ionizing radiation, non-ionizing radiation and noise hazards.

Ergonomic Hazards

Laboratory workers are at risk for repetitive motion injuries during routine laboratory procedures such as pipetting, working at microscopes, operating microtomes, using cell counters and keyboarding at computer workstations. Repetitive motion injuries develop over time and occur when muscles and joints are stressed, tendons are inflamed, nerves are pinched and the flow of blood is restricted. Standing and working in awkward positions in front of laboratory hoods/biological safety cabinets can also present ergonomic problems.

Ionizing Radiation

Ionizing radiation sources are found in a wide range of occupational settings, including laboratories. These radiation sources can pose a considerable health risk to affected workers if not properly controlled. Any laboratory possessing or using radioactive isotopes must be licensed by the Nuclear Regulatory Commission (NRC) and/or by a state agency that has been approved by the NRC, 10 CFR 31.11 and 10 CFR 35.12.

The fundamental objectives of radiation protection measures are:

- To limit entry of radionuclides into the human body (via ingestion, inhalation, absorption, or through open wounds) to quantities as low as reasonably achievable (ALARA) and always within the established limits;
- To limit exposure to external radiation to levels that are within established dose limits and as far below these limits as is reasonably achievable.

Safety Hazards

Autoclaves and Sterilizers

Workers should be trained to recognize the potential for exposure to burns or cuts that can occur from handling or sorting hot sterilized items or sharp instruments when removing them from autoclaves/sterilizers or from steam lines that service the autoclaves.

Centrifuges

Centrifuges, due to the high speed at which they operate, have great potential for injuring users if not operated properly. Unbalanced centrifuge rotors can result in injury, even death. Sample container breakage can generate aerosols that may be harmful if inhaled. The majority of all centrifuge accidents are the result of user error.

Compressed Gases

Laboratory standard for compressed gas:

- Is a gas or mixture of gases in a container having an absolute pressure exceeding 40 pounds per square inch (psi) at 70 °F (21.1 °C); or
- Is a gas or mixture of gases having an absolute pressure exceeding 104 psi at 130 °F (54.4 °C) regardless of the pressure at 70 °F (21.1 °C); or
- Is a liquid having a vapor pressure exceeding 40 psi at 100 °F (37.8 °C) as determined by ASTM (American Society for Testing and Materials)

Within laboratories, compressed gases are usually supplied either through fixed piped gas systems or individual cylinders of gases. Compressed gases can be toxic, flammable, oxidizing, corrosive, or inert. Leakage of any of these gases can be hazardous.

Store, Handle, and use Compressed Gases

- All cylinders whether empty or full must be stored upright.
- Secure cylinders of compressed gases. Cylinders should never be dropped or allowed to strike each other with force.
- Transport compressed gas cylinders with protective caps in place and do not roll or drag the cylinders.

Cryogens and Dry Ice

Cryogens, substances used to produce very low temperatures [below -153 °C (-243 °F)], such as liquid nitrogen (LN_2) which has a boiling point of -196 °C (-321 °F), are commonly used in laboratories. Although not a cryogen, solid carbon dioxide or dry ice which converts directly to carbon dioxide gas at -78 °C (-109 °F) is also often used in laboratories. Shipments packed with dry ice, samples preserved with liquid nitrogen, and in some cases, techniques that use cryogenic liquids, such as cryogenic grinding of samples, present potential hazards in the laboratory. Hand protection is required to guard against the hazard of touching cold surfaces. It is recommended that Cryogen Safety Gloves be used by the worker. Eye protection is required at all times when working with cryogenic fluids. When pouring a cryogen, working with a wide-mouth Dewar flask or around the exhaust of cold boil-off gas, use of a full face shield is recommended.

Personal Protective Equipments

Personal protective equipment or PPE are equipments worn to prevent against exposure of hazardous substances. Although, PPE does not eliminate the risks of hazards but it helps protect the user from the exposure. To make a workplace safer, it should provide instructions and training of how to use and choose proper PPE in different situations.

PPE includes:

- Long-sleeved shirts, lab coats, aprons.
- Goggles

- Safety gloves;
  (a) There are 2 common types of safety gloves that are widely used in high school or university laboratory, Latex and Nitrile gloves. Latex gloves have a high sensitivity when it comes to contact and fine control which is very suitable for surgery. On the other hands, Nitrile gloves are the gloves that do not have latex protein which cost twice. It was known as the most durable, resisted to tear and many chemicals. Beside all the benefits, Nitrile gloves also have drawbacks since it can oxidize silver and high reactive metals as these metals can react with sulfur. Therefore, wearer should have an extra care while wearing this type of protective gloves.
- Face shield or safety

Electrical

In the laboratory, there is the potential for workers to be exposed to electrical hazards including electric shock, electrocutions, fires and explosions. Damaged electrical cords can lead to possible shocks or electrocutions. A flexible electrical cord may be damaged by door or window edges, by staples and fastenings, by equipment rolling over it, or simply by aging.

The potential for possible electrocution or electric shock or contact with electrical hazards can result from a number of factors, including the following:

- Faulty electrical equipment/instrumentation or wiring;
- Damaged receptacles and connectors; and
- Unsafe work practices.

Fire

Fire is the most common serious hazard that one faces in a typical laboratory. While proper procedures and training can minimize the chances of an accidental fire, laboratory workers should still be prepared to deal with a fire emergency should it occur. In dealing with a laboratory fire, all containers of infectious materials should be placed into autoclaves, incubators, refrigerators, or freezers for containment. Small bench-top fires in laboratory spaces are not uncommon. Large laboratory fires are rare. However, the risk of severe injury or death is significant because fuel load and hazard levels in labs are typically very high. Laboratories, especially those using solvents in any quantity, have the potential for flash fires, explosion, rapid spread of fire, and high toxicity of products of combustion (heat, smoke, and flame).

Glassware

- Broken glass is a hazard for a sharps related injury.
- Correct eye protection should be worn in most experiments involving glassware.
- Inserting a glass rod through a stopper can introduce the possibility of a stab wound or sharps injury if the rod breaks. The hands must be protected.
- Tubing should be cut from a barbed connection so as not to shatter the connection. A quick disconnect is preferable to a barbed fitting.
- Ground glass joints can become a breaking hazard if they freeze.
- Broken and other waste glass should be discarded in a separate container specially marked to indicate its contents.
- Glassware should always be labeled as to its contents.
- Rapid heating (or cooling) may cause uneven thermal expansion putting too much mechanical stress on the surface and cause it to fracture. Fracturing is a concern when people new to laboratory become

impatient and heat glassware, especially the larger pieces, too fast. Heating of glassware should be slowed using an insulating material, such as metal foil or wool, or specialized equipment such as heated baths, heating mantles or laboratory grade hot plates to avoid fracturing.

- Hot glass looks like cold glass, so a person must be careful to avoid grabbing hot glassware.
- Glassware can explode if the exhaust is in any way restricted, so any apparatus should be vented.
- Glassware can implode under negative pressure
- When connecting joints, it is the responsibility of the person overseeing the experiment to select the correct seal. For example, PTFE tape, bands, and fluoroether-based grease or oils may emit toxic perfluoroisobutylene fumes if the rated temperature limits are exceed.

INFORMED CONSENT

Informed consent is a process for getting permission before conducting a health care intervention on a person, or for disclosing personal information. A health care provider may ask a patient to consent to receive therapy before providing it, or a clinical researcher may ask a research participant before enrolling that person into a clinical trial. Informed consent is collected according to guidelines from the fields of medical ethics and research ethics. An informed consent can be said to have been given based upon a clear appreciation and understanding of the facts, implications, and consequences of an action. Adequate informed consent is rooted in respecting a person's dignity. To give informed consent, the individual concerned must have adequate reasoning faculties and be in possession of all relevant facts. Impairments to reasoning and judgement that may prevent informed consent include basic intellectual or emotional immaturity, high levels of stress such as posttraumatic stress disorder (PTSD) or a severe intellectual disability, severe mental disorder, intoxication, severe sleep deprivation, Alzheimer's disease, or being in a coma.

Obtaining informed consent is not always required. If an individual is considered unable to give informed consent, another person is generally authorized to give consent on his behalf, *e.g.*, parents or legal guardians of a child (though in this circumstance the child may be required to provide informed assent) and conservators for the mentally disordered, or consent can be assumed through the doctrine of implied consent, *e.g.*, when an unconscious person will die without immediate medical treatment.

In cases where an individual is provided insufficient information to form a reasoned decision, serious ethical issues arise. Such cases in a clinical trial in medical research are anticipated and prevented by an ethics committee or Institutional Review Board.

Informed Consent Form Templates can be found on the World Health Organization Web site for practical use.

ASSESSMENT

Informed consent can be complex to evaluate, because neither expressions of consent, nor expressions of understanding of implications, necessarily mean that full adult consent was in fact given, nor that full comprehension of relevant issues is internally digested. Consent may be implied within the usual subtleties of human communication, rather than explicitly negotiated verbally or in writing. In some cases consent cannot legally be possible, even if the person protests he does indeed understand and wish. There are also structured instruments for evaluating capacity to give informed consent, although no ideal instrument presently exists. Thus, there is always a degree to which informed consent must be assumed or inferred based upon observation, or knowledge, or legal reliance. This especially is the case in sexual or relational issues. In medical or formal circumstances, explicit agreement by means of signature—normally relied on legally—regardless of actual consent, is the norm. This is the case with certain procedures, such as a "do not resuscitate" directive that a patient signed before onset of their illness.

Brief examples of each of the above:

- A person may verbally agree to something from fear, perceived social pressure, or psychological difficulty in asserting true feelings. The person requesting the action may honestly be unaware of this and believe the consent is genuine, and rely on it. *Consent is expressed, but not internally given.*
- A person may claim to understand the implications of some action, as part of consent, but in fact has failed to appreciate the possible consequences fully and may later deny the validity of the consent for this reason. *Understanding needed for informed consent is present but is, in fact (through ignorance), not present.*
- A person signs a legal release form for a medical procedure, and later feels he did not really consent. Unless he can show actual misinformation, the release is usually persuasive or conclusive in law, in that the clinician may rely legally upon it for consent. *In formal circumstances, a written consent usually legally overrides later denial of informed consent (unless obtained by misrepresentation).*
- Informed consent in the U.S. can be overridden in emergency medical situations pursuant to 21CFR50.24, which was first brought to the general public's attention via the controversy surrounding the study of Polyheme.

VALID ELEMENTS

For an individual to give valid informed consent, three components must be present: disclosure, capacity and voluntariness.

- *Disclosure* requires the researcher to supply each prospective subject with the information necessary to make an autonomous decision and also to ensure that the subject adequate understands the information provided. This latter requirement implies that a written consent form be written in lay language suited for the comprehension skills of subject population, as well as assessing the level of understanding through conversation.
- *Capacity* pertains to the ability of the subject to both understand the information provided and form a reasonable judgement based on the potential consequences of his/her decision.
- *Voluntariness* refers to the subject's right to freely exercise his/her decision making without being subjected to external pressure such as coercion, manipulation, or undue influence.

WAIVER OF REQUIREMENT

Waiver of the consent requirement may be applied in certain circumstances where no foreseeable harm is expected to result from the study or when permitted by law, federal regulations, or if an ethical review committee has approved the non-disclosure of certain information.

Besides studies with minimal risk, waivers of consent may be obtained in a military setting. According to 10 USC 980, the United States Code for the Armed Forces, Limitations on the Use of Humans as Experimental Subjects, a waiver of advanced informed consent may be granted by the Secretary of Defence if a research project would:

- Directly benefit subjects.
- Advance the development of a medical product necessary to the military.
- Be carried out under all laws and regulations (*i.e.*, Emergency Research Consent Waiver) including those pertinent to the FDA.

While informed consent is a basic right and should be carried out effectively, if a patient is incapacitated due to injury or illness, it is still important that patients benefit from emergency experimentation. The Food and Drug Administration (FDA) and the Department of Health and Human Services (DHHS) joined together to create federal guidelines to permit emergency research, without informed consent.

However, they can only proceed with the research if they obtain a waiver of informed consent (WIC) or an emergency exception from informed consent (EFIC).

HISTORY

Informed consent is a technical term first used by attorney, Paul G. Gebhard, in a medical malpractice United States court case in 1957. In tracing its history, some scholars have suggested tracing the history of checking for any of these practices:

- A patient agrees to a health intervention based on an understanding of it.
- The patient has multiple choices and is not compelled to choose a particular one.
- The consent includes giving permission.

These practices are part of what constitutes informed consent, and their history is the history of informed consent. They combine to form the modern concept of informed consent—which rose in response to particular incidents in modern research. Whereas various cultures in various places practiced informed consent, the modern concept of informed consent was developed by people who drew influence from Western tradition.

Medical History

Historians cite a series of medical guidelines to trace the history of informed consent in medical practice.

The Hippocratic Oath, a 500 BC Greek text, was the first set of Western writings giving guidelines for the conduct of medical professionals. It advises that physicians conceal most information from patients to give the patients the best care. The rationale is a beneficence model for care—the doctor knows better than the patient, and therefore should direct the patient's care, because the patient is not likely to have better ideas than the doctor.

Henri de Mondeville, a French surgeon who in the 14th century, wrote about medical practice. He traced his ideas to the Hippocratic Oath. Among his recommendations were that doctors "promise a cure to every patient" in hopes that the good prognosis would inspire a good outcome to treatment. Mondeville never mentioned getting consent, but did emphasize the need for the patient to have confidence in the doctor. He also advised that when deciding therapeutically unimportant details the doctor should meet the patients' requests "so far as they do not interfere with treatment".

Benjamin Rush was an 18th-century United States physician who was influenced by the Age of Enlightenment cultural movement. Because of this, he advised that doctors ought to share as much information as possible with patients. He recommended that doctors educate the public and respect a patient's informed decision to accept therapy. There is no evidence that he supported seeking a consent from patients. In a lecture titled "On the duties of patients to their physicians", he stated that patients should be strictly obedient to the physician's orders; this was representative of much of his writings. John Gregory, Rush's teacher, wrote similar views that a doctor could best practice beneficence by making decisions for the patients without their consent.

Thomas Percival was a British physician who published a book called *Medical Ethics* in 1803. Percival was a student of the works of Gregory and various earlier Hippocratic physicians. Like all previous works, Percival's *Medical Ethics* makes no mention of soliciting for the consent of patients or respecting their decisions. Percival said that patients have a right to truth, but when the physician could provide better treatment by lying or withholding information, he advised that the physician do as he thought best. When the American Medical Association was founded they in 1847 produced a work called the first edition of the *American Medical Association Code of Medical Ethics*. Many sections of this book are verbatim copies of passages from Percival's *Medical Ethics*. A new concept in this book was the idea that physicians should fully disclose all patient details truthfully when talking to other physicians, but the text does not also apply this idea to disclosing information to patients. Through this text, Percival's ideas became pervasive guidelines throughout the United States as

other texts were derived from them. Worthington Hooker was an American physician who in 1849 published *Physician and Patient*. This medical ethics book was radical demonstrating understanding of the AMA's guidelines and Percival's philosophy and soundly rejecting all directives that a doctor should lie to patients. In Hooker's view, benevolent deception is not fair to the patient, and he lectured widely on this topic. Hooker's ideas were not broadly influential.

Research History

Historians cite a series of human subject research experiments to trace the history of informed consent in research. The U.S. Army Yellow Fever Commission "is considered the first research group in history to use consent forms." In 1900, Major Walter Reed was appointed head of the four man U.S. Army Yellow Fever Commission in Cuba that determined mosquitoes were the vector for yellow fever transmission. His earliest experiments were probably done without formal documentation of informed consent. In later experiments he obtained support from appropriate military and administrative authorities. He then drafted what is now "one of the oldest series of extant informed consent documents." The three surviving examples are in Spanish with English translations; two have an individual's signature and one is marked with an X.

Tearoom Trade is the name of a book by American psychologist Laud Humphreys. In it he describes his research into male homosexual acts. In conducting this research he never sought consent from his research subjects and other researchers raised concerns that he violated the right to privacy for research participants. The Milgram experiment is the name of a 1961 experiment conducted by American psychologist Stanley Milgram. In the experiment Milgram had an authority figure order research participants to commit a disturbing act of harming another person. After the experiment he would reveal that he had deceived the participants and that they had not hurt anyone, but the research participants were upset at the experience of having participated in the research. The experiment raised broad discussion on the ethics of recruiting participants for research without giving them full information about the nature of the research. Chester M. Southam used HeLa cells to inject into cancer patients and Ohio State Penitentiary inmates without informed consent to determine if people could become immune to cancer and if cancer could be transmitted.

MEDICAL PROCEDURES

The doctrine of informed consent relates to professional negligence and establishes a breach of the duty of care owed to the patient. The doctrine of informed consent also has significant implications for medical trials of medications, devices, or procedures.

Requirements of the Professional

Until 2015 in the United Kingdom and in countries such as Malaysia and Singapore, informed consent in medical procedures requires proof as to the standard of care to expect as a recognised standard of acceptable professional practice (the Bolam Test), that is, what risks would a medical professional usually disclose in the circumstances. Arguably, this is "sufficient consent" rather than "informed consent." The UK has since departed from the Bolam test for judging standards of informed consent, due to the landmark ruling in *Montgomery v Lanarkshire Health Board*. This moves away from the concept of a reasonable physician and instead uses the standard of a reasonable patient, and what risks an individual would attach significance to.

Medicine in the United States, Australia, and Canada also takes this patient-centric approach to "informed consent." Informed consent in these jurisdictions requires health care providers to disclose significant risks, as well as risks of particular importance to that patient. This approach combines an objective (a hypothetical reasonable patient) and subjective (this particular patient) approach.

The doctrine of informed consent should be contrasted with the general doctrine of medical consent, which applies to assault or battery. The consent standard here is only that the person understands, in general terms, the nature of and purpose of the intended intervention. As the higher standard of informed consent applies to negligence, not battery, the other elements of negligence must be made out. Significantly, causation must be shown: That had the individual been made aware of the risk he would not have proceeded with the operation (or perhaps with that surgeon).

Optimal establishment of an informed consent requires adaptation to cultural or other individual factors of the patient. For example, people from Mediterranean and Arab appear to rely more on the context of the delivery of the information, with the information being carried more by who is saying it and where, when, and how it's being said, rather than *what* is said, which is of relatively more importance in typical "Western" countries. The informed consent doctrine is generally implemented through good health care practice: pre-operation discussions with patients and the use of medical consent forms in hospitals. However, reliance on a signed form should not undermine the basis of the doctrine in giving the patient an opportunity to weigh and respond to the risk. In one British case, a doctor performing routine surgery on a woman noticed that she had cancerous tissue in her womb. He took the initiative to remove the woman's womb; however, as she had not given informed consent for this operation, the doctor was judged by the General Medical Council to have acted negligently. The council stated that the woman should have been informed of her condition, and allowed to make her own decision.

Obtaining Informed Consents

To capture and manage informed consents, hospital management systems typically use paper-based consent forms which are scanned and stored in a document handling system after obtaining the necessary signatures. Hospital systems and research organizations are adopting an electronic way of capturing informed consents to enable indexing, to improve comprehension, search and retrieval of consent data, thus enhancing the ability to honour to patient intent and identify willing research participants. More recently, Health Sciences South Carolina, a statewide research collaborative focused on transforming health care quality, health information systems and patient outcomes, developed an open-source system called Research Permissions Management System (RPMS).

Competency of the Patient

The ability to give informed consent is governed by a general requirement of competency. In common law jurisdictions, adults are presumed competent to consent. This presumption can be rebutted, for instance, in circumstances of mental illness or other incompetence. This may be prescribed in legislation or based on a common-law standard of inability to understand the nature of the procedure. In cases of incompetent adults, a health care proxy makes medical decisions. In the absence of a proxy, the medical practitioner is expected to act in the patient's best interests until a proxy can be found. By contrast, 'minors' (which may be defined differently in different jurisdictions) are generally presumed incompetent to consent, but depending on their age and other factors may be required to provide Informed assent. In some jurisdictions (*e.g.* much of the U.S.), this is a strict standard. In other jurisdictions (*e.g.* England, Australia, Canada), this presumption may be rebutted through proof that the minor is 'mature' (the 'Gillick standard'). In cases of incompetent minors, informed consent is usually required from the parent (rather than the 'best interests standard') although a parens patriae order may apply, allowing the court to dispense with parental consent in cases of refusal.

Deception

Research involving deception is controversial given the requirement for informed consent. Deception typically arises in social psychology, when researching a particular psychological process requires that investigators

deceive subjects. For example, in the Milgram experiment, researchers wanted to determine the willingness of participants to obey authority figures despite their personal conscientious objections. They had authority figures demand that participants deliver what they thought was an electric shock to another research participant. For the study to succeed, it was necessary to deceive the participants so they believed that the subject was a peer and that their electric shocks caused the peer actual pain.

Nonetheless, research involving deception prevents the subject/patient from exercising his/her basic right of autonomous informed decision-making and conflicts with the ethical principle of respect for persons. The Ethical Principles of Psychologists and Code of Conduct set by the American Psychological Association says that psychologists may not conduct research that includes a deceptive compartment unless they can justify the act by the value and importance of the study's results, and show they couldn't obtain the results by some other way. Moreover, the research should bear no potential harm to the subject as an outcome of deception, be it physical pain or emotional distress. Finally, the code requires a debriefing session, in which the experimenter tells the subject about the deception, and gives subjects the option of withdrawing their data.

Abortion

In some U.S. states, informed consent laws (sometimes called "right to know" laws) require that a woman seeking an elective abortion receive information from the abortion provider about her legal rights, alternatives to abortion (such as adoption), available public and private assistance, and other information specified in the law, before the abortion is performed. Other countries with such laws (*e.g.* Germany) require that the information giver be properly certified to make sure that no abortion is carried out for the financial gain of the abortion provider and to ensure that the decision to have an abortion is not swayed by any form of incentive. Some informed consent laws have been criticized for allegedly using "loaded language in an apparently deliberate attempt to 'personify' the fetus," but those critics acknowledge that "most of the information in the [legally mandated] materials about abortion comports with recent scientific findings and the principles of informed consent", although "some content is either misleading or altogether incorrect."

CHILDREN

As children often lack the decision making ability or legal power (competence) to provide true informed consent for medical decisions, it often falls on parents or legal guardians to provide *informed permission* for medical decisions. This "consent by proxy" usually works reasonably well, but can lead to ethical dilemmas when the judgement of the parents or guardians and the medical professional differ with regard to what constitutes appropriate decisions "in the best interest of the child". Children who are legally emancipated, and certain situations such as decisions regarding sexually transmitted diseases or pregnancy, or for unemancipated minors who are deemed to have medical decision making capacity, may be able to provide consent without the need for parental permission depending on the laws of the jurisdiction the child lives in. The American Academy of Pediatrics encourages medical professionals also to seek the assent of older children and adolescents by providing age appropriate information to these children to help empower them in the decision making process. Research on children has benefited society in many ways. The only effective way to establish normal patterns of growth and metabolism is to do research on infants and young children. When addressing the issue of informed consent with children, the primary response is parental consent. This is valid, although only legal guardians are able to consent for a child, not adult siblings. Additionally, parents may not order the termination of a treatment that is required to keep a child alive, even if they feel it is in the best interest. Guardians are typically involved in the consent of children, however a number of doctrines have developed that allow children to receive health treatments without parental consent. For example, emancipated minors may consent to medical treatment, and minors can also consent in an emergency.

RESEARCH

Informed consent is part of the ethical clinical research as well, in which a human subject voluntarily confirms his or her willingness to participate in a particular clinical trial, after having been informed of all aspects of the trial that are relevant to the subject's decision to participate. Informed consent is documented by means of a written, signed, and dated informed consent form. In medical research, the Nuremberg Code set a base international standard in 1947, which continued to develop, for example in response to the ethical violation in the Holocaust. Nowadays, medical research is overseen by an ethics committee that also oversees the informed consent process. As the medical guidelines established in the Nuremberg Code were imported into the ethical guidelines for the social sciences, informed consent became a common part of the research procedure. However, while informed consent is the default in medical settings, it is not always required in the social science. Here, research often involves low or no risk for participants, unlike in many medical experiments. Second, the mere knowledge that they participate in a study can cause people to alter their behaviour, as in the Hawthorne Effect: "In the typical lab experiment, subjects enter an environment in which they are keenly aware that their behaviour is being monitored, recorded, and subsequently scrutinized." In such cases, seeking informed consent directly interferes with the ability to conduct the research, because the very act of revealing that a study is being conducted is likely to alter the behaviour studied. List exemplifies the potential dilemma that can result: "if one were interested in exploring whether, and to what extent, race or gender influences the prices that buyers pay for used cars, it would be difficult to measure accurately the degree of discrimination among used car dealers who know that they are taking part in an experiment." In cases where such interference is likely, and after careful consideration, a researcher may forgo the informed consent process. This is commonly done after weighting the risk to study participants versus the benefit to society and whether participants are present in the study out of their own wish and treated fairly. Researchers often consult with an ethics committee or institutional review board to render a decision.

The birth of new online media, such as social media, has complicated the idea of informed consent. In an online environment people pay little attention to Terms of Use agreements and can subject themselves to research without thorough knowledge.

This issue came to the public light following a study conducted by Facebook Inc. in 2014, and published by that company and Cornell University. Facebook conducted a study where they altered the Facebook News Feeds of roughly 700,000 users to reduce either the amount of positive or negative posts they saw for a week. The study then analyzed if the users status updates changed during the different conditions. The study was published in the Proceedings of the National Academy of Sciences.

The lack of informed consent led to outrage among many researchers and users. Many believed that by potentially altering the mood of users by altering what posts they see, Facebook put at-risk individuals at higher dangers for depression and suicide. However, supports of Facebook claim that Facebook details that they have the right to use information for research in their terms of use. Others say the experiment is just a part of Facebook's current work, which alters News Feeds algorithms continually to keep people interested and coming back to the site. Others pointed out that this specific study is not along but that news organizations constantly try out different headlines using algorithms to elicit emotions and garner clicks or Facebook shares. They say this Facebook study is no different than things people already accept.

Still, others say that Facebook broke the law when conducting the experiment on user that didn't give informed consent. The Facebook study controversy raises numerous questions about informed consent and the differences in the ethical review process between publicly and privately funded research. Some say Facebook was within its limits and others see the need for more informed consent and/or the establishment of in-house private review boards.

CONFLICTS OF INTEREST

Other, long-standing controversies underscore the role for conflicts of interest among medical school faculty and researchers. For example, coverage of University of California (UC) medical school faculty members has included news of ongoing corporate payments to researchers and practitioners from companies that market and produce the very devices and treatments they recommend to patients. Robert Pedowitz, the former chairman of UCLA's orthopedic surgery department, reported concern that his colleague's financial conflicts of interest could negatively affect patient care or research into new treatments. In a subsequent lawsuit about whistleblower retaliation, the University provided a $10 million settlement to Pedowitz while acknowledging no wrongdoing. Consumer Watchdog, an oversight group, observed that University of CA policies were "either inadequate or unenforced…Patients in UC hospitals deserve the most reliable surgical devices and medication…and they shouldn't be treated as subjects in expensive experiments." Other UC incidents include taking the eggs of women for implantation into other women without consent and injecting live bacteria into human brains, resulting in potentially premature deaths.

ETHICS COMMITTEE

An ethics committee is a body responsible for ensuring that medical experimentation and human research are carried out in an ethical manner in accordance with national and international law.

SPECIFIC REGIONS

An ethics committee in the European Union is a body responsible for oversight of medical or human research studies in EU member states. Local terms for a European ethics committee include:

- A Research Ethics Committee (REC) in the United Kingdom
- A Medical Research Ethics Committee (MREC) in the Netherlands.
- A Comités de Protection des Personnes (CPP) in France.

In the United States, an ethics committee is usually known as an institutional review board and is dedicated to overseeing the rights and well-being of research subjects participating in scientific studies in the US. Similarly in Canada, the committee is called a Research Ethics Board (REB).

In Australia, an ethics committee in medical research refers to a Human Research Ethics Committee (HREC).

HISTORY

One of the most fundamental ethical principles in human experimentation is that the experimenter should not subject the participants in the experiment to any procedure they would not be willing to undertake themselves. This idea was first codified in the Nuremberg Code in 1947, which was a result of the trials of Nazi doctors at the Nuremberg trials accused of murdering and torturing victims in valueless experiments. Several of these doctors were hanged. Point five of the Nuremberg Code requires that no experiment should be conducted that is dangerous to the subjects unless the experimenters themselves also take part. The Nuremberg Code has influenced medical experiment codes of practice around the world, as has the exposure of experiments that have since failed to follow it such as the notorious Tuskegee syphilis experiment. Another ethical principle is that volunteers must stand to gain some benefit from the research, even if that is only a remote future possibility of treatment being found for a disease that they only have a small chance of contracting. Tests on experimental drugs are sometimes conducted on sufferers of an untreatable condition. If the researcher does not have that condition then there can be no possible benefit to them personally. For instance, Ronald C. Desrosiers in responding to why he did not test an AIDS vaccine he was developing on himself said that he was not at risk of AIDS so could not possibly benefit.

An important element of an ethics committee's oversight is to ensure that informed consent of the subjects has been given. Informed consent is the principle that the volunteers in the experiment should fully understand the procedure that is going to take place, be aware of all the risks involved, and give their consent to taking part in the experiment beforehand. The principle of informed consent was first enacted in the U.S. Army's research into Yellow fever in Cuba in 1901. However, there was no general or official guidance at this time. That remained the case until the yellow fever programme was referenced in the drafting of the Nuremberg Code. This was further developed in the Declaration of Helsinki in 1964 by the World Medical Association which has since become the foundation for ethics committees' guidelines.

The convening of ethics committees to approve the research protocol in human experiments was first written into international guidelines in the first revision to the Declaration of Helsinki (Helsinki II, 1975). A controversy arose over the fourth revision (1996) concerning placebo trials in developing countries. It was claimed that US trials of the anti-HIV drug zidovudine in India was in breach of this requirement. This led the US Food and Drug Administration to cease incorporating new revisions of Helsinki and refer instead to the 1989 revision.

Ethics committees are also made a requirement in *International Ethical Guidelines for Biomedical Research Involving Human Subjects*, produced by the Council for International Organizations of Medical Sciences (CIOMS), a body set up by the World Health Organization. First published in 1993, the CIOMS guidelines have no legal force but they have been influential in the drafting of national regulations for ethics committees. The COIMS guidelines are focused on practice in developing countries.

QUESTIONNAIRE

A questionnaire is a research instrument consisting of a series of questions (or other types of prompts) for the purpose of gathering information from respondents. The questionnaire was invented by the Statistical Society of London in 1838. Although questionnaires are often designed for statistical analysis of the responses, this is not always the case.

Questionnaires have advantages over some other types of surveys in that they are cheap, do not require as much effort from the questioner as verbal or telephone surveys, and often have standardized answers that make it simple to compile data.

However, such standardized answers may frustrate users. Questionnaires are also sharply limited by the fact that respondents must be able to read the questions and respond to them. Thus, for some demographic groups conducting a survey by questionnaire may not be concrete.

TYPES

A distinction can be made between questionnaires with questions that measure separate variables, and questionnaires with questions that are aggregated into either a scale or index. Questionnaires with questions that measure separate variables, could for instance include questions on:

- Preferences (*e.g.* political party)
- Behaviors (*e.g.* food consumption)
- Facts (*e.g.* gender)

Questionnaires with questions that are aggregated into either a scale or index, include for instance questions that measure:

- Latent traits
- Attitudes (*e.g.* towards immigration)
- An index (*e.g.* Social Economic Status)

Examples

- A food frequency questionnaire (FFQ) is a questionnaire the type of diet consumed in people, and may be used as a research instrument. Examples of usages include assessment of intake of vitamins or toxins such as acrylamide.

QUESTIONNAIRE CONSTRUCTION

Question Type

Usually, a questionnaire consists of a number of questions that the respondent has to answer in a set format. A distinction is made between open-ended and closed-ended questions. An open-ended question asks the respondent to formulate his own answer, whereas a closed-ended question has the respondent pick an answer from a given number of options. The response options for a closed-ended question should be exhaustive and mutually exclusive. Four types of response scales for closed-ended questions are distinguished:

- Dichotomous, where the respondent has two options
- Nominal-polytomous, where the respondent has more than two unordered options
- Ordinal-polytomous, where the respondent has more than two ordered options
- (Bounded)Continuous, where the respondent is presented with a continuous scale

A respondent's answer to an open-ended question is coded into a response scale afterwards. An example of an open-ended question is a question where the testie has to complete a sentence (sentence completion item).

Question Sequence

In general, questions should flow logically from one to the next. To achieve the best response rates, questions should flow from the least sensitive to the most sensitive, from the factual and behavioural to the attitudinal, and from the more general to the more specific.

There typically is a flow that should be followed when constructing a questionnaire in regards to the order that the questions are asked. The order is as follows:

- Screens
- Warm-ups
- Transitions
- Skips
- Difficult
- Classification

Screens are used as a screening method to find out early whether or not someone should complete the questionnaire. Warm-ups are simple to answer, help capture interest in the survey, and may not even pertain to research objectives. Transition questions are used to make different areas flow well together. Skips include questions similar to "If yes, then answer question 3. If no, then continue to question 5." Difficult questions are towards the end because the respondent is in "response mode." Also, when completing an online questionnaire, the progress bars lets the respondent know that they are almost done so they are more willing to answer more difficult questions. Classification, or demographic question should be at the end because typically they can feel like personal questions which will make respondents uncomfortable and not willing to finish survey.

Basic Rules for Questionnaire Item Construction

- Use statements which are interpreted in the same way by members of different subpopulations of the population of interest.

- Use statements where persons that have different opinions or traits will give different answers.
- Think of having an "open" answer category after a list of possible answers.
- Use only one aspect of the construct you are interested in per item.
- Use positive statements and avoid negatives or double negatives.
- Do not make assumptions about the respondent.
- Use clear and comprehensible wording, easily understandable for all educational levels
- Use correct spelling, grammar and punctuation.
- Avoid items that contain more than one question per item (*e.g.* Do you like strawberries and potatoes?).
- Question should not be biased or even leading the participant towards an answer.

Multi-item Scales

Within social science research and practice, questionnaires are most frequently used to collect quantitative data using multi-item scales with the following characteristics:

- Multiple statements or questions (minimum ≥3; usually ≥5) are presented for each variable being examined.
- Each statement or question has an accompanying set of equidistant response-points (usually 5-7).
- Each response point has an accompanying verbal anchor (*e.g.*, "strongly agree") ascending from left to right.
- Verbal anchors should be balanced to reflect equal intervals between response-points.
- Collectively, a set of response-points and accompanying verbal anchors are referred to as a rating scale. One very frequently-used rating scale is a Likert scale.
- Usually, for clarity and efficiency, a single set of anchors is presented for multiple rating scales in a questionnaire.
- Collectively, a statement or question with an accompanying rating scale is referred to as an item.
- When multiple items measure the same variable in a reliable and valid way, they are collectively referred to as a multi-item scale, or a psychometric scale.
- The following types of reliability and validity should be established for a multi-item scale: internal reliability, test-retest reliability (if the variable is expected to be stable over time), content validity, construct validity, and criterion validity.
- Factor analysis is used in the scale development process.
- Questionnaires used to collect quantitative data usually comprise several multi-item scales, together with an introductory and concluding section.

QUESTIONNAIRE ADMINISTRATION MODES

Main modes of questionnaire administration include:

- Face-to-face questionnaire administration, where an interviewer presents the items orally.
- Paper-and-pencil questionnaire administration, where the items are presented on paper.
- Computerized questionnaire administration, where the items are presented on the computer.
- Adaptive computerized questionnaire administration, where a selection of items is presented on the computer, and based on the answers on those items, the computer selects following items optimized for the testee's estimated ability or trait.

CONCERNS WITH QUESTIONNAIRES

While questionnaires are inexpensive, quick, and easy to analyze, often the questionnaire can have more problems than benefits. For example, unlike interviews, the people conducting the research may never know if

the respondent understood the question that was being asked. Also, because the questions are so specific to what the researchers are asking, the information gained can be minimal. Often, questionnaires such as the Myers-Briggs Type Indicator, give too few options to answer; respondents can answer either option but must choose only one response. Questionnaires also produce very low return rates, whether they are mail or online questionnaires. The other problem associated with return rates is that often the people who do return the questionnaire are those who have a really positive or a really negative viewpoint and want their opinion heard. The people who are most likely unbiased either way typically don't respond because it is not worth their time.

One key concern with questionnaires is that there may contain quite large measurement errors (). These errors can be random or systematic. Random errors are caused by unintended mistakes by respondents, interviewers and/or coders. Systematic error can occur if there is a systematic reaction of the respondents to the scale used to formulate the survey question. Thus, the exact formulation of a survey question and its scale are crucial, since they affect the level of measurement error (). Different tools are available for the researchers to help them decide about this exact formulation of their questions, for instance estimating the quality of a question using MTMM experiments or predicting this quality using the Survey Quality Predictor software (SQP). This information about the quality can also be used in order to correct for measurement errors ().

Further, if the questionnaires are not collected using sound sampling techniques, often the results can be non-representative of the population—as such a good sample is critical to getting representative results based on questionnaires.

RELIABILITY

Reliability in statistics and psychometrics is the overall consistency of a measure. A measure is said to have a high reliability if it produces similar results under consistent conditions. "It is the characteristic of a set of test scores that relates to the amount of random error from the measurement process that might be embedded in the scores. Scores that are highly reliable are accurate, reproducible, and consistent from one testing occasion to another. That is, if the testing process were repeated with a group of test takers, essentially the same results would be obtained. Various kinds of reliability coefficients, with values ranging between 0.00 (much error) and 1.00 (no error), are usually used to indicate the amount of error in the scores." For example, measurements of people's height and weight are often extremely reliable.

TYPES

There are several general classes of reliability estimates:

- Inter-rater reliability assesses the degree of agreement between two or more raters in their appraisals.
- Test-retest reliability assesses the degree to which test scores are consistent from one test administration to the next. Measurements are gathered from a single rater who uses the same methods or instruments and the same testing conditions. This includes intra-rater reliability.
- Inter-method reliability assesses the degree to which test scores are consistent when there is a variation in the methods or instruments used. This allows inter-rater reliability to be ruled out. When dealing with forms, it may be termed parallel-forms reliability.
- Internal consistency reliability, assesses the consistency of results across items within a test.

DIFFERENCE FROM VALIDITY

Reliability does not imply validity. That is, a reliable measure that is measuring something consistently is not necessarily measuring what you want to be measured. For example, while there are many reliable tests of specific abilities, not all of them would be valid for predicting, say, job performance.

While reliability does not imply validity, reliability does place a limit on the overall validity of a test. A test that is not perfectly reliable cannot be perfectly valid, either as a means of measuring attributes of a person or as a means of predicting scores on a criterion. While a reliable test may provide useful valid information, a test that is not reliable cannot possibly be valid. For example, if a set of weighing scales consistently measured the weight of an object as 500 grams over the true weight, then the scale would be very reliable, but it would not be valid (as the returned weight is not the true weight). For the scale to be valid, it should return the true weight of an object. This example demonstrates that a perfectly reliable measure is not necessarily valid, but that a valid measure necessarily must be reliable.

GENERAL MODEL

In practice, testing measures are never perfectly consistent. Theories of test reliability have been developed to estimate the effects of inconsistency on the accuracy of measurement. The basic starting point for almost all theories of test reliability is the idea that test scores reflect the influence of two sorts of factors:

- *Factors that contribute to consistency:* stable characteristics of the individual or the attribute that one is trying to measure
- *Factors that contribute to inconsistency:* features of the individual or the situation that can affect test scores but have nothing to do with the attribute being measured.

These factors include:

- Temporary but general characteristics of the individual: health, fatigue, motivation, emotional strain
- Temporary and specific characteristics of individual: comprehension of the specific test task, specific tricks or techniques of dealing with the particular test materials, fluctuations of memory, attention or accuracy
- Aspects of the testing situation: freedom from distractions, clarity of instructions, interaction of personality, sex, or race of examiner
- Chance factors: luck in selection of answers by sheer guessing, momentary distractions

A true score is the replicable feature of the concept being measured. It is the part of the observed score that would recur across different measurement occasions in the absence of error. Errors of measurement are composed of both random error and systematic error. It represents the discrepancies between scores obtained on tests and the corresponding true scores.

This conceptual breakdown is typically represented by the simple equation:

Observed test score = true score + errors of measurement

ITEM RESPONSE THEORY

It was well-known to classical test theorists that measurement precision is not uniform across the scale of measurement. Tests tend to distinguish better for test-takers with moderate trait levels and worse among high- and low-scoring test-takers. Item response theory extends the concept of reliability from a single index to a function called the *information function*. The IRT information function is the inverse of the conditional observed score standard error at any given test score.

ESTIMATION

The goal of estimating reliability is to determine how much of the variability in test scores is due to errors in measurement and how much is due to variability in true scores.

Four practical strategies have been developed that provide workable methods of estimating test reliability.

- Test-retest reliability method: directly assesses the degree to which test scores are consistent from one test administration to the next.

It involves:

- Administering a test to a group of individuals
- Re-administering the same test to the same group at some later time
- Correlating the first set of scores with the second

The correlation between scores on the first test and the scores on the retest is used to estimate the reliability of the test using the Pearson product-moment correlation coefficient.

- Parallel-forms method:

The key to this method is the development of alternate test forms that are equivalent in terms of content, response processes and statistical characteristics. For example, alternate forms exist for several tests of general intelligence, and these tests are generally seen equivalent. With the parallel test model it is possible to develop two forms of a test that are equivalent in the sense that a person's true score on form A would be identical to their true score on form B. If both forms of the test were administered to a number of people, differences between scores on form A and form B may be due to errors in measurement only.

It involves:

- Administering one form of the test to a group of individuals
- At some later time, administering an alternate form of the same test to the same group of people
- Correlating scores on form A with scores on form B

The correlation between scores on the two alternate forms is used to estimate the reliability of the test.

This method provides a partial solution to many of the problems inherent in the test-retest reliability method. For example, since the two forms of the test are different, carryover effect is less of a problem. Reactivity effects are also partially controlled; although taking the first test may change responses to the second test. However, it is reasonable to assume that the effect will not be as strong with alternate forms of the test as with two administrations of the same test.

However, this technique has its disadvantages:

- It may be very difficult to create several alternate forms of a test
- It may also be difficult if not impossible to guarantee that two alternate forms of a test are parallel measures

- Split-half method:

This method treats the two halves of a measure as alternate forms. It provides a simple solution to the problem that the parallel-forms method faces: the difficulty in developing alternate forms.

It involves:

- Administering a test to a group of individuals
- Splitting the test in half
- Correlating scores on one half of the test with scores on the other half of the test

The correlation between these two split halves is used in estimating the reliability of the test. This halves reliability estimate is then stepped up to the full test length using the Spearman–Brown prediction formula. There are several ways of splitting a test to estimate reliability. For example, a 40-item vocabulary test could be split into two subtests, the first one made up of items 1 through 20 and the second made up of items 21 through 40. However, the responses from the first half may be systematically different from responses in the second half due to an increase in item difficulty and fatigue. In splitting a test, the two halves would need to be as similar as possible, both in terms of their content and in terms of the probable state of the respondent. The simplest method is to adopt an odd-even split, in which the odd-numbered items form one half of the test and the even-numbered items form the other. This arrangement guarantees that each half will contain an equal number of items from the beginning, middle, and end of the original test.

- Internal consistency: assesses the consistency of results across items within a test. The most common internal consistency measure is Cronbach's alpha, which is usually interpreted as the mean of all

possible split-half coefficients. Cronbach's alpha is a generalization of an earlier form of estimating internal consistency, Kuder–Richardson Formula 20. Although the most commonly used, there are some misconceptions regarding Cronbach's alpha.

These measures of reliability differ in their sensitivity to different sources of error and so need not be equal. Also, reliability is a property of the *scores of a measure* rather than the measure itself and are thus said to be *sample dependent*. Reliability estimates from one sample might differ from those of a second sample (beyond what might be expected due to sampling variations) if the second sample is drawn from a different population because the true variability is different in this second population. (This is true of measures of all types—yardsticks might measure houses well yet have poor reliability when used to measure the lengths of insects.)

Reliability may be improved by clarity of expression (for written assessments), lengthening the measure, and other informal means. However, formal psychometric analysis, called item analysis, is considered the most effective way to increase reliability. This analysis consists of computation of item difficulties and item discrimination indices, the latter index involving computation of correlations between the items and sum of the item scores of the entire test. If items that are too difficult, too easy, and/or have near-zero or negative discrimination are replaced with better items, the reliability of the measure will increase.

- $R(t) = 1- F(t)$.
- $R(t) = \exp(-\lambda t)$. (where λ is the failure rate)

VALIDITY (STATISTICS)

Validity is the extent to which a concept, conclusion or measurement is well-founded and likely corresponds accurately to the real world based on probability. The word "valid" is derived from the Latin validus, meaning strong.

This should not be confused with notions of certainty nor necessity. The validity of a measurement tool (for example, a test in education) is considered to be the degree of probability to which the tool measures what it claims to measure; in this case, the validity is an equivalent to a percent of how accurately the claim corresponds to reality. In psychometrics, validity has a particular application known as test validity: "the degree to which evidence and theory support the interpretations of test scores" ("as entailed by proposed uses of tests").

It is generally accepted that the concept of scientific validity addresses the nature of reality in terms of statistical probability and as such is an epistemological and philosophical issue as well as a question of measurement of the possibility that a scientific claim is true. The use of the term in logic is narrower, relating to the truth of inferences made from premises. In logic, and therefore as the term is applied to any epistemological claim, validity refers to the consistency of an argument flowing from the premises to the conclusion; as such, the truth of the claim in logic is not only reliant on validity. Rather, argumentative claim is true if and only if it is both valid and sound. This means the argument flows without contradiction from the premises or the conclusion, and all of the premises and the conclusion correspond to known facts. As such, "scientific or statistical validity" is not a deductive claim that is necessarily truth preserving, but is an inductive claim that remains true or false in an undecided manner. This is why "scientific or statistical validity" is a claim that is qualified as being either strong or weak in its nature, it is never necessarily nor certainly true. This has the effect of making claims of "scientific or statistical validity" open to interpretation as to what, in fact, the facts of the matter mean.

Validity is important because it can help determine what types of tests to use, and help to make sure researchers are using methods that are not only ethical, and cost-effective, but also a method that truly measures the idea or construct in question.

TEST VALIDITY

Validity (Accuracy)

Validity of an assessment is the degree to which it measures what it is supposed to measure. This is not the same as reliability, which is the extent to which a measurement gives results that are very consistent. Within validity, the measurement does not always have to be similar, as it does in reliability. However, just because a measure is reliable, it is not necessarily valid eg.

A scale that is 5 pounds off is reliable but not valid. A test cannot be valid unless it is reliable. Validity is also dependent on the measurement measuring what it was designed to measure, and not something else instead. Validity (similar to reliability) is a relative concept; validity is not an all-or-nothing idea. There are many different types of validity.

Construct Validity

Construct validity refers to the extent to which operationalizations of a construct (*e.g.*, practical tests developed from a theory) measure a construct as defined by a theory. It subsumes all other types of validity. For example, the extent to which a test measures intelligence is a question of construct validity. A measure of intelligence presumes, among other things, that the measure is associated with things it should be associated with (convergent validity), not associated with things it should not be associated with (discriminant validity).

Construct validity evidence involves the empirical and theoretical support for the interpretation of the construct. Such lines of evidence include statistical analyses of the internal structure of the test including the relationships between responses to different test items. They also include relationships between the test and measures of other constructs. As currently understood, construct validity is not distinct from the support for the substantive theory of the construct that the test is designed to measure. As such, experiments designed to reveal aspects of the causal role of the construct also contribute to construct validity evidence.

Content Validity

Content validity is a non-statistical type of validity that involves "the systematic examination of the test content to determine whether it covers a representative sample of the behaviour domain to be measured" (Anastasi and Urbina, 1997 p. 114). For example, does an IQ questionnaire have items covering all areas of intelligence discussed in the scientific literature?

Content validity evidence involves the degree to which the content of the test matches a content domain associated with the construct. For example, a test of the ability to add two numbers should include a range of combinations of digits. A test with only one-digit numbers, or only even numbers, would not have good coverage of the content domain. Content related evidence typically involves a subject matter expert (SME) evaluating test items against the test specifications. Before going to final administration of questionnaires, the researcher should consult the validity of items against each of the constructs or variables and accordingly modify measurement instruments on the basis of SME's opinion.

A test has content validity built into it by careful selection of which items to include (Anastasi and Urbina, 1997). Items are chosen so that they comply with the test specification which is drawn up through a thorough examination of the subject domain. Foxcroft, Paterson, le Roux and Herbst (2004, p. 49) note that by using a panel of experts to review the test specifications and the selection of items the content validity of a test can be improved. The experts will be able to review the items and comment on whether the items cover a representative sample of the behaviour domain.

Face Validity

Face validity is an estimate of whether a test appears to measure a certain criterion; it does not guarantee that the test actually measures phenomena in that domain. Measures may have high validity, but when the test does not appear to be measuring what it is, it has low face validity. Indeed, when a test is subject to faking (malingering), low face validity might make the test more valid. Considering one may get more honest answers with lower face validity, it is sometimes important to make it appear as though there is low face validity whilst administering the measures. Face validity is very closely related to content validity. While content validity depends on a theoretical basis for assuming if a test is assessing all domains of a certain criterion (*e.g.* does assessing addition skills yield in a good measure for mathematical skills? To answer this you have to know, what different kinds of arithmetic skills mathematical skills include) face validity relates to whether a test appears to be a good measure or not. This judgement is made on the "face" of the test, thus it can also be judged by the amateur.

Face validity is a starting point, but should never be assumed to be probably valid for any given purpose, as the "experts" have been wrong before—the Malleus Malificarum (Hammer of Witches) had no support for its conclusions other than the self-imagined competence of two "experts" in "witchcraft detection," yet it was used as a "test" to condemn and burn at the stake tens of thousands men and women as "witches."

Criterion Validity

Criterion validity evidence involves the correlation between the test and a criterion variable (or variables) taken as representative of the construct. In other words, it compares the test with other measures or outcomes (the criteria) already held to be valid. For example, employee selection tests are often validated against measures of job performance (the criterion), and IQ tests are often validated against measures of academic performance (the criterion). If the test data and criterion data are collected at the same time, this is referred to as concurrent validity evidence. If the test data are collected first in order to predict criterion data collected at a later point in time, then this is referred to as predictive validity evidence.

Concurrent Validity

Concurrent validity refers to the degree to which the operationalization correlates with other measures of the same construct that are measured at the same time. When the measure is compared to another measure of the same type, they will be related (or correlated). Returning to the selection test example, this would mean that the tests are administered to current employees and then correlated with their scores on performance reviews.

Predictive Validity

Predictive validity refers to the degree to which the operationalization can predict (or correlate with) other measures of the same construct that are measured at some time in the future. Again, with the selection test example, this would mean that the tests are administered to applicants, all applicants are hired, their performance is reviewed at a later time, and then their scores on the two measures are correlated.

This is also when measurement predicts a relationship between what is measured and something else; predicting whether or not the other thing will happen in the future. This type of validity is important from a public view standpoint; is this going to look acceptable to the public or not?

Experimental Validity

The validity of the design of experimental research studies is a fundamental part of the scientific method, and a concern of research ethics. Without a valid design, valid scientific conclusions cannot be drawn.

Statistical Conclusion Validity

Statistical conclusion validity is the degree to which conclusions about the relationship among variables based on the data are correct or 'reasonable'. This began as being solely about whether the statistical conclusion about the relationship of the variables was correct, but now there is a movement towards moving to 'reasonable' conclusions that use: quantitative, statistical, and qualitative data. Statistical conclusion validity involves ensuring the use of adequate sampling procedures, appropriate statistical tests, and reliable measurement procedures. As this type of validity is concerned solely with the relationship that is found among variables, the relationship may be solely a correlation.

Internal Validity

Internal validity is an inductive estimate of the degree to which conclusions about *causal* relationships can be made (*e.g.* cause and effect), based on the measures used, the research setting, and the whole research design. Good experimental techniques, in which the effect of an independent variable on a dependent variable is studied under highly controlled conditions, usually allow for higher degrees of internal validity than, for example, single-case designs.

Eight kinds of confounding variable can interfere with internal validity (i.e. with the attempt to isolate causal relationships):

- History, the specific events occurring between the first and second measurements in addition to the experimental variables
- Maturation, processes within the participants as a function of the passage of time (not specific to particular events), *e.g.*, growing older, hungrier, more tired, and so on.
- Testing, the effects of taking a test upon the scores of a second testing.
- Instrumentation, changes in calibration of a measurement tool or changes in the observers or scorers may produce changes in the obtained measurements.
- Statistical regression, operating where groups have been selected on the basis of their extreme scores.
- Selection, biases resulting from differential selection of respondents for the comparison groups.
- Experimental mortality, or differential loss of respondents from the comparison groups.
- Selection-maturation interaction, etc. *e.g.*, in multiple-group quasi-experimental designs.

External Validity

External validity concerns the extent to which the (internally valid) results of a study can be held to be true for other cases, for example to different people, places or times. In other words, it is about whether findings can be validly generalized. If the same research study was conducted in those other cases, would it get the same results?

A major factor in this is whether the study sample (e.g. the research participants) are representative of the general population along relevant dimensions. Other factors jeopardizing external validity are:

- Reactive or interaction effect of testing, a pretest might increase the scores on a posttest
- Interaction effects of selection biases and the experimental variable.
- Reactive effects of experimental arrangements, which would preclude generalization about the effect of the experimental variable upon persons being exposed to it in non-experimental settings.
- Multiple-treatment interference, where effects of earlier treatments are not erasable.

Ecological Validity

Ecological validity is the extent to which research results can be applied to real-life situations outside of research settings. This issue is closely related to external validity but covers the question of to what degree

experimental findings mirror what can be observed in the real world (ecology = the science of interaction between organism and its environment). To be ecologically valid, the methods, materials and setting of a study must approximate the real-life situation that is under investigation.

Ecological validity is partly related to the issue of experiment versus observation. Typically in science, there are two domains of research: observational (passive) and experimental (active). The purpose of experimental designs is to test causality, so that you can infer A causes B or B causes A. But sometimes, ethical and/or methological restrictions prevent you from conducting an experiment (*e.g.* how does isolation influence a child's cognitive functioning?). Then you can still do research, but it is not causal, it is correlational. You can only conclude that A occurs together with B. Both techniques have their strengths and weaknesses.

Relationship to Internal Validity

On first glance, internal and external validity seem to contradict each other – to get an experimental design you have to control for all interfering variables. That is why you often conduct your experiment in a laboratory setting. While gaining internal validity (excluding interfering variables by keeping them constant) you lose ecological or external validity because you establish an artificial laboratory setting. On the other hand, with observational research you can not control for interfering variables (low internal validity) but you can measure in the natural (ecological) environment, at the place where behaviour normally occurs. However, in doing so, you sacrifice internal validity. The apparent contradiction of internal validity and external validity is, however, only superficial. The question of whether results from a particular study generalize to other people, places or times arises only when one follows an inductivist research strategy. If the goal of a study is to deductively test a theory, one is only concerned with factors which might undermine the rigour of the study, *i.e.* threats to internal validity.

Diagnostic Validity

In psychiatry there is a particular issue with assessing the validity of the diagnostic categories themselves. In this context:

- Content validity may refer to symptoms and diagnostic criteria;
- Concurrent validity may be defined by various correlates or markers, and perhaps also treatment response;
- Predictive validity may refer mainly to diagnostic stability over time;
- Discriminant validity may involve delimitation from other disorders.

Robins and Guze proposed in 1970 what were to become influential formal criteria for establishing the validity of psychiatric diagnoses. They listed five criteria:

- Distinct clinical description (including symptom profiles, demographic characteristics, and typical precipitants)
- Laboratory studies (including psychological tests, radiology and postmortem findings)
- Delimitation from other disorders (by means of exclusion criteria)
- Follow-up studies showing a characteristic course (including evidence of diagnostic stability)
- Family studies showing familial clustering

These were incorporated into the Feighner Criteria and Research Diagnostic Criteria that have since formed the basis of the DSM and ICD classification systems.

Kendler in 1980 distinguished between:

- Antecedent validators (familial aggregation, premorbid personality, and precipitating factors)
- Concurrent validators (including psychological tests)
- Predictive validators (diagnostic consistency over time, rates of relapse and recovery, and response to treatment)

Nancy Andreasen (1995) listed several additional validators – molecular genetics and molecular biology, neurochemistry, neuroanatomy, neurophysiology, and cognitive neuroscience – that are all potentially capable of linking symptoms and diagnoses to their neural substrates.

Kendell and Jablinsky (2003) emphasized the importance of distinguishing between validity and utility, and argued that diagnostic categories defined by their syndromes should be regarded as valid only if they have been shown to be discrete entities with natural boundaries that separate them from other disorders.

Kendler (2006) emphasized that to be useful, a validating criterion must be sensitive enough to validate most syndromes that are true disorders, while also being specific enough to invalidate most syndromes that are not true disorders. On this basis, he argues that a Robins and Guze criterion of "runs in the family" is inadequately specific because most human psychological and physical traits would qualify - for example, an arbitrary syndrome comprising a mixture of "height over 6 ft, red hair, and a large nose" will be found to "run in families" and be "hereditary", but this should not be considered evidence that it is a disorder. Kendler has further suggested that "essentialist" gene models of psychiatric disorders, and the hope that we will be able to validate categorical psychiatric diagnoses by "carving nature at its joints" solely as a result of gene discovery, are implausible. In the United States Federal Court System validity and reliability of evidence is evaluated using the Daubert Standard. Perri and Lichtenwald (2010) provide a starting point for a discussion about a wide range of reliability and validity topics in their analysis of a wrongful murder conviction.

5

Bias

Statistical bias is a feature of a statistical technique or of its results whereby the expected value of the results differs from the true underlying quantitative parameter being estimated.

SELECTION BIAS

Selection bias is the bias introduced by the selection of individuals, groups or data for analysis in such a way that proper randomization is not achieved, thereby ensuring that the sample obtained is not representative of the population intended to be analyzed. It is sometimes referred to as the selection effect. The phrase "selection bias" most often refers to the distortion of a statistical analysis, resulting from the method of collecting samples. If the selection bias is not taken into account, then some conclusions of the study may not be accurate.

TYPES

There are many types of possible selection bias, including:

Sampling Bias

Sampling bias is systematic error due to a non-random sample of a population, causing some members of the population to be less likely to be included than others, resulting in a biased sample, defined as a statistical sample of a population (or non-human factors) in which all participants are not equally balanced or objectively represented. It is mostly classified as a subtype of selection bias, sometimes specifically termed *sample selection bias*, but some classify it as a separate type of bias. A distinction of sampling bias (albeit not a universally accepted one) is that it undermines the external validity of a test (the ability of its results to be generalized to the rest of the population), while selection bias mainly addresses internal validity for differences or similarities found in the sample at hand. In this sense, errors occurring in the process of gathering the sample or cohort cause sampling bias, while errors in any process thereafter cause selection bias.

Examples of sampling bias include self-selection, pre-screening of trial participants, discounting trial subjects/tests that did not run to completion and migration bias by excluding subjects who have recently moved into or out of the study area.

Time Interval

- Early termination of a trial at a time when its results support the desired conclusion.
- A trial may be terminated early at an extreme value (often for ethical reasons), but the extreme value is likely to be reached by the variable with the largest variance, even if all variables have a similar mean.

Exposure

- *Susceptibility bias*
  (a) *Clinical susceptibility bias*, when one disease predisposes for a second disease, and the treatment for the first disease erroneously appears to predispose to the second disease. For example, postmenopausal syndrome gives a higher likelihood of also developing endometrial cancer, so estrogens given for the postmenopausal syndrome may receive a higher than actual blame for causing endometrial cancer.
  (b) *Protopathic bias*, when a treatment for the first symptoms of a disease or other outcome appear to cause the outcome. It is a potential bias when there is a lag time from the first symptoms and start of treatment before actual diagnosis. It can be mitigated by lagging, that is, exclusion of exposures that occurred in a certain time period before diagnosis.
  (c) *Indication bias*, a potential mixup between cause and effect when exposure is dependent on indication, *e.g.* a treatment is given to people in high risk of acquiring a disease, potentially causing a preponderance of treated people among those acquiring the disease. This may cause an erroneous appearance of the treatment being a cause of the disease.

Data

- Partitioning (dividing) data with knowledge of the contents of the partitions, and then analyzing them with tests designed for blindly chosen partitions.
- Post hoc alteration of data inclusion based on arbitrary or subjective reasons, including:
  (a) Cherry picking, when specific subsets of data are chosen to support a conclusion (*e.g.* citing examples of plane crashes as evidence of airline flight being unsafe, while ignoring the far more common example of flights that complete safely)
  (b) Rejection of bad data on (1) arbitrary grounds, instead of according to previously stated or generally agreed criteria or (2) discarding "outliers" on statistical grounds that fail to take into account important information that could be derived from "wild" observations.

Studies

- Selection of which studies to include in a meta-analysis.
- Performing repeated experiments and reporting only the most favorable results, perhaps relabelling lab records of other experiments as "calibration tests", "instrumentation errors" or "preliminary surveys".
- Presenting the most significant result of a data dredge as if it were a single experiment (which is logically the same as the previous item, but is seen as much less dishonest).

Attrition

Attrition bias is a kind of selection bias caused by attrition (loss of participants), discounting trial subjects/tests that did not run to completion. It is closely related to the survivorship bias, where only the subjects that "survived" a process are included in the analysis or the failure bias, where only the subjects that "failed" a process are included. It includes *dropout*, *non response* (lower response rate), *withdrawal* and *protocol deviators*. It gives biased results where it is unequal in regard to exposure and/or outcome. For example, in a test of a dieting programme, the researcher may simply reject everyone who drops out of the trial, but most of those who drop out are those for whom it was not working. Different loss of subjects in intervention and comparison group may change the characteristics of these groups and outcomes irrespective of the studied intervention.

Observer Selection

Data are filtered not only by study design and measurement, but by the necessary precondition that there has to be someone doing a study. In situations where the existence of the observer or the study is correlated with the data, observation selection effects occur, and anthropic reasoning is required. An example is the past impact event record of Earth: if large impacts cause mass extinctions and ecological disruptions precluding the evolution of intelligent observers for long periods, no one will observe any evidence of large impacts in the recent past (since they would have prevented intelligent observers from evolving). Hence there is a potential bias in the impact record of Earth. Astronomical existential risks might similarly be underestimated due to selection bias, and an anthropic correction has to be introduced.

MITIGATION

In the general case, selection biases cannot be overcome with statistical analysis of existing data alone, though Heckman correction may be used in special cases. An assessment of the degree of selection bias can be made by examining correlations between exogenous (background) variables and a treatment indicator. However, in regression models, it is correlation between *unobserved* determinants of the outcome and *unobserved* determinants of selection into the sample which bias estimates, and this correlation between unobservables cannot be directly assessed by the observed determinants of treatment.

RELATED ISSUES

Selection bias is closely related to:

- Publication bias or reporting bias, the distortion produced in community perception or meta-analyses by not publishing uninteresting (usually negative) results, or results which go against the experimenter's prejudices, a sponsor's interests, or community expectations.
- Confirmation bias, the distortion produced by experiments that are designed to seek confirmatory evidence instead of trying to disprove the hypothesis.
- Exclusion bias, results from applying different criteria to cases and controls in regards to participation eligibility for a study/different variables serving as basis for exclusion.

STATISTICAL HYPOTHESIS TESTING

A statistical hypothesis, sometimes called confirmatory data analysis, is an hypothesis that is testable on the basis of observing a process that is modeled via a set of random variables. A statistical hypothesis test is a method of statistical inference. Commonly, two statistical data sets are compared, or a data set obtained by sampling is compared against a synthetic data set from an idealized model. A hypothesis is proposed for the statistical relationship between the two data sets, and this is compared as an alternative to an idealized null hypothesis that proposes no relationship between two data sets. The comparison is deemed *statistically significant* if the relationship between the data sets would be an unlikely realization of the null hypothesis according to a threshold probability—the significance level. Hypothesis tests are used in determining what outcomes of a study would lead to a rejection of the null hypothesis for a pre-specified level of significance. The process of distinguishing between the null hypothesis and the alternative hypothesis is aided by identifying two conceptual types of errors, type 1 and type 2, and by specifying parametric limits on *e.g.* how much type 1 error will be permitted. An alternative framework for statistical hypothesis testing is to specify a set of statistical models, one for each candidate hypothesis, and then use model selection techniques to choose the most appropriate model. The most common selection techniques are based on either Akaike information criterion or Bayes factor. Confirmatory data analysis can be contrasted with exploratory data analysis, which may not have pre-specified hypotheses.

VARIATIONS AND SUB-CLASSES

Statistical hypothesis testing is a key technique of both frequentist inference and Bayesian inference, although the two types of inference have notable differences. Statistical hypothesis tests define a procedure that controls (fixes) the probability of incorrectly *deciding* that a default position (null hypothesis) is incorrect. The procedure is based on how likely it would be for a set of observations to occur if the null hypothesis were true. Note that this probability of making an incorrect decision is *not* the probability that the null hypothesis is true, nor whether any specific alternative hypothesis is true. This contrasts with other possible techniques of decision theory in which the null and alternative hypothesis are treated on a more equal basis. One naïve Bayesian approach to hypothesis testing is to base decisions on the posterior probability, but this fails when comparing point and continuous hypotheses. Other approaches to decision making, such as Bayesian decision theory, attempt to balance the consequences of incorrect decisions across all possibilities, rather than concentrating on a single null hypothesis. A number of other approaches to reaching a decision based on data are available via decision theory and optimal decisions, some of which have desirable properties. Hypothesis testing, though, is a dominant approach to data analysis in many fields of science. Extensions to the theory of hypothesis testing include the study of the power of tests, *i.e.* the probability of correctly rejecting the null hypothesis given that it is false. Such considerations can be used for the purpose of sample size determination prior to the collection of data.

THE TESTING PROCESS

In the statistics literature, statistical hypothesis testing plays a fundamental role. The usual line of reasoning is as follows:

- There is an initial research hypothesis of which the truth is unknown.
- The first step is to state the relevant null and alternative hypotheses. This is important, as mis-stating the hypotheses will muddy the rest of the process.
- The second step is to consider the statistical assumptions being made about the sample in doing the test; for example, assumptions about the statistical independence or about the form of the distributions of the observations. This is equally important as invalid assumptions will mean that the results of the test are invalid.
- Decide which test is appropriate, and state the relevant test statistic*T*.
- Derive the distribution of the test statistic under the null hypothesis from the assumptions. In standard cases this will be a well-known result. For example, the test statistic might follow a Student's t distribution or a normal distribution.
- Select a significance level (a), a probability threshold below which the null hypothesis will be rejected. Common values are 5 per cent and 1 per cent.
- The distribution of the test statistic under the null hypothesis partitions the possible values of T into those for which the null hypothesis is rejected—the so-called *critical region*—and those for which it is not. The probability of the critical region is a.
- Compute from the observations the observed value t_{obs} of the test statistic T.
- Decide to either reject the null hypothesis in favour of the alternative or not reject it. The decision rule is to reject the null hypothesis H_0 if the observed value t_{obs} is in the critical region, and to accept or "fail to reject" the hypothesis otherwise.

An alternative process is commonly used:

- Compute from the observations the observed value t_{obs} of the test statistic T.
- Calculate the p-value. This is the probability, under the null hypothesis, of sampling a test statistic at least as extreme as that which was observed.

- Reject the null hypothesis, in favour of the alternative hypothesis, if and only if the p-value is less than the significance level (the selected probability) threshold.

The two processes are equivalent. The former process was advantageous in the past when only tables of test statistics at common probability thresholds were available. It allowed a decision to be made without the calculation of a probability. It was adequate for classwork and for operational use, but it was deficient for reporting results.

The latter process relied on extensive tables or on computational support not always available. The explicit calculation of a probability is useful for reporting. The calculations are now trivially performed with appropriate software.

The difference in the two processes applied to the Radioactive suitcase example (below):

- "The Geiger-counter reading is 10. The limit is 9. Check the suitcase."
- "The Geiger-counter reading is high; 97 per cent of safe suitcases have lower readings. The limit is 95 per cent. Check the suitcase."

The former report is adequate, the latter gives a more detailed explanation of the data and the reason why the suitcase is being checked. It is important to note the difference between accepting the null hypothesis and simply failing to reject it. The "fail to reject" terminology highlights the fact that the null hypothesis is assumed to be true from the start of the test; if there is a lack of evidence against it, it simply continues to be assumed true. The phrase "accept the null hypothesis" may suggest it has been proved simply because it has not been disproved, a logical fallacy known as the argument from ignorance. Unless a test with particularly high power is used, the idea of "accepting" the null hypothesis may be dangerous. Nonetheless the terminology is prevalent throughout statistics, where the meaning actually intended is well understood. The processes described here are perfectly adequate for computation. They seriously neglect the design of experiments considerations.

It is particularly critical that appropriate sample sizes be estimated before conducting the experiment.

The phrase "test of significance" was coined by statistician Ronald Fisher.

Interpretation

The *p*-value is the probability that a given result (or a more significant result) would occur under the null hypothesis. For example, say that a fair coin is tested for fairness (the null hypothesis). At a significance level of 0.05, the fair coin would be expected to (incorrectly) reject the null hypothesis in about 1 out of every 20 tests. The *p*-value does not provide the probability that either hypothesis is correct (a common source of confusion). If the *p*-value is less than the chosen significance threshold (equivalently, if the observed test statistic is in the critical region), then we say the null hypothesis is rejected at the chosen level of significance. Rejection of the null hypothesis is a conclusion. This is like a "guilty" verdict in a criminal trial: the evidence is sufficient to reject innocence, thus proving guilt. We might accept the alternative hypothesis (and the research hypothesis). If the *p*-value is *not* less than the chosen significance threshold (equivalently, if the observed test statistic is outside the critical region), then the evidence is insufficient to support a conclusion. (This is similar to a "not guilty" verdict.) The researcher typically gives extra consideration to those cases where the *p*-value is close to the significance level. Some people find it helpful to think of the hypothesis testing framework as analogous to a mathematical proof by contradiction. In the Lady tasting tea example (below), Fisher required the Lady to properly categorize all of the cups of tea to justify the conclusion that the result was unlikely to result from chance. His test revealed that if the lady was effectively guessing at random (the null hypothesis), there was a 1.4 per cent chance that the observed results (perfectly ordered tea) would occur.

Whether rejection of the null hypothesis truly justifies acceptance of the research hypothesis depends on the structure of the hypotheses. Rejecting the hypothesis that a large paw print originated from a bear does not immediately prove the existence of Bigfoot. Hypothesis testing emphasizes the rejection, which is based on a probability, rather than the acceptance, which requires extra steps of logic.

"The probability of rejecting the null hypothesis is a function of five factors: whether the test is one- or two tailed, the level of significance, the standard deviation, the amount of deviation from the null hypothesis, and the number of observations." These factors are a source of criticism; factors under the control of the experimenter/analyst give the results an appearance of subjectivity.

Use and Importance

Statistics are helpful in analyzing most collections of data. This is equally true of hypothesis testing which can justify conclusions even when no scientific theory exists. In the Lady tasting tea example, it was "obvious" that no difference existed between (milk poured into tea) and (tea poured into milk). The data contradicted the "obvious".

Real world applications of hypothesis testing include:

- Testing whether more men than women suffer from nightmares
- Establishing authorship of documents
- Evaluating the effect of the full moon on behaviour
- Determining the range at which a bat can detect an insect by echo
- Deciding whether hospital carpeting results in more infections
- Selecting the best means to stop smoking
- Checking whether bumper stickers reflect car owner behaviour
- Testing the claims of handwriting analysts

Statistical hypothesis testing plays an important role in the whole of statistics and in statistical inference. For example, Lehmann (1992) in a review of the fundamental paper by Neyman and Pearson (1933) says: "Nevertheless, despite their shortcomings, the new paradigm formulated in the 1933 paper, and the many developments carried out within its framework continue to play a central role in both the theory and practice of statistics and can be expected to do so in the foreseeable future". Significance testing has been the favored statistical tool in some experimental social sciences (over 90 per cent of articles in the *Journal of Applied Psychology* during the early 1990s). Other fields have favored the estimation of parameters (*e.g.*, effect size). Significance testing is used as a substitute for the traditional comparison of predicted value and experimental result at the core of the scientific method. When theory is only capable of predicting the sign of a relationship, a directional (one-sided) hypothesis test can be configured so that only a statistically significant result supports theory. This form of theory appraisal is the most heavily criticized application of hypothesis testing.

Cautions

"If the government required statistical procedures to carry warning labels like those on drugs, most inference methods would have long labels indeed." This caution applies to hypothesis tests and alternatives to them.

The successful hypothesis test is associated with a probability and a type-I error rate. The conclusion *might* be wrong.

The conclusion of the test is only as solid as the sample upon which it is based. The design of the experiment is critical. A number of unexpected effects have been observed including:

- The clever Hans effect. A horse appeared to be capable of doing simple arithmetic.
- The Hawthorne effect. Industrial workers were more productive in better illumination, and most productive in worse.
- The placebo effect. Pills with no medically active ingredients were remarkably effective.

A statistical analysis of misleading data produces misleading conclusions. The issue of data quality can be more subtle. In forecasting for example, there is no agreement on a measure of forecast accuracy. In the absence of a consensus measurement, no decision based on measurements will be without controversy.

The book *How to Lie with Statistics* is the most popular book on statistics ever published. It does not much consider hypothesis testing, but its cautions are applicable, including: Many claims are made on the basis of samples too small to convince. If a report does not mention sample size, be doubtful. Hypothesis testing acts as a filter of statistical conclusions; only those results meeting a probability threshold are publishable. Economics also acts as a publication filter; only those results favorable to the author and funding source may be submitted for publication. The impact of filtering on publication is termed publication bias. A related problem is that of multiple testing (sometimes linked to data mining), in which a variety of tests for a variety of possible effects are applied to a single data set and only those yielding a significant result are reported. These are often dealt with by using multiplicity correction procedures that control the family wise error rate (FWER) or the false discovery rate (FDR). Those making critical decisions based on the results of a hypothesis test are prudent to look at the details rather than the conclusion alone. In the physical sciences most results are fully accepted only when independently confirmed. The general advice concerning statistics is, "Figures never lie, but liars figure" (anonymous).

EXAMPLES

Human Sex Ratio

The earliest use of statistical hypothesis testing is generally credited to the question of whether male and female births are equally likely (null hypothesis), which was addressed in the 1700s by John Arbuthnot (1710), and later by Pierre-Simon Laplace (1770s). Arbuthnot examined birth records in London for each of the 82 years from 1629 to 1710, and applied the sign test, a simple non-parametric test. In every year, the number of males born in London exceeded the number of females. Considering more male or more female births as equally likely, the probability of the observed outcome is 0.5, or about 1 in 4,8360,0000,0000,0000,0000,0000; in modern terms, this is the *p*-value. This is vanishingly small, leading Arbuthnot that this was not due to chance, but to divine providence: "From whence it follows, that it is Art, not Chance, that governs." In modern terms, he rejected the null hypothesis of equally likely male and female births at the $p = 1/2$ significance level.

Laplace considered the statistics of almost half a million births. The statistics showed an excess of boys compared to girls. He concluded by calculation of a *p*-value that the excess was a real, but unexplained, effect.

Lady Tasting Tea

In a famous example of hypothesis testing, known as the *Lady tasting tea*, Dr. Muriel Bristol, a female colleague of Fisher claimed to be able to tell whether the tea or the milk was added first to a cup. Fisher proposed to give her eight cups, four of each variety, in random order. One could then ask what the probability was for her getting the number she got correct, but just by chance. The null hypothesis was that the Lady had no such ability. The test statistic was a simple count of the number of successes in selecting the 4 cups. The critical region was the single case of 4 successes of 4 possible based on a conventional probability criterion (< 5 per cent; 1 of 70 ≈ 1.4%). Fisher asserted that no alternative hypothesis was (ever) required. The lady correctly identified every cup, which would be considered a statistically significant result.

Philosopher's Beans

The following example was produced by a philosopher describing scientific methods generations before hypothesis testing was formalized and popularized. Few beans of this handful are white. Most beans in this bag are white. Therefore: Probably, these beans were taken from another bag. This is an hypothetical inference.

The beans in the bag are the population. The handful are the sample. The null hypothesis is that the sample originated from the population. The criterion for rejecting the null-hypothesis is the "obvious" difference in appearance (an informal difference in the mean). The interesting result is that consideration of a real population and a real sample produced an imaginary bag. The philosopher was considering logic rather than probability. To be a real statistical hypothesis test, this example requires the formalities of a probability calculation and a comparison of that probability to a standard. A simple generalization of the example considers a mixed bag of beans and a handful that contain either very few or very many white beans. The generalization considers both extremes. It requires more calculations and more comparisons to arrive at a formal answer, but the core philosophy is unchanged; If the composition of the handful is greatly different from that of the bag, then the sample probably originated from another bag. The original example is termed a one-sided or a one-tailed test while the generalization is termed a two-sided or two-tailed test. The statement also relies on the inference that the sampling was random. If someone had been picking through the bag to find white beans, then it would explain why the handful had so many white beans, and also explain why the number of white beans in the bag was depleted (although the bag is probably intended to be assumed much larger than one's hand).

Radioactive Suitcase

As an example, consider determining whether a suitcase contains some radioactive material. Placed under a Geiger counter, it produces 10 counts per minute. The null hypothesis is that no radioactive material is in the suitcase and that all measured counts are due to ambient radioactivity typical of the surrounding air and harmless objects. We can then calculate how likely it is that we would observe 10 counts per minute if the null hypothesis were true. If the null hypothesis predicts (say) on average 9 counts per minute, then according to the Poisson distribution typical for radioactive decay there is about 41 per cent chance of recording 10 or more counts. Thus we can say that the suitcase is compatible with the null hypothesis (this does not guarantee that there is no radioactive material, just that we don't have enough evidence to suggest there is). On the other hand, if the null hypothesis predicts 3 counts per minute (for which the Poisson distribution predicts only 0.1 per cent chance of recording 10 or more counts) then the suitcase is not compatible with the null hypothesis, and there are likely other factors responsible to produce the measurements.

The test does not directly assert the presence of radioactive material. A *successful* test asserts that the claim of no radioactive material present is unlikely given the reading (and therefore...). The double negative (disproving the null hypothesis) of the method is confusing, but using a counter-example to disprove is standard mathematical practice. The attraction of the method is its practicality. We know (from experience) the expected range of counts with only ambient radioactivity present, so we can say that a measurement is *unusually* large. Statistics just formalizes the intuitive by using numbers instead of adjectives. We probably do not know the characteristics of the radioactive suitcases; We just assume that they produce larger readings. To slightly formalize intuition: radioactivity is suspected if the Geiger-count with the suitcase is among or exceeds the greatest (5% or 1%) of the Geiger-counts made with ambient radiation alone. This makes no assumptions about the distribution of counts. Many ambient radiation observations are required to obtain good probability estimates for rare events.

The test described here is more fully the null-hypothesis statistical significance test. The null hypothesis represents what we would believe by default, before seeing any evidence. Statistical significance is a possible finding of the test, declared when the observed sample is unlikely to have occurred by chance if the null hypothesis were true.

The name of the test describes its formulation and its possible outcome. One characteristic of the test is its crisp decision: to reject or not reject the null hypothesis. A calculated value is compared to a threshold, which is determined from the tolerable risk of error.

DEFINITION OF TERMS

The following definitions are mainly based on the exposition in the book by Lehmann and Romano:

- *Statistical hypothesis:* A statement about the parameters describing a population (not a sample).
- *Statistic:* A value calculated from a sample, often to summarize the sample for comparison purposes.
- *Simple hypothesis:* Any hypothesis which specifies the population distribution completely.
- *Composite hypothesis:* Any hypothesis which does *not* specify the population distribution completely.
- *Null hypothesis (H_0):* A hypothesis associated with a contradiction to a theory one would like to prove.
- *Alternative hypothesis (H_1):* A hypothesis (often composite) associated with a theory one would like to prove.
- *Statistical test:* A procedure whose inputs are samples and whose result is a hypothesis.
- *Region of acceptance:* The set of values of the test statistic for which we fail to reject the null hypothesis.
- *Region of rejection/ Critical region:* The set of values of the test statistic for which the null hypothesis is rejected.
- *Critical value:* The threshold value delimiting the regions of acceptance and rejection for the test statistic.
- *Power of a test (1- β):* The test's probability of correctly rejecting the null hypothesis. The complement of the false negative rate, β. Power is termed sensitivity in biostatistics. ("This is a sensitive test. Because the result is negative, we can confidently say that the patient does not have the condition.")
- *Size:* For simple hypotheses, this is the test's probability of *incorrectly* rejecting the null hypothesis. The false positive rate. For composite hypotheses this is the supremum of the probability of rejecting the null hypothesis over all cases covered by the null hypothesis. The complement of the false positive rate is termed specificity in biostatistics. ("This is a specific test. Because the result is positive, we can confidently say that the patient has the condition.")
- *Significance level of a test (α):* It is the upper bound imposed on the size of a test. Its value is chosen by the statistician prior to looking at the data or choosing any particular test to be used. It is the maximum exposure to erroneously rejecting H_0 he/she is ready to accept. Testing H_0 at significance level α means testing H_0 with a test whose size does not exceed α. In most cases, one uses tests whose size is equal to the significance level.
- *p-value:* The probability, assuming the null hypothesis is true, of observing a result at least as extreme as the test statistic.
- *Statistical significance test:* A predecessor to the statistical hypothesis test. An experimental result was said to be statistically significant if a sample was sufficiently inconsistent with the (null) hypothesis. This was variously considered common sense, a pragmatic heuristic for identifying meaningful experimental results, a convention establishing a threshold of statistical evidence or a method for drawing conclusions from data. The statistical hypothesis test added mathematical rigour and philosophical consistency to the concept by making the alternative hypothesis explicit. The term is loosely used to describe the modern version which is now part of statistical hypothesis testing.
- *Conservative test:* A test is conservative if, when constructed for a given nominal significance level, the true probability of *incorrectly* rejecting the null hypothesis is never greater than the nominal level.
- *Exact test:* A test in which the significance level or critical value can be computed exactly, *i.e.*, without any approximation. In some contexts this term is restricted to tests applied to categorical data and to permutation tests, in which computations are carried out by complete enumeration of all

possible outcomes and their probabilities. A statistical hypothesis test compares a test statistic (z or t for examples) to a threshold. The test statistic (the formula found in the table below) is based on optimality. For a fixed level of Type I error rate, use of these statistics minimizes Type II error rates (equivalent to maximizing power). The following terms describe tests in terms of such optimality:

- *Most powerful test:* For a given *size* or *significance level*, the test with the greatest power (probability of rejection) for a given value of the parameter(s) being tested, contained in the alternative hypothesis.
- *Uniformly most powerful test (UMP):* A test with the greatest *power* for all values of the parameter(s) being tested, contained in the alternative hypothesis.

HISTORY

Early Use

While hypothesis testing was popularized early in the 20th century, early forms were used in the 1700s. The first use is credited to John Arbuthnot (1710), followed by Pierre-Simon Laplace (1770s), in analyzing the human sex ratio at birth.

Modern Origins and Early Controversy

Modern significance testing is largely the product of Karl Pearson (p-value, Pearson's chi-squared test), William Sealy Gosset (Student's t-distribution), and Ronald Fisher ("null hypothesis", analysis of variance, "significance test"), while hypothesis testing was developed by Jerzy Neyman and Egon Pearson (son of Karl). Ronald Fisher began his life in statistics as a Bayesian (Zabell 1992), but Fisher soon grew disenchanted with the subjectivity involved (namely use of the principle of indifference when determining prior probabilities), and sought to provide a more "objective" approach to inductive inference.

Fisher was an agricultural statistician who emphasized rigorous experimental design and methods to extract a result from few samples assuming Gaussian distributions. Neyman (who teamed with the younger Pearson) emphasized mathematical rigour and methods to obtain more results from many samples and a wider range of distributions. Modern hypothesis testing is an inconsistent hybrid of the Fisher vs Neyman/Pearson formulation, methods and terminology developed in the early 20th century.

Fisher popularized the "significance test". He required a null-hypothesis (corresponding to a population frequency distribution) and a sample. His (now familiar) calculations determined whether to reject the null-hypothesis or not. Significance testing did not utilize an alternative hypothesis so there was no concept of a Type II error. The p-value was devised as an informal, but objective, index meant to help a researcher determine (based on other knowledge) whether to modify future experiments or strengthen one's faith in the null hypothesis. Hypothesis testing (and Type I/II errors) was devised by Neyman and Pearson as a more objective alternative to Fisher's p-value, also meant to determine researcher behaviour, but without requiring any inductive inference by the researcher.

Neyman and Pearson considered a different problem (which they called "hypothesis testing"). They initially considered two simple hypotheses (both with frequency distributions). They calculated two probabilities and typically selected the hypothesis associated with the higher probability (the hypothesis more likely to have generated the sample). Their method always selected a hypothesis. It also allowed the calculation of both types of error probabilities.

Fisher and Neyman/Pearson clashed bitterly. Neyman/Pearson considered their formulation to be an improved generalization of significance testing.(The defining paper was abstract. Mathematicians have generalized and refined the theory for decades.) Fisher thought that it was not applicable to scientific research because often,

during the course of the experiment, it is discovered that the initial assumptions about the null hypothesis are questionable due to unexpected sources of error. He believed that the use of rigid reject/accept decisions based on models formulated before data is collected was incompatible with this common scenario faced by scientists and attempts to apply this method to scientific research would lead to mass confusion.

The dispute between Fisher and Neyman–Pearson was waged on philosophical grounds, characterized by a philosopher as a dispute over the proper role of models in statistical inference. Events intervened: Neyman accepted a position in the western hemisphere, breaking his partnership with Pearson and separating disputants (who had occupied the same building) by much of the planetary diameter. World War II provided an intermission in the debate. The dispute between Fisher and Neyman terminated (unresolved after 27 years) with Fisher's death in 1962. Neyman wrote a well-regarded eulogy. Some of Neyman's later publications reported p-values and significance levels. The modern version of hypothesis testing is a hybrid of the two approaches that resulted from confusion by writers of statistical textbooks (as predicted by Fisher) beginning in the 1940s. (But signal detection, for example, still uses the Neyman/Pearson formulation.) Great conceptual differences and many caveats in addition to those mentioned above were ignored. Neyman and Pearson provided the stronger terminology, the more rigorous mathematics and the more consistent philosophy, but the subject taught today in introductory statistics has more similarities with Fisher's method than theirs. This history explains the inconsistent terminology (example: the null hypothesis is never accepted, but there is a region of acceptance). Sometime around 1940, in an apparent effort to provide researchers with a "non-controversial" way to have their cake and eat it too, the authors of statistical text books began anonymously combining these two strategies by using the p-value in place of the test statistic (or data) to test against the Neyman–Pearson "significance level". Thus, researchers were encouraged to infer the strength of their data against some null hypothesis using p-values, while also thinking they are retaining the post-data collection objectivity provided by hypothesis testing. It then became customary for the null hypothesis, which was originally some realistic research hypothesis, to be used almost solely as a strawman "nil" hypothesis (one where a treatment has no effect, regardless of the context).

A Comparison between Fisherian, Frequentist (Neyman–Pearson)

Fisher's null hypothesis testing	Neyman–Pearson decision theory
1. Set up a statistical null hypothesis. The null need not be a nil hypothesis (i.e., zero difference).	1. Set up two statistical hypotheses, H1 and H2, and decide about α, β, and sample size before the experiment, based on subjective cost-benefit considerations. These define a rejection region for each hypothesis.
2. Report the exact level of significance (e.g., p = 0.051 or p = 0.049). Do not use a conventional 5% level, and do not talk about accepting or rejecting hypotheses. If the result is "not significant", draw no conclusions and make no decisions, but suspend judgement until further data is available.	2. If the data falls into the rejection region of H1, accept H2; otherwise accept H1. Note that accepting a hypothesis does not mean that you believe in it, but only that you act as if it were true.
3. Use this procedure only if little is known about the problem at hand, and only to draw provisional conclusions in the context of an attempt to understand the experimental situation.	3. The usefulness of the procedure is limited among others to situations where you have a disjunction of hypotheses (e.g., either $\mu 1 = 8$ or $\mu 2 = 10$ is true) and where you can make meaningful cost-benefit trade-offs for choosing alpha and beta.

Early Choices of Null Hypothesis

Paul Meehl has argued that the epistemological importance of the choice of null hypothesis has gone largely unacknowledged. When the null hypothesis is predicted by theory, a more precise experiment will be a more severe test of the underlying theory.

When the null hypothesis defaults to "no difference" or "no effect", a more precise experiment is a less severe test of the theory that motivated performing the experiment. An examination of the origins of the latter practice may therefore be useful:

1778:Pierre Laplace compares the birthrates of boys and girls in multiple European cities. He states: "it is natural to conclude that these possibilities are very nearly in the same ratio". Thus Laplace's null hypothesis that the birthrates of boys and girls should be equal given "conventional wisdom".

1900:Karl Pearson develops the chi squared test to determine "whether a given form of frequency curve will effectively describe the samples drawn from a given population." Thus the null hypothesis is that a population is described by some distribution predicted by theory. He uses as an example the numbers of five and sixes in the Weldon dice throw data.

1904:Karl Pearson develops the concept of "contingency" in order to determine whether outcomes are independent of a given categorical factor. Here the null hypothesis is by default that two things are unrelated (*e.g.* scar formation and death rates from smallpox). The null hypothesis in this case is no longer predicted by theory or conventional wisdom, but is instead the principle of indifference that led Fisher and others to dismiss the use of "inverse probabilities".

NULL HYPOTHESIS STATISTICAL SIGNIFICANCE TESTING

An example of Neyman–Pearson hypothesis testing can be made by a change to the radioactive suitcase example. If the "suitcase" is actually a shielded container for the transportation of radioactive material, then a test might be used to select among three hypotheses: no radioactive source present, one present, two (all) present. The test could be required for safety, with actions required in each case. The Neyman–Pearson lemma of hypothesis testing says that a good criterion for the selection of hypotheses is the ratio of their probabilities (a likelihood ratio). A simple method of solution is to select the hypothesis with the highest probability for the Geiger counts observed. The typical result matches intuition: few counts imply no source, many counts imply two sources and intermediate counts imply one source. Notice also that usually there are problems for proving a negative. Null hypotheses should be at least falsifiable.

Neyman–Pearson theory can accommodate both prior probabilities and the costs of actions resulting from decisions. The former allows each test to consider the results of earlier tests (unlike Fisher's significance tests). The latter allows the consideration of economic issues (for example) as well as probabilities. A likelihood ratio remains a good criterion for selecting among hypotheses. The two forms of hypothesis testing are based on different problem formulations. The original test is analogous to a true/false question; the Neyman–Pearson test is more like multiple choice. In the view of Tukey the former produces a conclusion on the basis of only strong evidence while the latter produces a decision on the basis of available evidence. While the two tests seem quite different both mathematically and philosophically, later developments lead to the opposite claim. Consider many tiny radioactive sources. The hypotheses become 0,1,2,3... grains of radioactive sand. There is little distinction between none or some radiation (Fisher) and 0 grains of radioactive sand versus all of the alternatives (Neyman–Pearson). The major Neyman–Pearson paper of 1933 also considered composite hypotheses (ones whose distribution includes an unknown parameter). An example proved the optimality of the (Student's) *t*-test, "there can be no better test for the hypothesis under consideration" (p 321). Neyman–Pearson theory was proving the optimality of Fisherian methods from its inception.

Fisher's significance testing has proven a popular flexible statistical tool in application with little mathematical growth potential. Neyman–Pearson hypothesis testing is claimed as a pillar of mathematical statistics, creating a new paradigm for the field. It also stimulated new applications in statistical process control, detection theory, decision theory and game theory. Both formulations have been successful, but the successes have been of a different character.

The dispute over formulations is unresolved. Science primarily uses Fisher's (slightly modified) formulation as taught in introductory statistics. Statisticians study Neyman–Pearson theory in graduate school. Mathematicians are proud of uniting the formulations. Philosophers consider them separately. Learned opinions deem the formulations variously competitive (Fisher vs Neyman), incompatible or complementary. The dispute has become more complex since Bayesian inference has achieved respectability.

The terminology is inconsistent. Hypothesis testing can mean any mixture of two formulations that both changed with time. Any discussion of significance testing vs hypothesis testing is doubly vulnerable to confusion.

Fisher thought that hypothesis testing was a useful strategy for performing industrial quality control, however, he strongly disagreed that hypothesis testing could be useful for scientists. Hypothesis testing provides a means of finding test statistics used in significance testing. The concept of power is useful in explaining the consequences of adjusting the significance level and is heavily used in sample size determination. The two methods remain philosophically distinct. They usually (but *not always*) produce the same mathematical answer. The preferred answer is context dependent. While the existing merger of Fisher and Neyman–Pearson theories has been heavily criticized, modifying the merger to achieve Bayesian goals has been considered.

CRITICISM

Criticism of statistical hypothesis testing fills volumes citing 300–400 primary references. Much of the criticism can be summarized by the following issues:

- The interpretation of a *p*-value is dependent upon stopping rule and definition of multiple comparison. The former often changes during the course of a study and the latter is unavoidably ambiguous. (*i.e.* "p values depend on both the (data) observed and on the other possible (data) that might have been observed but weren't").
- Confusion resulting (in part) from combining the methods of Fisher and Neyman–Pearson which are conceptually distinct.
- Emphasis on statistical significance to the exclusion of estimation and confirmation by repeated experiments.
- Rigidly requiring statistical significance as a criterion for publication, resulting in publication bias. Most of the criticism is indirect. Rather than being wrong, statistical hypothesis testing is misunderstood, overused and misused.
- When used to detect whether a difference exists between groups, a paradox arises. As improvements are made to experimental design (*e.g.*, increased precision of measurement and sample size), the test becomes more lenient. Unless one accepts the absurd assumption that all sources of noise in the data cancel out completely, the chance of finding statistical significance in either direction approaches 100 per cent.
- Layers of philosophical concerns. The probability of statistical significance is a function of decisions made by experimenters/analysts. If the decisions are based on convention they are termed arbitrary or mindless while those not so based may be termed subjective. To minimize type II errors, large samples are recommended. In psychology practically all null hypotheses are claimed to be false for sufficiently large samples so "...it is usually nonsensical to perform an experiment with the *sole* aim of rejecting the null hypothesis.". "Statistically significant findings are often misleading" in psychology. Statistical significance does not imply practical significance and correlation does not imply causation. Casting doubt on the null hypothesis is thus far from directly supporting the research hypothesis.
- "It does not tell us what we want to know". Lists of dozens of complaints are available.

Critics and supporters are largely in factual agreement regarding the characteristics of null hypothesis significance testing (NHST): While it can provide critical information, it is *inadequate as the sole tool for statistical analysis*. *Successfully rejecting the null hypothesis may offer no support for the research hypothesis.* The continuing controversy concerns the selection of the best statistical practices for the near-term future given the (often poor) existing practices. Critics would prefer to ban NHST completely, forcing a complete departure from those practices, while supporters suggest a less absolute change.

Controversy over significance testing, and its effects on publication bias in particular, has produced several results. The American Psychological Association has strengthened its statistical reporting requirements after review, medical journal publishers have recognized the obligation to publish some results that are not statistically significant to combat publication bias and a journal (*Journal of Articles in Support of the Null Hypothesis*) has been created to publish such results exclusively. Textbooks have added some cautions and increased coverage of the tools necessary to estimate the size of the sample required to produce significant results. Major organizations have not abandoned use of significance tests although some have discussed doing so.

ALTERNATIVES

The numerous criticisms of significance testing do not lead to a single alternative. A unifying position of critics is that statistics should not lead to a conclusion or a decision but to a probability or to an estimated value with a confidence interval rather than to an accept-reject decision regarding a particular hypothesis. It is unlikely that the controversy surrounding significance testing will be resolved in the near future. Its supposed flaws and unpopularity do not eliminate the need for an objective and transparent means of reaching conclusions regarding studies that produce statistical results. Critics have not unified around an alternative. Other forms of reporting confidence or uncertainty could probably grow in popularity. One strong critic of significance testing suggested a list of reporting alternatives: effect sizes for importance, prediction intervals for confidence, replications and extensions for replicability, meta-analyses for generality. None of these suggested alternatives produces a conclusion/decision. Lehmann said that hypothesis testing theory can be presented in terms of conclusions/decisions, probabilities, or confidence intervals. "The distinction between the... approaches is largely one of reporting and interpretation." On one "alternative" there is no disagreement: Fisher himself said, "In relation to the test of significance, we may say that a phenomenon is experimentally demonstrable when we know how to conduct an experiment which will rarely fail to give us a statistically significant result." Cohen, an influential critic of significance testing, concurred, "... don't look for a magic alternative to NHST *[null hypothesis significance testing]*... It doesn't exist." "... given the problems of statistical induction, we must finally rely, as have the older sciences, on replication." The "alternative" to significance testing is repeated testing. The easiest way to decrease statistical uncertainty is by obtaining more data, whether by increased sample size or by repeated tests. Nickerson claimed to have never seen the publication of a literally replicated experiment in psychology. An indirect approach to replication is meta-analysis.

Bayesian inference is one proposed alternative to significance testing. (Nickerson cited 10 sources suggesting it, including Rozeboom (1960)). For example, Bayesian parameter estimation can provide rich information about the data from which researchers can draw inferences, while using uncertain priors that exert only minimal influence on the results when enough data is available. Psychologist John K. Kruschke has suggested Bayesian estimation as an alternative for the *t*-test. Alternatively two competing models/hypothesis can be compared using Bayes factors. Bayesian methods could be criticized for requiring information that is seldom available in the cases where significance testing is most heavily used. Neither the prior probabilities nor the probability distribution of the test statistic under the alternative hypothesis are often available in the social sciences.

Advocates of a Bayesian approach sometimes claim that the goal of a researcher is most often to objectively assess the probability that a hypothesis is true based on the data they have collected. Neither Fisher's significance

testing, nor Neyman–Pearson hypothesis testing can provide this information, and do not claim to. The probability a hypothesis is true can only be derived from use of Bayes' Theorem, which was unsatisfactory to both the Fisher and Neyman–Pearson camps due to the explicit use of subjectivity in the form of the prior probability. Fisher's strategy is to sidestep this with the p-value (an objective *index* based on the data alone) followed by *inductive inference*, while Neyman–Pearson devised their approach of *inductive behaviour*.

PHILOSOPHY

Hypothesis testing and philosophy intersect. Inferential statistics, which includes hypothesis testing, is applied probability. Both probability and its application are intertwined with philosophy. Philosopher David Hume wrote, "All knowledge degenerates into probability." Competing practical definitions of probability reflect philosophical differences. The most common application of hypothesis testing is in the scientific interpretation of experimental data, which is naturally studied by the philosophy of science.

Fisher and Neyman opposed the subjectivity of probability. Their views contributed to the objective definitions. The core of their historical disagreement was philosophical. Many of the philosophical criticisms of hypothesis testing are discussed by statisticians in other contexts, particularly correlation does not imply causation and the design of experiments. Hypothesis testing is of continuing interest to philosophers.

EDUCATION

Statistics is increasingly being taught in schools with hypothesis testing being one of the elements taught. Many conclusions reported in the popular press (political opinion polls to medical studies) are based on statistics. Some writers have stated that statistical analysis of this kind allows for thinking clearly about problems involving mass data, as well as the effective reporting of trends and inferences from said data, but caution that writers for a broad public should have a solid understanding of the field in order to use the terms and concepts correctly. An introductory college statistics class places much emphasis on hypothesis testing – perhaps half of the course. Such fields as literature and divinity now include findings based on statistical analysis. An introductory statistics class teaches hypothesis testing as a cookbook process. Hypothesis testing is also taught at the postgraduate level. Statisticians learn how to create good statistical test procedures (like z, Student's t, F and chi-squared). Statistical hypothesis testing is considered a mature area within statistics, but a limited amount of development continues. An academic study states that the cookbook method of teaching introductory statistics leaves no time for history, philosophy or controversy. Hypothesis testing has been taught as received unified method. Surveys showed that graduates of the class were filled with philosophical misconceptions (on all aspects of statistical inference) that persisted among instructors. While the problem was addressed more than a decade ago, and calls for educational reform continue, students still graduate from statistics classes holding fundamental misconceptions about hypothesis testing. Ideas for improving the teaching of hypothesis testing include encouraging students to search for statistical errors in published papers, teaching the history of statistics and emphasizing the controversy in a generally dry subject.

EDUCATIONAL MEASUREMENT

Educational measurement refers to the use of educational assessments and the analysis of data such as scores obtained from educational assessments to infer the abilities and proficiencies of students. The approaches overlap with those in psychometrics. Educational measurement is the assigning of numerals to traits such as achievement, interest, attitudes, aptitudes, intelligence and performance.

The aim of theory and practice in educational measurement is typically to measure abilities and levels of attainment by students in areas such as reading, writing, mathematics, science and so forth. Traditionally,

attention focuses on whether assessments are reliable and valid. In practice, educational measurement is largely concerned with the analysis of data from educational assessments or tests. Typically, this means using total scores on assessments, whether they are multiple choice or open-ended and marked using marking rubrics or guides. In technical terms, the pattern of scores by individual students to individual items is used to infer so-called scale locations of students, the "measurements". This process is one form of scaling. Essentially, higher total scores give higher scale locations, consistent with the traditional and everyday use of total scores. If certain theory is used, though, there is not a strict correspondence between the ordering of total scores and the ordering of scale locations. The Rasch model provides a strict correspondence provided all students attempt the same test items, or their performances are marked using the same marking rubrics. In terms of the broad body of purely mathematical theory drawn on, there is substantial overlap between educational measurement and psychometrics. However, certain approaches considered to be a part of psychometrics, including Classical test theory, Item Response Theory and the Rasch model, were originally developed more specifically for the analysis of data from educational assessments. One of the aims of applying theory and techniques in educational measurement is to try to place the results of different tests administered to different groups of students on a single or common scale through processes known as test equating. The rationale is that because different assessments usually have different difficulties, the total scores cannot be directly compared. The aim of trying to place results on a common scale is to allow comparison of the scale locations inferred from the totals via scaling processes.

FUNDING BIAS

Funding bias, also known as sponsorship bias, funding outcome bias, funding publication bias, and funding effect, refers to the tendency of a scientific study to support the interests of the study's financial sponsor. This phenomenon is recognized sufficiently that researchers undertake studies to examine bias in past published studies. Funding bias has been associated, in particular, with research into chemical toxicity, tobacco, and pharmaceutical drugs. It is an instance of experimenter's bias.

CAUSES

Human Nature

The psychology text *Influence: Science and Practice* describes the act of reciprocity as a trait in which a person feels obliged to return favors. This trait is embodied in all human cultures. Human nature may influence even the most ethical researchers to be affected by their sponsors, although they may genuinely deny it.

Misconduct

Scientific malpractice involving shoddy research or data manipulation does occur in rare instances. Often, however, the quality of manufacturers' studies are at least as good as studies that were not funded by a special interest. Therefore, bias usually occurs for other reasons.

Predetermined Conclusion

Research results can be selected or discarded to support a predetermined conclusion. The tobacco industry, for example, would publish their own internal research that invariably found minimal adverse health effects of passive smoking. A company that hires researchers to perform a study may require the researchers to sign a nondisclosure agreement before they are funded, by which researchers waive their right to release any results

independently and release them only to the sponsor. The sponsor may fund several studies at the same time, suppressing results found contrary to their business interests while publicizing the results that support their interests. Indeed, a review of pharmaceutical studies revealed that research funded by drug companies was less likely to be published, but the drug-company-funded research that was published was more likely to report outcomes favorable to the sponsor.

A double-blind study with only objective measures is less likely to be biased to support a given conclusion. However, the researchers or the sponsors still have opportunities to skew the results by discarding or ignoring undesirable data, qualitatively characterizing the results, and ultimately deciding whether to publish at all. Also, not all studies are possible to conduct double-blind.

Publication Bias

Scientist researcher Anders Sandberg writes that funding bias may be a form of publication bias. Because it is easier to publish positive results than inconclusive or no results, positive results may be correlated with being positive for the sponsor. Outcome reporting bias is related to publication bias and selection bias, in which multiple outcomes are measured but only the significant outcomes are reported, while insignificant or unfavorable outcomes are ignored.

Selection of Subjects or Comparators

Selection bias may result in a non-representative population of test subjects in spite of best efforts to obtain a representative sample. Even a double-blind study may be subject to biased selection of dependent variables, population (via inclusion and exclusion criteria), sample size, statistical methods, or inappropriate comparators, any of which can bias the outcome of a study to favour a particular conclusion.

EXAMPLES

- A 1996 study on the effects of nicotine on cognitive performance revealed that findings of nicotine or smoking improving performance were more likely to be published by scientists who acknowledged tobacco industry support.
- A 2003 study of published research on antidepressants found that studies sponsored by manufacturers of selective serotonin reuptake inhibitors (SSRI) and newer antidepressants tended to favour their products over alternatives when compared to non-industry-funded studies. Also, modelling studies funded by industry were more favorable to industry than studies funded by non-industry sponsors. In general, studies funded by drug companies are four times more likely to favour the drug under trial than studies funded by other sponsors.
- A 2006 review of experimental studies examining the health effects of cell phone use found that studies funded exclusively by industry were least likely to report a statistically significant result.
- The US Food and Drug Administration (FDA) determined in 2008 that the bisphenol A (BPA) in plastic containers is safe when leeched into food, citing chemical industry studies. Independent research studies reached different conclusions, with over 90 percent of them finding health effects from low doses of BPA.
- Two opposing commercial sponsors can be at odds with the published findings of research they sponsor. A 2008 Duke University study on rats, funded by the Sugar Association, found adverse effects of consuming the artificial sweetener Splenda. The manufacturer, Johnson and Johnson subsidiary McNeil Nutritionals LLC, responded by sponsoring its own team of experts to refute the study.

- In 2016, an analysis of studies exploring health effects of sugary soda consumption published between 2001 and 2016 found a 100 per cent probability that a study was funded by sugar-sweetened beverage companies if it found no link between sugar-sweetened beverage consumption and poorer metabolic health. Only 2.9 per cent of studies that found sugary beverages linked to higher rates of diabetes and obesity were underwritten by the sugar-sweetened beverage industry. The authors concluded "This industry seems to be manipulating contemporary scientific processes to create controversy and advance their business interests at the expense of the public's health."
- A 2017 Cochrane review analysis of outcomes of studies pertaining to drugs and medical devices revealed that manufacturing company sponsorship "leads to more favorable results and conclusions than sponsorship by other sources."

REPORTING BIAS

In epidemiology, reporting bias is defined as "selective revealing or suppression of information" by subjects (for example about past medical history, smoking, sexual experiences). In artificial intelligence research, the term reporting bias is used to refer to people's tendency to under-report all the information available. In empirical research, the term may be used to refer to authors under-reporting unexpected or undesirable experimental results, attributing the results to sampling or measurement error, while being more trusting of expected or desirable results, though these may be subject to the same sources of error. In this context, reporting bias can eventually lead to a status quo where multiple investigators discover and discard the same results, and later experimenters justify their own reporting bias by observing that previous experimenters reported different results. Thus, each incident of reporting bias can make future incidents more likely.

REPORTING BIASES IN RESEARCH

Research can only contribute to knowledge if it is communicated from investigators to the community. The generally accepted primary means of communication is "full" publication of the study methods and results in an article published in a scientific journal. Sometimes, investigators choose to present their findings at a scientific meeting as well, either through an oral or poster presentation. These presentations are included as part of the scientific record as brief "abstracts" which may or may not be recorded in publicly accessible documents typically found in libraries or the World Wide Web. Sometimes, investigators fail to publish the results of entire studies. The Declaration of Helsinki[1] and other consensus documents have outlined the ethical obligation to make results from clinical research publicly available.

Reporting bias occurs when the dissemination of research findings is influenced by the nature and direction of the results, for instance in systematic reviews. Positive results is a commonly used term to describe a study finding that one intervention is better than another.

Various attempts have been made to overcome the effects of the reporting biases, including statistical adjustments to the results of published studies. None of these approaches has proved satisfactory, however, and there is increasing acceptance that reporting biases must be tackled by establishing registers of controlled trials and by promoting good publication practice. Until these problems have been addressed, estimates of the effects of treatments based on published evidence may be biased.

CASE STUDY

Litigation brought upon by consumers and health insurers against Pfizer for the fraudulent sales practices in marketing of the drug gabapentin in 2004 revealed a comprehensive publication strategy that employed elements of reporting bias. Spin was used to put emphasis on favorable findings that favored gabapentin, and also to

explain away unfavorable findings towards the drug. In this case, favorable secondary outcomes became the focus over the original primary outcome, which was unfavorable. Other changes found in outcome reporting include the introduction of a new primary outcome, failure to distinguish between primary and secondary outcomes, and failure to report one or more protocol-defined primary outcomes. The decision to publish certain findings in certain journals is another strategy. Trials with statistically significant findings were generally published in academic journals with higher circulation more often than trials with nonsignificant findings. Timing of publication results of trials was influenced, in that the company tried to optimize the timing between the release of two studies. Trials with nonsignificant findings were found to be published in a staggered fashion, as to not have two consecutive trials published without salient findings. Ghost authorship was also an issue, where professional medical writers who drafted the published reports were not properly acknowledged. Fallout from this case is still being settled by Pfizer in 2014, 10 years after the initial litigation.

TYPES OF REPORTING BIAS

Publication Bias

The publication or nonpublication of research findings, depending on the nature and direction of the results. Although medical writers have acknowledged the problem of reporting biases for over a century, it was not until the second half of the 20th century that researchers began to investigate the sources and size of the problem of reporting biases. Over the past two decades, evidence has accumulated that failure to publish research studies, including clinical trials testing intervention effectiveness, is pervasive. Almost all failure to publish is due to failure of the investigator to submit; only a small proportion of studies are not published because of rejection by journals. The most direct evidence of publication bias in the medical field comes from follow-up studies of research projects identified at the time of funding or ethics approval. These studies have shown that "positive findings" is the principal factor associated with subsequent publication: researchers say that the reason they don't write up and submit reports of their research for publication is usually because they are "not interested" in the results (editorial rejection by journals is a rare cause of failure to publish).

Even those investigators who have initially published their results as conference abstracts are less likely to publish their findings in full unless the results are "significant". This is a problem because data presented in abstracts are frequently preliminary or interim results and thus may not be reliable representations of what was found once all data were collected and analyzed. In addition, abstracts are often not accessible to the public through journals, MEDLINE, or easily accessed databases. Many are published in conference programmes, conference proceedings, or on CD-ROM, and are made available only to meeting registrants.

The main factor associated with failure to publish is negative or null findings. Controlled trials that are eventually reported in full are published more rapidly if their results are positive. Publication bias leads to overestimates of treatment effect in meta-analyses, which in turn can lead doctors and decision makers to believe a treatment is more useful than it is. It is now well-established that publication bias is associated with the source of funding for the study.

Time Lag Bias

The rapid or delayed publication of research findings, depending on the nature and direction of the results. In a systematic review of the literature, Hopewell and her colleagues found that overall, trials with "positive results" (statistically significant in favour of the experimental arm) were published about a year sooner than trials with "null or negative results" (not statistically significant or statistically significant in favour of the control arm).

Multiple (Duplicate) Publication Bias

The multiple or singular publication of research findings, depending on the nature and direction of the results. Investigators may also publish the same findings multiple times using a variety of patterns of "duplicate" publication. Many duplicates are published in journal supplements, potentially difficult to access literature. Positive results appear to be published more often in duplicate, which can lead to overestimates of a treatment effect.

Location Bias

The publication of research findings in journals with different ease of access or levels of indexing in standard databases, depending on the nature and direction of results. There is also evidence that, compared to negative or null results, statistically significant results are on average published in journals with greater impact factors, and that publication in the mainstream (non- grey) literature is associated with an overall greater treatment effect compared to the grey literature.

Citation Bias

The citation or non-citation of research findings, depending on the nature and direction of the results. Authors tend to cite positive results over negative or null results, and this has been established over a broad cross section of topics. Differential citation may lead to a perception in the community that an intervention is effective when it is not, and it may lead to over-representation of positive findings in systematic reviews if those left uncited are difficult to locate. Selective pooling of results in a meta-analysis is a form of citation bias that is particularly insidious in its potential to influence knowledge. To minimize bias, pooling of results from similar but separate studies requires an exhaustive search for all relevant studies. That is, a meta-analysis (or pooling of data from multiple studies) must always have emerged from a systematic review (not a selective review of the literature), even though a systematic review does not always have an associated meta-analysis.

Language Bias

The publication of research findings in a particular language, depending on the nature and direction of the results. There is longstanding question about whether there is a language bias such that investigators choose to publish their negative findings in non-English language journals and reserve their positive findings for English language journals. Some research has shown that language restrictions in systematic reviews can change the results of the review and in other cases, authors have not found that such a bias exists.

Knowledge Reporting Bias

The frequency with which people write about actions, outcomes, or properties is not a reflection of real-world frequencies or the degree to which a property is characteristic of a class of individuals. People write about only some parts of the world around them; much of the information is left unsaid.

Outcome Reporting Bias

The selective reporting of some outcomes but not others, depending on the nature and direction of the results. A study may be published in full, but pre-specified outcomes omitted or misrepresented. Efficacy outcomes that are statistically significant have a higher chance of being fully published compared to those that are not statistically significant.

Selective reporting of suspected or confirmed adverse treatment effects is an area for particular concern because of the potential for patient harm. In a study of adverse drug events submitted to Scandinavian drug licensing authorities, reports for published studies were less likely than unpublished studies to record adverse events (for example, 56 vs 77 per cent respectively for Finnish trials involving psychotropic drugs). Recent attention in the lay and scientific media on failure to accurately report adverse events for drugs (*e.g.*, selective serotonin uptake inhibitors, rosiglitazone, rofecoxib) has resulted in additional publications, too numerous to review, indicating substantial selective outcome reporting (mainly suppression) of known or suspected adverse events.

RECALL BIAS

In epidemiological research, recall bias is a systematic error caused by differences in the accuracy or completeness of the recollections retrieved ("recalled") by study participants regarding events or experiences from the past. Sometimes also referred to as response bias, responder bias or reporting bias, this type of measurement bias can be a methodological issue in research involving interviews or questionnaires, in which case it could lead to misclassification of various types of exposure. Recall bias is of particular concern in retrospective studies that use a case-control design to investigate the etiology of a disease or psychiatric condition. For example, in studies of risk factors for breast cancer, women who have had the disease may search their memories more thoroughly than members of the unaffected control group; those who had the disease may recall a greater variety of risk factors they had been exposed to, including those falsely attributed to the disease in the media, such as use of oral contraceptives. To minimize recall bias, some clinical trials have adopted a "wash out period", *i.e.*, a substantial time period that must elapse between the subject's first observation and their subsequent observation of the same event.

OBSERVER-EXPECTANCY EFFECT

The observer-expectancy effect (also called the experimenter-expectancy effect, expectancy bias, observer effect, or experimenter effect) is a form of reactivity in which a researcher's cognitive bias causes them to subconsciously influence the participants of an experiment. Confirmation bias can lead to the experimenter interpreting results incorrectly because of the tendency to look for information that conforms to their hypothesis, and overlook information that argues against it. It is a significant threat to a study's internal validity, and is therefore typically controlled using a double-blind experimental design. An example of the observer-expectancy effect is demonstrated in music backmasking, in which hidden verbal messages are said to be audible when a recording is played backwards. Some people expect to hear hidden messages when reversing songs, and therefore hear the messages, but to others it sounds like nothing more than random sounds. Often when a song is played backwards, a listener will fail to notice the "hidden" lyrics until they are explicitly pointed out, after which they are obvious. Other prominent examples include facilitated communication and dowsing.

In research, experimenter bias occurs when experimenter expectancies regarding study results bias the research outcome. Examples of experimenter bias include conscious or unconscious influences on subject behaviour including creation of demand characteristics that influence subjects, and altered or selective recording of experimental results themselves.

OBSERVER-EXPECTANCY EFFECT

The experimenter may introduce cognitive bias into a study in several ways. In what is called the observer-expectancy effect, the experimenter may subtly communicate their expectations for the outcome of the study to the participants, causing them to alter their behaviour to conform to those expectations. Such observer bias effects are near-universal in human data interpretation under expectation and in the presence of imperfect

cultural and methodological norms that promote or enforce objectivity. The classic example of experimenter bias is that of "Clever Hans" (in German, *der Kluge Hans*), an Orlov Trotterhorse claimed by his owner von Osten to be able to do arithmetic and other tasks. As a result of the large public interest in Clever Hans, philosopher and psychologistCarl Stumpf, along with his assistant Oskar Pfungst, investigated these claims. Ruling out simple fraud, Pfungst determined that the horse could answer correctly even when von Osten did not ask the questions. However, the horse was unable to answer correctly when either it could not see the questioner, or if the questioner themselves was unaware of the correct answer: When von Osten knew the answers to the questions, Hans answered correctly 89 per cent of the time. However, when von Osten did not know the answers, Hans guessed only 6 per cent of questions correctly.

Pfungst then proceeded to examine the behaviour of the questioner in detail, and showed that as the horse's taps approached the right answer, the questioner's posture and facial expression changed in ways that were consistent with an increase in tension, which was released when the horse made the final, correct tap. This provided a cue that the horse had learned to use as a reinforced cue to stop tapping.

Experimenter-bias also influences human subjects. As an example, researchers compared performance of two groups given the same task (rating portrait pictures and estimating how successful each individual was on a scale of -10 to 10), but with different experimenter expectations. In one group, ("Group A"), experimenters were told to expect positive ratings while in another group, ("Group B"), experimenters were told to expect negative ratings. Data collected from Group A was a significant and substantially more optimistic appraisal than the data collected from Group B. The researchers suggested that experimenters gave subtle but clear cues with which the subjects complied.

WHERE BIAS CAN EMERGE

A review of bias in clinical studies concluded that bias can occur at any or all of the seven stages of research. These include:

- Selective background reading
- Specifying and selecting the study sample
- Executing the experimental manoeuvre (or exposure)
- Measuring exposures and outcomes
- Data analysis
- Interpretation and discussion of results
- Publishing the results (or not)

The ultimate source of bias lies in a lack of objectivity. It may occur more often in sociological and medical studies, perhaps due to incentives. Experimenter bias can also be found in some physical sciences, for instance, where an experimenter selectively rounds off measurements. Double blind techniques may be employed to combat bias.

CLASSIFICATION

Modern electronic or computerized data acquisition techniques have greatly reduced the likelihood of such bias, but it can still be introduced by a poorly designed analysis technique. Experimenter's bias was not well recognized until the 1950s and 60s, and then it was primarily in medical experiments and studies. Sackett (1979) catalogued 56 biases that can arise in sampling and measurement in clinical research, among the above-stated first six stages of research. These are as follows:

- In reading-up on the field
  - (a) The biases of rhetoric
  - (b) The "all's well" literature bias

(c) One-sided reference bias
(d) Positive results bias
(e) Hot stuff bias

- In specifying and selecting the study sample
  (a) Popularity bias
  (b) Centripetal bias
  (c) Referral filter bias
  (d) Diagnostic access bias
  (e) Diagnostic suspicion bias
  (f) Unmasking (detection signal) bias
  (g) Mimicry bias
  (h) Previous opinion bias
  (i) Wrong sample size bias
  (j) Admission rate (Berkson) bias
  (k) Prevalence-incidence (Neyman) bias
  (l) Diagnostic vogue bias
  (m) Diagnostic purity bias
  (n) Procedure selection bias
  (o) Missing clinical data bias
  (p) Non-contemporaneous control bias
  (q) Starting time bias
  (r) Unacceptable disease bias
  (s) Migrator bias
  (u) Membership bias
  (v) Non-respondent bias
  (w) Volunteer bias
- In executing the experimental manoeuvre (or exposure)
  (a) Contamination bias
  (b) Withdrawal bias
  (c) Compliance bias
  (d) Therapeutic personality bias
  (e) Bogus control bias
- In measuring exposures and outcomes
  (a) Insensitive measure bias
  (b) Underlying cause bias (rumination bias)
  (c) End-digit preference bias
  (d) Apprehension bias
  (e) Unacceptability bias
  (f) Obsequiousness bias
  (g) Expectation bias
  (h) Substitution game
  (i) Family information bias
  (j) Exposure suspicion bias
  (k) Recall bias
  (l) Attention bias
  (m) Instrument bias

- In analyzing the data
  (a) Post-hoc significance bias
  (b) Data dredging bias (looking for the pony)
  (c) Scale degradation bias
  (d) Tidying-up bias
  (e) Repeated peeks bias
- In interpreting the analysis
  (a) Mistaken identity bias
  (b) Cognitive dissonance bias
  (c) Magnitude bias
  (d) Significance bias
  (e) Correlation bias
  (f) Under-exhaustion bias

PREVENTION

Double blind techniques may be employed to combat bias by causing the experimenter and subject to be ignorant of which condition data flows from. It might be thought that, due to the central limit theorem of statistics, collecting more independent measurements will improve the precision of estimates, thus decreasing bias. However this assumes that the measurements are statistically independent. In the case of experimenter bias, the measures share correlated bias: simply averaging such data will not lead to a better statistic but may merely reflect the correlations among the individual measurements and their non-independent nature.

EXAMPLES

Medical Sciences

In medical sciences, the complexity of living systems and ethical constraints may limit the ability of researchers to perform controlled experiments. In such circumstances scientific knowledge about the phenomenon under study, and the systematic elimination of probable causes of bias, by detecting confounding factors, is the only way to isolate true cause-effect relationships. Experimenter bias in epidemiology has been better studied than in other sciences.

A number of studies into Spiritual Healing illustrate how the design of the study can introduce experimenter bias into the results. A comparison of two studies illustrates that subtle differences in the design of the tests can adversely affect the results of one. The difference was due to the intended result: a positive or negative outcome rather than positive or neutral. A 1995 paper by Hodges and Scofield of spiritual healing used the growth rate of cress seeds as their independent variable in order to eliminate a placebo response or participant bias. The study reported positive results as the test results for each sample were consistent with the healers intention that healing *should* or *should not* occur. However the healer involved in the experiment was a personal acquaintance of the study authors raising the distinct possibility of experimenter bias. A randomized clinical trial, published in 2001, investigated the efficacy of spiritual healing (both at a distance and face-to-face) on the treatment of chronic pain in 120 patients. Healers were observed by "simulated healers" who then mimicked the healers movements on a control group while silently counting backwards in fives - a *neutral* rather than *should not heal* intention. The study found a decrease in pain in all patient groups but "no statistically significant differences between healing and control groups... it was concluded that a specific effect of face-to-face or distant healing on chronic pain could not be demonstrated."

In Physical Sciences

When a signal under study is smaller than the rounding error of measurement and data are over-averaged, a positive result may be found where none exists (*i.e.* a more precise experimental apparatus would conclusively show no signal). For instance a study of variation in sidereal time, subject to rounding of measures by a human who is aware of the measurement value may lead to selectivity in rounding, effectively generating a false signal. In such cases a single-blind experimental protocol is required; if the human observer does not know the sidereal time of the measurements, then even though the round-off is non-random it cannot introduce a spurious sidereal variation.

In Forensic Sciences

Results of a scientific test may be distorted when the underlying data are ambiguous and the scientist is exposed to domain-irrelevant cues which engage emotion. For instance, forensic DNA results are ambiguous, and resolving these ambiguities, particularly when interpreting difficult evidence samples such as those that contain mixtures of DNA from two or more individuals, degraded or inhibited DNA, or limited quantities of DNA template may introduce bias. The full potential of forensic DNA testing can only be realized if observer effects are minimized.

IN SOCIAL SCIENCE

After the data are collected, bias may be introduced during data interpretation and analysis. For example, in deciding which variables to control in analysis, social scientists often face a trade-off between omitted-variable bias and post-treatment bias.

BIAS OF AN ESTIMATOR

In statistics, the bias (or bias function) of an estimator is the difference between this estimator's expected value and the true value of the parameter being estimated. An estimator or decision rule with zero bias is called unbiased. Otherwise the estimator is said to be biased. In statistics, "bias" is an objective property of an estimator, and while not a desired property, it is not pejorative, unlike the ordinary English use of the term "bias".

Bias can also be measured with respect to the median, rather than the mean (expected value), in which case one distinguishes *median*-unbiased from the usual *mean*-unbiasedness property. Bias is related to consistency in that consistent estimators are convergent and *asymptotically* unbiased (hence converge to the correct value as the number of data points grows arbitrarily large), though individual estimators in a consistent sequence may be biased (so long as the bias converges to zero). All else being equal, an unbiased estimator is preferable to a biased estimator, but in practice all else is not equal, and biased estimators are frequently used, generally with small bias. When a biased estimator is used, bounds of the bias are calculated. A biased estimator may be used for various reasons: because an unbiased estimator does not exist without further assumptions about a population or is difficult to compute (as in unbiased estimation of standard deviation); because an estimator is median-unbiased but not mean-unbiased (or the reverse); because a biased estimator gives a lower value of some loss function (particularly mean squared error) compared with unbiased estimators (notably in shrinkage estimators); or because in some cases being unbiased is too strong a condition, and the only unbiased estimators are not useful. Further, mean-unbiasedness is not preserved under non-linear transformations, though median-unbiasedness is; for example, the sample variance is an unbiased estimator for the population variance, but its square root, the sample standard deviation, is a biased estimator for the population standard deviation. These are all illustrated below.

MEDIAN-UNBIASED ESTIMATORS

The theory of median-unbiased estimators was revived by George W. Brown in 1947:

- An estimate of a one-dimensional parameter θ will be said to be median-unbiased, if, for fixed θ, the median of the distribution of the estimate is at the value θ; *i.e.*, the estimate underestimates just as often as it overestimates. This requirement seems for most purposes to accomplish as much as the mean-unbiased requirement and has the additional property that it is invariant under one-to-one transformation.

Further properties of median-unbiased estimators have been noted by Lehmann, Birnbaum, van der Vaart and Pfanzagl. In particular, median-unbiased estimators exist in cases where mean-unbiased and maximum-likelihood estimators do not exist. They are invariant under one-to-one transformations. There are methods of construction median-unbiased estimators for probability distributions that have monotone likelihood-functions, such as one-parameter exponential families, to ensure that they are optimal (in a sense analogous to minimum-variance property considered for mean-unbiased estimators). One such procedure is an analogue of the Rao—Blackwell procedure for mean-unbiased estimators: The procedure holds for a smaller class of probability distributions than does the Rao—Blackwell procedure for mean-unbiased estimation but for a larger class of loss-functions.

BIAS WITH RESPECT TO OTHER LOSS FUNCTIONS

Any minimum-variance *mean*-unbiased estimator minimizes the risk (expected loss) with respect to the squared-error loss function (among mean-unbiased estimators), as observed by Gauss. A minimum-average absolute deviation *median*-unbiased estimator minimizes the risk with respect to the absolute loss function (among median-unbiased estimators), as observed by Laplace. Other loss functions are used in statistics, particularly in robust statistics.

EFFECT OF TRANSFORMATIONS

As stated above, for univariate parameters, median-unbiased estimators remain median-unbiased under transformations that preserve order (or reverse order). Note that, when a transformation is applied to a mean-unbiased estimator, the result need not be a mean-unbiased estimator of its corresponding population statistic. By Jensen's inequality, a convex function as transformation will introduce positive bias, while a concave function will introduce negative bias, and a function of mixed convexity may introduce bias in either direction, depending on the specific function and distribution. That is, for a non-linear function f and a mean-unbiased estimator U of a parameter p, the composite estimator $f(U)$ need not be a mean-unbiased estimator of $f(p)$. For example, the square root of the unbiased estimator of the population variance is *not* a mean-unbiased estimator of the population standard deviation: the square root of the unbiased sample variance, the corrected sample standard deviation, is biased. The bias depends both on the sampling distribution of the estimator and on the transform, and can be quite involved to calculate.

FORECAST BIAS

A forecast bias occurs when there are consistent differences between actual outcomes and previously generated forecasts of those quantities; that is: forecasts may have a general tendency to be too high or too low. A normal property of a good forecast is that it is not biased.

As a quantitative measure, the "forecast bias" can be specified as a probabilistic or statistical property of the forecast error. A typical measure of bias of forecasting procedure is the arithmetic mean or expected value of the forecast errors, but other measures of bias are possible. For example, a median-unbiased forecast would be one where half of the forecasts are too low and half too high.

In contexts where forecasts are being produced on a repetitive basis, the performance of the forecasting system may be monitored using a tracking signal, which provides an automatically maintained summary of the forecasts produced up to any given time. This can be used to monitor for deteriorating performance of the system.

HEALTHY USER BIAS

The healthy user bias is a bias that can damage the validity of epidemiologic studies testing the efficacy of particular therapies or interventions. Specifically, it is a sampling bias: the kind of subjects that voluntarily enroll in a clinical trial and actually follow the experimental regimen are not representative of the general population. They can be expected, on average, to be healthier as they are concerned for their health and are predisposed to follow medical advice, both factors that would aid one's health. In a sense, being healthy or active about one's health is a precondition for becoming a subject of the study, an effect that can appear under other conditions such as studying particular groups of workers (*i.e.* someone in ill health is unlikely to have a job as manual laborer).

INFORMATION BIAS (EPIDEMIOLOGY)

In epidemiology, Information bias refers to bias arising from measurement error. Information bias is also referred to as *observational bias* and *misclassification*. *A Dictionary of Epidemiology*, sponsored by the International Epidemiological Association, defines this as the following:

- A flaw in measuring exposure, covariate, or outcome variables that results in different quality (accuracy) of information between comparison groups. The occurrence of information biases may not be independent of the occurrence of selection biases.
- Bias in an estimate arising from measurement errors."

MISCLASSIFICATION

Misclassification thus refers to measurement error. There are two types of misclassification in epidemiological research: *non- differential* misclassification and *differential* misclassification.

Nondifferential Misclassification

Nondifferential misclassification is when all classes, groups, or categories of a variable (whether exposure, outcome, or covariate) have the same error rate or probability of being misclassified for all study subjects. It has traditionally been assumed that in the case of binary or dichotomous variables nondifferential misclassification would result in an 'underestimation' of the hypothesized relationship between exposure and outcome. However, this has more recently been challenged in that results of individual studies represent a single estimate and not the average of repeated measurements and thus can be farther (or nearer) from the null value (*i.e.* zero) than the true value.

Differential Misclassification

Differential misclassification occurs when the error rate or probability of being misclassified differs across groups of study subjects. For example, the accuracy of blood pressure measurement may be lower for heavier than for lighter study subjects, or a study of elderly persons may find that reports from elderly persons with dementia are less reliable than those without dementia. The effect(s) of such misclassification can vary from an overestimation to an underestimation of the true value. Statisticians have developed methods to adjust for this type of bias, which may assist somewhat in compensating for this problem when known and when it is quantifiable.

LEAD TIME BIAS

Lead time is the length of time between the detection of a disease (usually based on new, experimental criteria) and its usual clinical presentation and diagnosis (based on traditional criteria). It is the time between early diagnosis with screening and the time in which diagnosis would have been made without screening. It is an important factor when evaluating the effectiveness of a specific test.

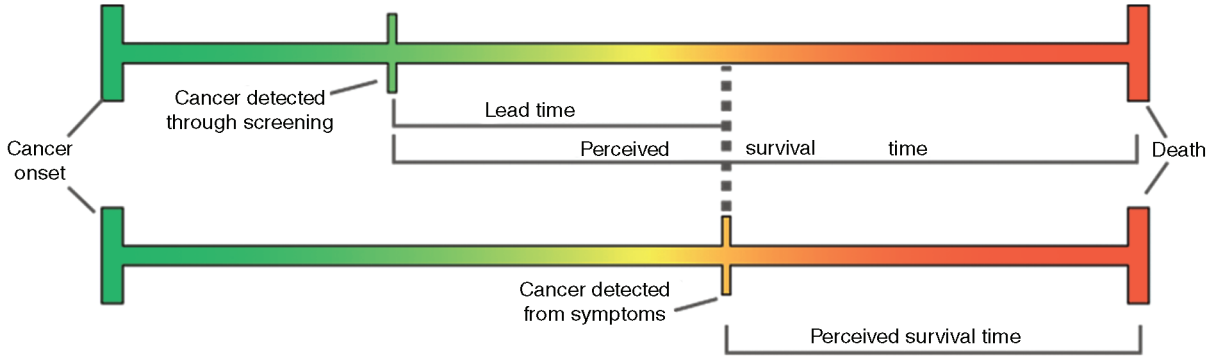


Fig. Lead time bias occurs if testing increases the perceived survival time without affecting the course of the disease.

RELATIONSHIP BETWEEN SCREENING AND SURVIVAL

By screening, the intention is to diagnose a disease earlier than it would be without screening. Without screening, the disease may be discovered later, when symptoms appear.

Early diagnosis by screening may not prolong the life of someone but just determine the propensity of the person to a disease or medical condition such as by DNA testing. No additional life span has been gained and the patient may even be subject to added anxiety as the patient must live for longer with knowledge of the disease. For example, the genetic disorder Huntington's disease is diagnosed when symptoms appear at around 50, and the person dies at around 65. The typical patient, therefore, lives about 15 years after diagnosis. A genetic test at birth makes it possible to diagnose this disorder earlier. If this newborn baby dies at around 65, the person will have "survived" 65 years after diagnosis, without having actually lived any longer than those diagnosed without DNA detection.

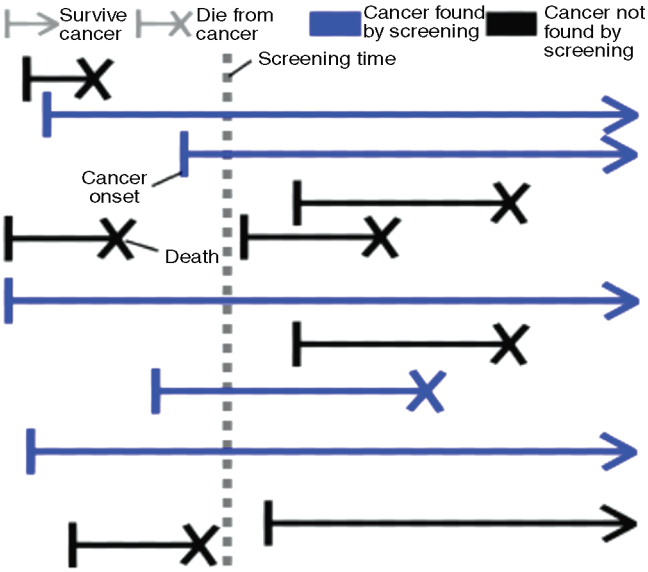
Raw statistics can make screening appear to increase survival time (called lead time). If the person dies at a time in life that previously has been the usual course of the disease than when detected by early screening, the person's life has not been prolonged. Detection by advanced screening does not always mean prolong survival. Lead time bias can affect interpretation of the five-year survival rate.

LENGTH TIME BIAS

Length time bias is a form of selection bias, a statistical distortion of results that can lead to incorrect conclusions about the data. Length time bias can occur when the lengths of intervals are analysed by selecting intervals that occupy randomly chosen points in time or space. That process favors longer intervals and so skews the data.

Length time bias is often discussed in the context of the benefits of cancer screening, and it can lead to the perception that screening leads to better outcomes when in reality it has no effect. Fast-growing tumors generally have a shorter asymptomatic phase than slower-growing tumors. Thus, there is a shorter period of time during which the cancer is present in the body (and so might be detected by screening) but not yet large enough to cause symptoms, that would cause the patient to seek medical care and be diagnosed without screening. As a result, if the same number of slow-growing and fast-growing tumors appear in a year, the screening test detects more slow-growers than fast-growers. If the slow growing tumors are less likely to be fatal than the fast growers,

the people whose cancer is detected by screening do better, on average, than the people whose tumors are detected from symptoms (or at autopsy) even if there is no real benefit to catching the cancer earlier. That can give the impression that detecting cancers by screening causes cancers to be less dangerous even if less dangerous cancers are simply more likely to be detected by screening.



	Death from cancer	Survive cancer	% Surviving cancer
Cancer discovered through screening	1	4	80%
Truly had cancer	7	5	41.7%

Fig. Length time bias in cancer screening. Screening appears to lead to better survival even if no effective treatment is given.

PARTICIPATION BIAS

Participation bias or non-response bias is a phenomenon in which the results of elections, studies, polls, etc. become non-representative because the participants disproportionately possess certain traits which affect the outcome. These traits mean the sample is systematically different from the target population, potentially resulting in biased estimates. For instance, a study found that those who refused to answer a survey on AIDS tended to be "older, attend church more often, are less likely to believe in the confidentiality of surveys, and have lower sexual self disclosure." It may occur due to several factors as outlined in Deming (1990).

Non-response bias can be a problem in longitudinal research due to attrition during the study.

EXAMPLE

If one selects a sample of 1000 managers in a field and polls them about their workload, the managers with a high workload may not answer the survey because they do not have enough time to answer it, and/or those with a low workload may decline to respond for fear that their supervisors or colleagues will perceive them as surplus employees (either immediately, if the survey is non-anonymous, or in the future, should their anonymity be compromised). Therefore, non-response bias may make the measured value for the workload too low, too high, or, if the effects of the above biases happen to offset each other, "right for the wrong reasons." For a simple example of this effect, consider a survey that includes, "Agree or disagree: I have enough time in my day to complete a survey." In the 1936 U.S. presidential election, *The Literary Digest* mailed out 10 million questionnaires, of which 2.3 million were returned. Based on this, they predicted that Republican Alf Landon

would win with 370 of 531 electoral votes; he actually got 8. Research published in 1976 and 1988 concluded that non-response bias was the primary source of this error, although their sampling frame was also quite different from the vast majority of voters. Non- responders have been shown to be associated with younger patients, poorer communities and those who are less satisfied and subsequently could be a source of bias by a study published by Imam et al. in 2014.

TEST

There are different ways to test for non-response bias. A common technique involves comparing the first and fourth quartiles of responses for differences in demographics and key constructs. In e-mail surveys some values are already known from all potential participants (*e.g.* age, branch of the firm,...) and can be compared to the values that prevail in the subgroup of those who answered. If there is no significant difference this is an indicator that there might be no non-response bias.

In e-mail surveys those who didn't answer can also systematically be phoned and a small number of survey questions can be asked. If their answers don't differ significantly from those who answered the survey, there might be no non-response bias. This technique is sometimes called non-response follow-up. Generally speaking, the lower the response rate, the greater the likelihood of a non-response bias in play.

RELATED TERMINOLOGY

- Self-selection bias is a type of bias in which individuals voluntarily select themselves into a group, thereby potentially biasing the response of that group.
- Response bias is not the opposite of non-response bias, but instead relates to a possible tendency of respondents to give inaccurate or untruthful answers for various reasons.

OMITTED-VARIABLE BIAS

In statistics, omitted-variable bias (OVB) occurs when a statistical model leaves out one or more relevant variables. The bias results in the model attributing the effect of the missing variables to the estimated effects of the included variables. More specifically, OVB is the bias that appears in the estimates of parameters in a regression analysis, when the assumed specification is incorrect in that it omits an independent variable that is correlated with both the dependent variable and one or more of the included independent variables.

EFFECT IN ORDINARY LEAST SQUARES

The Gauss–Markov theorem states that regression models which fulfill the classical linear regression model assumptions provide the most efficient, linear and unbiased estimators. In ordinary least squares, the relevant assumption of the classical linear regression model is that the error term is uncorrelated with the regressors.

The presence of omitted-variable bias violates this particular assumption. The violation causes the OLS estimator to be biased and inconsistent. The direction of the bias depends on the estimators as well as the covariance between the regressors and the omitted variables. A positive covariance of the omitted variable with both a regressor and the dependent variable will lead the OLS estimate of the included regressor's coefficient to be greater than the true value of that coefficient. This effect can be seen by taking the expectation of the parameter, as shown in the previous section.

SAMPLING BIAS

In statistics, sampling bias is a bias in which a sample is collected in such a way that some members of the intended population are less likely to be included than others. It results in a biased sample, a non-random

sample of a population (or non-human factors) in which all individuals, or instances, were not equally likely to have been selected. If this is not accounted for, results can be erroneously attributed to the phenomenon under study rather than to the method of sampling. Medical sources sometimes refer to sampling bias as ascertainment bias. Ascertainment bias has basically the same definition, but is still sometimes classified as a separate type of bias.

DISTINCTION FROM SELECTION BIAS

Sampling bias is mostly classified as a subtype of selection bias, sometimes specifically termed sample selection bias, but some classify it as a separate type of bias. A distinction, albeit not universally accepted, of sampling bias is that it undermines the external validity of a test (the ability of its results to be generalized to the entire population), while selection bias mainly addresses internal validity for differences or similarities found in the sample at hand.

In this sense, errors occurring in the process of gathering the sample or cohort cause sampling bias, while errors in any process thereafter cause selection bias. However, selection bias and sampling bias are often used synonymously.

TYPES

- Selection from a specific real area. For example, a survey of high school students to measure teenage use of illegal drugs will be a biased sample because it does not include home-schooled students or dropouts. A sample is also biased if certain members are underrepresented or overrepresented relative to others in the population. For example, a "man on the street" interview which selects people who walk by a certain location is going to have an overrepresentation of healthy individuals who are more likely to be out of the home than individuals with a chronic illness. This may be an extreme form of biased sampling, because certain members of the population are totally excluded from the sample (that is, they have zero probability of being selected).
- Self-selection bias, which is possible whenever the group of people being studied has any form of control over whether to participate (as current standards of human-subject research ethics require for many real-time and some longitudinal forms of study). Participants' decision to participate may be correlated with traits that affect the study, making the participants a non-representative sample. For example, people who have strong opinions or substantial knowledge may be more willing to spend time answering a survey than those who do not. Another example is online and phone-in polls, which are biased samples because the respondents are self-selected. Those individuals who are highly motivated to respond, typically individuals who have strong opinions, are overrepresented, and individuals that are indifferent or apathetic are less likely to respond. This often leads to a polarization of responses with extreme perspectives being given a disproportionate weight in the summary. As a result, these types of polls are regarded as unscientific.
- Pre-screening of trial participants, or advertising for volunteers within particular groups. For example, a study to "prove" that smoking does not affect fitness might recruit at the local fitness center, but advertise for smokers during the advanced aerobics class, and for non-smokers during the weight loss sessions.
- Exclusion bias results from exclusion of particular groups from the sample, *e.g.* exclusion of subjects who have recently migrated into the study area (this may occur when newcomers are not available in a register used to identify the source population). Excluding subjects who move out of the study area during follow-up is rather equivalent of dropout or nonresponse, a selection bias in that it rather affects the internal validity of the study.

- Healthy user bias, when the study population is likely healthier than the general population. For example, someone in poor health is unlikely to have a job as manual laborer.
- Berkson's fallacy, when the study population is selected from a hospital and so is less healthy than the general population. This can result in a spurious negative correlation between diseases: a hospital patient without diabetes is *more* likely to have another given disease such as cholecystitis, since they must have had some reason to enter the hospital in the first place.
- Overmatching, matching for an apparent confounder that actually is a result of the exposure. The control group becomes more similar to the cases in regard to exposure than does the general population.
- Survivorship bias, in which only "surviving" subjects are selected, ignoring those that fell out of view. For example, using the record of current companies as an indicator of business climate or economy ignores the businesses that failed and no longer exist.
- Malmquist bias, an effect in observational astronomy which leads to the preferential detection of intrinsically bright objects.

Symptom-based Sampling

The study of medical conditions begins with anecdotal reports. By their nature, such reports only include those referred for diagnosis and treatment. A child who can't function in school is more likely to be diagnosed with dyslexia than a child who struggles but passes. A child examined for one condition is more likely to be tested for and diagnosed with other conditions, skewing comorbidity statistics. As certain diagnoses become associated with behaviour problems or intellectual disability, parents try to prevent their children from being stigmatized with those diagnoses, introducing further bias. Studies carefully selected from whole populations are showing that many conditions are much more common and usually much milder than formerly believed.

Truncate Selection in Pedigree Studies

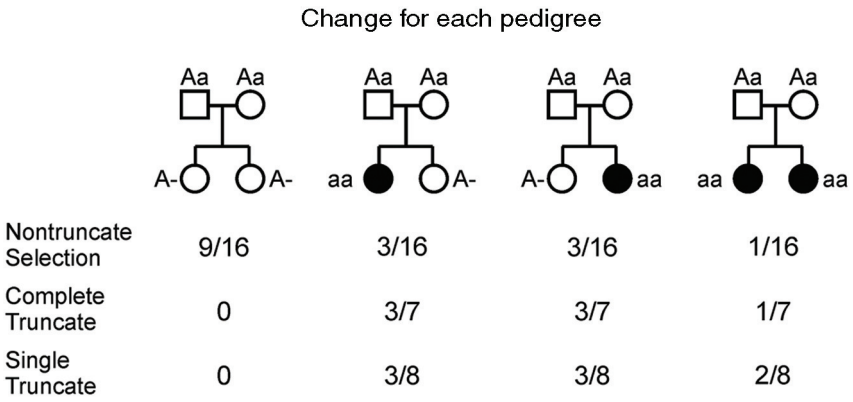


Fig. Simple pedigree example of sampling bias.

Geneticists are limited in how they can obtain data from human populations. As an example, consider a human characteristic. We are interested in deciding if the characteristic is inherited as a simple Mendelian trait. Following the laws of Mendelian inheritance, if the parents in a family do not have the characteristic, but carry the allele for it, they are carriers (*e.g.* a non-expressive heterozygote). In this case their children will each have a 25 per cent chance of showing the characteristic. The problem arises because we can't tell which families have both parents as carriers (heterozygous) unless they have a child who exhibits the characteristic. The description follows the textbook by Sutton.

The figure shows the pedigrees of all the possible families with two children when the parents are carriers (Aa).

- Nontruncate selection. In a perfect world we should be able to discover all such families with a gene including those who are simply carriers. In this situation the analysis would be free from ascertainment bias and the pedigrees would be under "nontruncate selection" In practice, most studies identify, and include, families in a study based upon them having affected individuals.
- Truncate selection. When afflicted *individuals* have an equal chance of being included in a study this is called truncate selection, signifying the inadvertent exclusion (truncation) of families who are carriers for a gene. Because selection is performed on the individual level, families with two or more affected children would have a higher probability of becoming included in the study.
- Complete truncate selection is a special case where each *family* with an affected child has an equal chance of being selected for the study.

The probabilities of each of the families being selected is given in the figure, with the sample frequency of affected children also given. In this simple case, the researcher will look for a frequency of $/_7$ or $/_8$ for the characteristic, depending on the type of truncate selection used.

The Caveman Effect

An example of selection bias is called the "caveman effect". Much of our understanding of prehistoric peoples comes from caves, such as cave paintings made nearly 40,000 years ago. If there had been contemporary paintings on trees, animal skins or hillsides, they would have been washed away long ago. Similarly, evidence of fire pits, middens, burial sites, etc. are most likely to remain intact to the modern era in caves. Prehistoric people are associated with caves because that is where the data still exists, not necessarily because most of them lived in caves for most of their lives.

PROBLEMS DUE TO SAMPLING BIAS

Sampling bias is problematic because it is possible that a statistic computed of the sample is systematically erroneous. Sampling bias can lead to a systematic over- or under-estimation of the corresponding parameter in the population. Sampling bias occurs in practice as it is practically impossible to ensure perfect randomness in sampling. If the degree of misrepresentation is small, then the sample can be treated as a reasonable approximation to a random sample. Also, if the sample does not differ markedly in the quantity being measured, then a biased sample can still be a reasonable estimate. The word bias has a strong negative connotation. Indeed, biases sometimes come from deliberate intent to mislead or other scientific fraud. In statistical usage, bias merely represents a mathematical property, no matter if it is deliberate or unconscious or due to imperfections in the instruments used for observation. While some individuals might deliberately use a biased sample to produce misleading results, more often, a biased sample is just a reflection of the difficulty in obtaining a truly representative sample, or ignorance of the bias in their process of measurement or analysis. An example of how ignorance of a bias can exist is in the widespread use of a ratio (a.k.a. 'fold change) as a measure of difference in biology. Because it is easier to achieve a large ratio with two small numbers with a given difference, and relatively more difficult to achieve a large ratio with two large numbers with a larger difference, large significant differences may be missed when comparing relatively large numeric measurements. Some have called this a 'demarcation bias' because the use of a ratio (division) instead of a difference (subtraction) removes the results of the analysis from science into pseudoscience. Some samples use a biased statistical design which nevertheless allows the estimation of parameters. The U.S. National Center for Health Statistics, for example, deliberately oversamples from minority populations in many of its nationwide surveys in order to gain sufficient precision for estimates within these groups. These surveys require the use of sample weights to produce proper estimates across all ethnic groups. Provided that certain conditions are met (chiefly that the weights are calculated and used correctly) these samples permit accurate estimation of population parameters.

HISTORICAL EXAMPLES

A classic example of a biased sample and the misleading results it produced occurred in 1936. In the early days of opinion polling, the American *Literary Digest* magazine collected over two million postal surveys and predicted that the Republican candidate in the U.S. presidential election, Alf Landon, would beat the incumbent president, Franklin Roosevelt, by a large margin. The result was the exact opposite. The Literary Digest survey represented a sample collected from readers of the magazine, supplemented by records of registered automobile owners and telephone users. This sample included an over-representation of individuals who were rich, who, as a group, were more likely to vote for the Republican candidate. In contrast, a poll of only 50 thousand citizens selected by George Gallup's organization successfully predicted the result, leading to the popularity of the Gallup poll. Another classic example occurred in the 1948 presidential election. On election night, the Chicago Tribune printed the headline *DEWEY DEFEATS TRUMAN*, which turned out to be mistaken. In the morning the grinning president-elect, Harry S. Truman, was photographed holding a newspaper bearing this headline. The reason the Tribune was mistaken is that their editor trusted the results of a phone survey. Survey research was then in its infancy, and few academics realized that a sample of telephone users was not representative of the general population. Telephones were not yet widespread, and those who had them tended to be prosperous and have stable addresses. (In many cities, the Bell System telephone directory contained the same names as the Social Register). In addition, the Gallup poll that the Tribune based its headline on was over two weeks old at the time of the printing.

STATISTICAL CORRECTIONS FOR A BIASED SAMPLE

If entire segments of the population are excluded from a sample, then there are no adjustments that can produce estimates that are representative of the entire population. But if some groups are underrepresented and the degree of underrepresentation can be quantified, then sample weights can correct the bias. However, the success of the correction is limited to the selection model chosen. If certain variables are missing the methods used to correct the bias could be inaccurate. For example, a hypothetical population might include 10 million men and 10 million women. Suppose that a biased sample of 100 patients included 20 men and 80 women. A researcher could correct for this imbalance by attaching a weight of 2.5 for each male and 0.625 for each female. This would adjust any estimates to achieve the same expected value as a sample that included exactly 50 men and 50 women, unless men and women differed in their likelihood of taking part in the survey.

SELF-SELECTION BIAS

In statistics, self-selection bias arises in any situation in which individuals select themselves into a group, causing a biased sample with nonprobability sampling. It is commonly used to describe situations where the characteristics of the people which cause them to select themselves in the group create abnormal or undesirable conditions in the group. It is closely related to the non-response bias, describing when the group of people responding has different responses than the group of people not responding.

Self-selection bias is a major problem in research in sociology, psychology, economics and many other social sciences. In such fields, a poll suffering from such bias is termed a self-selected listener opinion poll or "SLOP". The term is also used in criminology to describe the process by which specific predispositions may lead an offender to choose a criminal career and lifestyle. While the effects of self-selection bias are closely related to those of selection bias, the problem arises for rather different reasons; thus there may be a purposeful intent on the part of respondents leading to self-selection bias whereas other types of selection bias may arise more inadvertently, possibly as the result of mistakes by those designing any given study.

EXPLANATION

Self-selection makes determination of causation more difficult. For example, when attempting to assess the effect of a test preparation course in increasing participant's test scores, significantly higher test scores might be observed among students who choose to participate in the preparation course itself. Due to self-selection, there may be a number of differences between the people who choose to take the course and those who choose not to, such as motivation, socioeconomic status, or prior test-taking experience. Due to self-selection according to such factors, a significant difference in mean test scores could be observed between the two populations independent of any ability of the course to effect higher test scores. An outcome might be that those who elect to do the preparation course would have achieved higher scores in the actual test anyway. If the study measures an improvement in absolute test scores due to participation in the preparation course, they may be skewed to show a higher effect. A relative measure of 'improvement' might improve the reliability of the study somewhat, but only partially.

Self-selection bias causes problems for research about programmes or products. In particular, self-selection affects evaluation of whether or not a given programme has some effect, and complicates interpretation of market research. The Roy model provides one of the earliest academic illustrations of the self-selection problem.

SOCIAL DESIRABILITY BIAS

In social science research, social desirability bias is a type of response bias that is the tendency of survey respondents to answer questions in a manner that will be viewed favorably by others. It can take the form of over-reporting "good behaviour" or under-reporting "bad," or undesirable behaviour. The tendency poses a serious problem with conducting research with self-reports, especially questionnaires. This bias interferes with the interpretation of average tendencies as well as individual differences. Topics where socially desirable responding (SDR) is of special concern are self-reports of abilities, personality, sexual behaviour, and drug use. When confronted with the question "How often do you masturbate?," for example, respondents may be pressured by the societal taboo against masturbation, and either under-report the frequency or avoid answering the question. Therefore, the mean rates of masturbation derived from self-report surveys are likely to be severe underestimates.

When confronted with the question, "Do you use drugs/illicit substances?" the respondent may be influenced by the fact that controlled substances, including the more commonly used marijuana, are generally illegal. Respondents may feel pressured to deny any drug use or rationalize it, *e.g.* "I only smoke marijuana when my friends are around." The bias can also influence reports of number of sexual partners. In fact, the bias may operate in opposite directions for different subgroups: Whereas men tend to inflate the numbers, women tend to underestimate theirs. In either case, the mean reports from both groups are likely to be distorted by social desirability bias.

Other topics that are sensitive to social desirability bias:

- Self-reported personality traits will correlate strongly with social desirability bias
- Personal income and earnings, often inflated when low and deflated when high
- Feelings of low self-worth and/or powerlessness, often denied
- Excretory functions, often approached uncomfortably, if discussed at all
- Compliance with medicinal dosing schedules, often inflated
- Religion, often either avoided or uncomfortably approached
- Patriotism, either inflated or, if denied, done so with a fear of other party's judgement
- Bigotry and intolerance, often denied, even if it exists within the responder
- Intellectual achievements, often inflated
- Physical appearance, either inflated or deflated
- Acts of real or imagined physical violence, often denied

- Indicators of charity or "benevolence," often inflated
- Illegal acts, often denied
- Voter turnout

INDIVIDUAL DIFFERENCES

In 1953, Allen L. Edwards introduced the notion of social desirability to psychology, demonstrating the role of social desirability in the measurement of personality traits. He demonstrated that social desirability ratings of personality trait descriptions are very highly correlated with the probability that a subsequent group of people will endorse these trait self-descriptions. In his first demonstration of this pattern, the correlation between one group of college students' social desirability ratings of a set of traits and the probability that college students in a second group would endorse self-descriptions describing the same traits was so high that it could distort the meaning of the personality traits. In other words, do these self-descriptions describe personality traits or social desirability?

Edwards subsequently developed the first Social Desirability Scale, a set of 39, true-false questions extracted from the Minnesota Multiphasic Personality Inventory (MMPI), questions that judges could, with high agreement, order according to their social desirability. These items were subsequently found to be very highly correlated with a wide range of measurement scales, MMPI personality and diagnostic scales. The SDS is also highly correlated with the Beck Hopelessness Inventory. The fact that people differ in their tendency to engage in socially desirable responding (SDR) is a special concern to those measuring individual differences with self-reports. Individual differences in SDR make it difficult to distinguish those people with good traits who are responding factually from those distorting their answers in a positive direction.

When SDR cannot be eliminated, researchers may resort to evaluating the tendency and then control for it. A separate SDR measure must be administered together with the primary measure (test or interview) aimed at the subject matter of the research/investigation. The key assumption is that respondents who answer in a socially desirable manner on that scale are also responding desirably to all self-reports throughout the study.

In some cases the entire questionnaire package from high scoring respondents may simply be discarded. Alternatively, respondents' answers on the primary questionnaires may be statistically adjusted commensurate with their SDR tendencies. For example, this adjustment is performed automatically in the standard scoring of MMPI scales. The major concern with SDR scales is that they confound style with content. After all, people actually differ in the degree to which they possess desirable traits (*e.g.* nuns versus criminals). Consequently, measures of social desirability confound true differences with social-desirability bias.

STANDARD MEASURES

Until the 1990s, the most commonly used measure of socially desirable responding was the Marlowe–Crowne Social Desirability Scale. The original version comprised 33 True-False items. A shortened version, the Strahan–Gerbasi only comprises ten items, but some have raised questions regarding the reliability of this measure. In 1991, Delroy L. Paulhus published the Balanced Inventory of Desirable Responding (BIDR): a questionnaire designed to measure two forms of SDR. This forty-item instrument provides separate subscales for "impression management," the tendency to give inflated self-descriptions to an audience; and *self-deceptive enhancement*, the tendency to give honest but inflated self-descriptions. The commercial version of the BIDR called "Paulhus Deception Scales (PDS)."

NON-ENGLISH MEASURES

Scales designed to tap response styles are available in all major languages, including Italian and German. Another measure has been used in surveys or opinion polls carried out by interviewing people face-to-face or through the telephone.

OTHER RESPONSE STYLES

"Extreme response style" (ERS) takes the form of exaggerated extremity preference, *e.g.* for '1' or '7' on 7-point scales. Its converse, 'moderacy bias' entails a preference for middle range (or midpoint) responses (*e.g.* 3–5 on 7-point scales).

"Acquiescence" (ARS) is the tendency to respond to items with agreement/affirmation independent of their content ("yea"-saying). These kinds of response styles differ from social desirability bias in that they are unrelated to the question's content and may be present in both socially neutral and in socially favorable or unfavorable contexts, whereas SDR is, by definition, tied to the latter.

ANONYMITY AND CONFIDENTIALITY

When the subjects' details are not required, as in sample investigations and screenings, anonymous administration is preferably used as the person does not feel directly and personally involved in the answers he or she is going to give.

Anonymous self-administration provides neutrality, detachment and reassurance. An even better result is obtained by returning the questionnaires by mail or ballot boxes so as to further guarantee anonymity and the impossibility to identify the subjects who filled in the questionnaires. A further method to assess the prevalence of socially sensitive issues is the so-called randomized response technique. Therein, for example, respondents secretly throw a coin and respond "yes" if it comes up heads, and are instructed to respond truthfully (*e.g.* drug abuse) if it comes up tails. This enables the researcher to estimate the actual prevalence of the given behaviour without needing to know the true state of an individual respondent.

NEUTRALIZED ADMINISTRATION

SDR tends to be reduced by wording questions in a neutral fashion. Another is to use forced-choice questions where the two options have been equated for their desirability. One approach is to administer tests through a computer (self-administration software). A computer, even compared to the most competent interviewer, provides a higher sense of neutrality as it does not appear judgmental.

BEHAVIORAL MEASUREMENT

The most recent approach—the over-claiming technique—assesses the tendency to claim knowledge about non-existent items. More complex methods to promote honest answers include the randomized response and unmatched count techniques, as well as the bogus pipeline technique.

SURVIVORSHIP BIAS

Survivorship bias or survival bias is the logical error of concentrating on the people or things that made it past some selection process and overlooking those that did not, typically because of their lack of visibility. This can lead to false conclusions in several different ways. It is a form of selection bias. Survivorship bias can lead to overly optimistic beliefs because failures are ignored, such as when companies that no longer exist are excluded from analyses of financial performance. It can also lead to the false belief that the successes in a group have some special property, rather than just coincidence (correlation proves causality). For example, if three of the five students with the best college grades went to the same high school, that can lead one to believe that the high school must offer an excellent education. This could be true, but the question cannot be answered without looking at the grades of all the other students from that high school, not just the ones who "survived" the top-five selection process.

EXAMPLES

In Business, Finance and Economics

In finance, survivorship bias is the tendency for failed companies to be excluded from performance studies because they no longer exist. It often causes the results of studies to skew higher because only companies which were successful enough to survive until the end of the period are included. For example, a mutual fund company's selection of funds today will include only those that are successful now. Many losing funds are closed and merged into other funds to hide poor performance. In theory, 90 per cent of extant funds could truthfully claim to have performance in the first quartile of their peers, if the peer group includes funds that have closed.

In 1996, Elton, Gruber, and Blake showed that survivorship bias is larger in the small-fund sector than in large mutual funds (presumably because small funds have a high probability of folding). They estimate the size of the bias across the U.S. mutual fund industry as 0.9 per cent per annum, where the bias is defined and measured as:

- "Bias is defined as average α for surviving funds minus average α for all funds"
- (Where α is the risk-adjusted return over the S&P 500. This is the standard measure of mutual fund out-performance).

Additionally, in quantitative backtesting of market performance or other characteristics, survivorship bias is the use of a current index membership set rather than using the actual constituent changes over time. Consider a backtest to 1990 to find the average performance (total return) of S&P 500 members who have paid dividends within the previous year. To use the current 500 members only and create a historical equity line of the total return of the companies that met the criteria would be adding survivorship bias to the results. S&P maintains an index of healthy companies, removing companies that no longer meet their criteria as a representative of the large-cap U.S. stock market. Companies that had healthy growth on their way to inclusion in the S&P 500 would be counted as if they were in the index during that growth period, which they were not. Instead there may have been another company in the index that was losing market capitalization and was destined for the S&P 600 Small-cap Index that was later removed and would not be counted in the results. Using the actual membership of the index and applying entry and exit dates to gain the appropriate return during inclusion in the index would allow for a bias-free output.

Michael Shermer in *Scientific American* and Larry Smith of the University of Waterloo have described how advice about commercial success distorts perceptions of it by ignoring all of the businesses and college dropouts that failed. Journalist and author David McRaney observes that the "advice business is a monopoly run by survivors. When something becomes a non-survivor, it is either completely eliminated, or whatever voice it has is muted to zero". In his book *The Black Swan*, financial writer Nassim Taleb called the data obscured by survivorship bias "silent evidence."

In History

Diogenes was asked concerning paintings of those who had escaped shipwreck: "Look, you who think the gods have no care of human things, what do you say to so many persons preserved from death by their especial favour?", to which Diogenes replied: "Why, I say that their pictures are not here who were cast away, who are by much the greater number." Susan Mumm has described how survival bias leads historians to study organisations that are still in existence more than those which have closed. This means large, successful organisations such as the Women's Institute, which were well organised and still have accessible archives for historians to work from, are studied more than smaller charitable organisations, even though these may have done a great deal of work.

In Manufacturing and Goods Production

A commonly held opinion in many populations is that machinery, equipment, and goods manufactured in previous generations often is better built and lasts longer than similar contemporary items. (This perception is reflected in the common expression "They don't make 'em [them] like they used to"). Again, because of the selective pressures of time and use, it is inevitable that only those items which were built to last will have survived into the present day. Therefore, most of the old machinery still seen functioning well in the present day must necessarily have been built to a standard of quality necessary to survive. All of the machinery, equipment, and goods that have failed over the intervening years are no longer visible to the general population as they have been junked, scrapped, recycled, or otherwise disposed of. Though survivorship bias may explain a significant portion of the common perception that older manufacturing processes were more rigorous, there are other processes that may explain that perception, such as planned obsolescence and overengineering. It is difficult to directly compare and determine whether manufacturing has become overall better or worse. Manufactured goods are constantly changing, the same items are rarely built for more than a single generation, and even the raw materials change from one era to the next. Capabilities and processes in materials science, technology, manufacturing, and testing have all advanced immensely since the 20th century, undoubtedly raising the potential for similar increases in durability, but pressures on production costs and time have also increased, resulting in manufacturing shortcuts that often result in less durable products. Overall, the contemporary consumer probably has access to and experiences a much wider range of product durability than past generations. Again, bias arises from the fact that historical goods of poor quality are no longer visible, and only the best produced items of the past survive to today.

In Architecture and Construction

Just as new buildings are being built every day and older structures are constantly torn down, the story of most civil and urban architecture involves a process of constant renewal, renovation, and revolution. Only the most (subjectively, but popularly determined) beautiful, most useful, and most structurally sound buildings survive from one generation to the next. This creates another selection effect where the ugliest and weakest buildings of history have long been eradicated from existence and thus the public view, and so it leaves the visible impression, seemingly correct but factually flawed, that all buildings in the past were both more beautiful and better built.

In Highly Competitive Careers

Whether it be movie stars, or athletes, or musicians, or CEOs of multibillion-dollar corporations who dropped out of school, popular media often tells the story of the determined individual who pursues their dreams and beats the odds. There is much less focus on the many people that may be similarly skilled and determined but fail to ever find success because of factors beyond their control or other (seemingly) random events. This creates a false public perception that anyone can achieve great things if they have the ability and make the effort. The overwhelming majority of failures are not visible to the public eye, and only those who survive the selective pressures of their competitive environment are seen regularly.

In the Military

During World War II, the statistician Abraham Wald took survivorship bias into his calculations when considering how to minimize bomber losses to enemy fire. Researchers from the Center for Naval Analyses had conducted a study of the damage done to aircraft that had returned from missions, and had recommended that

armor be added to the areas that showed the most damage. Wald noted that the study only considered the aircraft that had *survived* their missions—the bombers that had been shot down were not present for the damage assessment.

The holes in the returning aircraft, then, represented areas where a bomber could take damage and still return home safely. Wald proposed that the Navy reinforce areas where the returning aircraft were unscathed, since those were the areas that, if hit, would cause the plane to be lost. His work is considered seminal in the then-fledgling discipline of operational research.

In Cats

In a study performed in 1987 it was reported that cats who fall from less than six stories, and are still alive, have greater injuries than cats who fall from higher than six stories. It has been proposed that this might happen because cats reach terminal velocity after righting themselves at about five stories, and after this point they relax, leading to less severe injuries in cats who have fallen from six or more stories. In 2008, *The Straight Dope* newspaper column proposed that another possible explanation for this phenomenon would be survivorship bias. Cats that die in falls are less likely to be brought to a veterinarian than injured cats, and thus many of the cats killed in falls from higher buildings are not reported in studies of the subject.

In Tropical Trees

Tropical vines and lianas are often viewed as macro-parasites of trees that reduce host tree survival. The proportion of trees infested with lianas was observed to be much greater in shade-tolerant, heavy wooded, slow-growing tree species while light-demanding, lighter wooded and fast-growing species are often liana free. Such observations led to the expectation that lianas have stronger negative effects on shade-tolerant species. However, further investigations revealed that liana infestation is far more harmful to light-demanding fast-growing tree species where liana infestation greatly decreases survival such that the observable sample is biased towards those that survived and are liana-free. Hence, the observable sample of trees with lianas in their crown is skewed due to survivorship bias.

AS A GENERAL EXPERIMENTAL FLAW

Survivorship bias (or survivor bias) is a statistical artifact in applications outside finance, where studies on the remaining population are fallaciously compared with the historic average despite the survivors having unusual properties. Mostly, the unusual property in question is a track record of success (like the successful funds). For example, the parapsychology researcher Joseph Banks Rhine believed he had identified the few individuals from hundreds of potential subjects who had powers of ESP. His calculations were based on the improbability of these few subjects guessing the Zener cards shown to a partner by chance.

A major criticism which surfaced against his calculations was the possibility of unconscious survivorship bias in subject selections. He was accused of failing to take into account the large effective size of his sample (all the people he rejected as not being "strong telepaths" because they failed at an earlier testing stage). Had he done this he might have seen that, from the large sample, one or two individuals would probably achieve the track record of success he had found purely by chance.

Writing about the Rhine case in *Fads and Fallacies in the Name of Science*, Martin Gardner explained that he did not think the experimenters had made such obvious mistakes out of statistical naïveté, but as a result of subtly disregarding some poor subjects. He said that, without trickery of any kind, there would always be some people who had improbable success, if a large enough sample were taken. To illustrate this, he speculates about what would happen if one hundred professors of psychology read Rhine's work and decided to make their own

tests; he said that survivor bias would winnow out the typical failed experiments, but encourage the lucky successes to continue testing. He thought that the common null hypothesis (of no result) would not be reported, but:

- Eventually, one experimenter remains whose subject has made high scores for six or seven successive sessions. Neither experimenter nor subject is aware of the other ninety-nine projects, and so both have a strong delusion that ESP is operating.

He concludes:

- The experimenter writes an enthusiastic paper, sends it to Rhine who publishes it in his magazine, and the readers are greatly impressed.

If enough scientists study a phenomenon, some will find statistically significant results by chance, and these are the experiments submitted for publication. Additionally, papers showing positive results may be more appealing to editors. This problem is known as *positive results bias*, a type of publication bias. To combat this, some editors now call for the submission of "negative" scientific findings, where "nothing happened". Survivorship bias is one of the issues discussed in the provocative 2005 paper "Why Most Published Research Findings Are False".

IN BUSINESS LAW

Survivorship bias can raise truth-in-advertising problems when the success rate advertised for a product or service is measured with respect to a population whose makeup differs from that of the target audience whom the company offering that product or service targets with advertising claiming that success rate. These problems become especially significant when

- The advertisement either fails to disclose the existence of relevant differences between the two populations or describes them in insufficient detail;
- These differences result from the company's deliberate "pre-screening" of prospective customers to ensure that only customers with traits increasing their likelihood of success are allowed to purchase the product or service, especially when the company's selection procedures or evaluation standards are kept secret; and
- The company offering the product or service charges a fee, especially one that is non-refundable or not disclosed in the advertisement, for the privilege of attempting to become a customer.

For example, the advertisements of online dating service eHarmony.com pass this test because they fail the first two prongs but not the third: They claim a success rate significantly higher than that of competing services while generally not disclosing that the rate is calculated with respect to a viewership subset who possess traits that increase their likelihood of finding and maintaining relationships and lack traits that pose obstacles to their doing so (1), and the company deliberately selects for these traits by administering a lengthy pre-screening process designed to reject prospective customers who lack the former traits or possess the latter ones (2), but the company does not charge a fee for administration of its pre-screening test, with the effect that its prospective customers face no "downside risk" other than losing the time and expending the effort involved in completing the pre-screening process (negating 3).

Similarly, many investors believe that chance is the main reason that most successful fund managers have the track records they do.

OBSERVATIONAL ERROR

Observational error (or measurement error) is the difference between a measured value of a quantity and its true value. In statistics, an error is not a "mistake". Variability is an inherent part of the results of measurements and of the measurement process.

SCIENCE AND EXPERIMENTS

When either randomness or uncertainty modeled by probability theory is attributed to such errors, they are "errors" in the sense in which that term is used in statistics.

Every time we repeat a measurement with a sensitive instrument, we obtain slightly different results. The common statistical model used is that the error has two additive parts:

- Systematic error which always occurs, with the same value, when we use the instrument in the same way and in the same case, and
- Random error which may vary from observation to another.

Systematic error is sometimes called statistical bias. It may often be reduced with standardized procedures. Part of the learning process in the various sciences is learning how to use standard instruments and protocols so as to minimize systematic error. Random error (or random variation) is due to factors which cannot or will not be controlled. Some possible reason to forgo controlling for these random errors is because it may be too expensive to control them each time the experiment is conducted or the measurements are made. Other reasons may be that whatever we are trying to measure is changing in time or is fundamentally probabilistic. Random error often occurs when instruments are pushed to the extremes of their operating limits. For example, it is common for digital balances to exhibit random error in their least significant digit. Three measurements of a single object might read something like 0.9111g, 0.9110g, and 0.9112g.

RANDOM ERRORS VERSUS SYSTEMATIC ERRORS

Measurement errors can be divided into two components: random error and systematic error.

Random error is always present in a measurement. It is caused by inherently unpredictable fluctuations in the readings of a measurement apparatus or in the experimenter's interpretation of the instrumental reading. Random errors show up as different results for ostensibly the same repeated measurement. They can be estimated by comparing multiple measurements, and reduced by averaging multiple measurements. Systematic error, however, is predictable and typically constant or proportional to the true value. If the cause of the systematic error can be identified, then it usually can be eliminated. Systematic errors are caused by imperfect calibration of measurement instruments or imperfect methods of observation, or interference of the environment with the measurement process, and always affect the results of an experiment in a predictable direction. Incorrect zeroing of an instrument leading to a zero error is an example of systematic error in instrumentation. The Performance Test Standard PTC 19.1-2005 "Test Uncertainty", published by the American Society of Mechanical Engineers (ASME), discusses systematic and random errors in considerable detail. In fact, it conceptualizes its basic uncertainty categories in these terms. Random error can be caused by unpredictable fluctuations in the readings of a measurement apparatus, or in the experimenter's interpretation of the instrumental reading; these fluctuations may be in part due to interference of the environment with the measurement process. The concept of random error is closely related to the concept of precision. The higher the precision of a measurement instrument, the smaller the variability (standard deviation) of the fluctuations in its readings.

SOURCES OF SYSTEMATIC ERROR

Imperfect Calibration

Sources of systematic error may be imperfect calibration of measurement instruments (zero error), changes in the environment which interfere with the measurement process and sometimes imperfect methods of observation can be either zero error or percentage error. If you consider an experimenter taking a reading of the time period of a pendulum swinging past a fiducial marker: If their stop-watch or timer starts with 1 second on

the clock then all of their results will be off by 1 second (zero error). If the experimenter repeats this experiment twenty times (starting at 1 second each time), then there will be a percentage error in the calculated average of their results; the final result will be slightly larger than the true period.

Distance measured by radar will be systematically overestimated if the slight slowing down of the waves in air is not accounted for. Incorrect zeroing of an instrument leading to a zero error is an example of systematic error in instrumentation.

Systematic errors may also be present in the result of an estimate based upon a mathematical model or physical law. For instance, the estimated oscillation frequency of a pendulum will be systematically in error if slight movement of the support is not accounted for.

Quantity

Systematic errors can be either constant, or related (*e.g.* proportional or a percentage) to the actual value of the measured quantity, or even to the value of a different quantity (the reading of a ruler can be affected by environmental temperature). When it is constant, it is simply due to incorrect zeroing of the instrument. When it is not constant, it can change its sign. For instance, if a thermometer is affected by a proportional systematic error equal to 2 per cent of the actual temperature, and the actual temperature is 200°, 0°, or -100°, the measured temperature will be 204° (systematic error = +4°), 0° (null systematic error) or -102° (systematic error = -2°), respectively. Thus, the temperature will be overestimated when it will be above zero, and underestimated when it will be below zero.

Drift

Systematic errors which change during an experiment (drift) are easier to detect. Measurements indicate trends with time rather than varying randomly about a mean. Drift is evident if a measurement of a constant quantity is repeated several times and the measurements drift one way during the experiment. If the next measurement is higher than the previous measurement as may occur if an instrument becomes warmer during the experiment then the measured quantity is variable and it is possible to detect a drift by checking the zero reading during the experiment as well as at the start of the experiment (indeed, the zero reading is a measurement of a constant quantity). If the zero reading is consistently above or below zero, a systematic error is present. If this cannot be eliminated, potentially by resetting the instrument immediately before the experiment then it needs to be allowed by subtracting its (possibly time-varying) value from the readings, and by taking it into account while assessing the accuracy of the measurement. If no pattern in a series of repeated measurements is evident, the presence of fixed systematic errors can only be found if the measurements are checked, either by measuring a known quantity or by comparing the readings with readings made using a different apparatus, known to be more accurate. For example, if you think of the timing of a pendulum using an accurate stopwatch several times you are given readings randomly distributed about the mean. A systematic error is present if the stopwatch is checked against the 'speaking clock' of the telephone system and found to be running slow or fast. Clearly, the pendulum timings need to be corrected according to how fast or slow the stopwatch was found to be running. Measuring instruments such as ammeters and voltmeters need to be checked periodically against known standards.

Systematic errors can also be detected by measuring already known quantities. For example, a spectrometer fitted with a diffraction grating may be checked by using it to measure the wavelength of the D-lines of the sodiumelectromagnetic spectrum which are at 600 nm and 589.6 nm. The measurements may be used to determine the number of lines per millimetre of the diffraction grating, which can then be used to measure the wavelength of any other spectral line. Constant systematic errors are very difficult to deal with as their effects are only

observable if they can be removed. Such errors cannot be removed by repeating measurements or averaging large numbers of results. A common method to remove systematic error is through calibration of the measurement instrument.

SOURCES OF RANDOM ERROR

The random or stochastic error in a measurement is the error that is random from one measurement to the next. Stochastic errors tend to be normally distributed when the stochastic error is the sum of many independent random errors because of the central limit theorem. Stochastic errors added to a regression equation account for the variation in *Y* that cannot be explained by the included *X*s.

SURVEYS

The term "Observational error" is also sometimes used to refer to response errors and some other types of non-sampling error. In survey-type situations, these errors can be mistakes in the collection of data, including both the incorrect recording of a response and the correct recording of a respondent's inaccurate response. These sources of non-sampling error are discussed in Salant and Dillman (1995) and Bland and Altman (1996). These errors can be random or systematic. Random errors are caused by unintended mistakes by respondents, interviewers and/or coders. Systematic error can occur if there is a systematic reaction of the respondents to the method used to formulate the survey question. Thus, the exact formulation of a survey question is crucial, since it affects the level of measurement error (). Different tools are available for the researchers to help them decide about this exact formulation of their questions, for instance estimating the quality of a question using MTMM experiments or predicting this quality using the Survey Quality Predictor software (SQP). This information about the quality can also be used in order to correct for measurement error ().

EFFECT ON REGRESSION ANALYSIS

If the dependent variable in a regression is measured with error, regression analysis and associated hypothesis testing are unaffected, except that the R will be lower than it would be with perfect measurement. However, if one or more independent variables is measured with error, then the regression coefficients and standard hypothesis tests are invalid.

SYSTEMIC BIAS

Systemic bias, also called institutional bias, is the inherent tendency of a process to support particular outcomes. The term generally refers to human systems such as institutions; the equivalent bias in non-human systems (such as measurement instruments or mathematical models used to estimate physical quantities) is often called systematic bias, and leads to systematic error in measurements or estimates. The issues of systemic bias are dealt with extensively in the field of industrial organization economics.

IN HUMAN INSTITUTIONS

Cognitive bias is inherent in the experiences, loyalties, and relationships of people in their daily lives, and new biases are constantly being discovered and addressed on both an ethical and political level. For example, the goal of affirmative action in the United States is to counter biases concerning gender, race, and ethnicity, by opening up institutional participation to people with a wider range of backgrounds, and hence a wider range of points of view. In India, the system of scheduled castes and tribes intends to address systemic bias caused by the controversial caste system, a system centered on organized discrimination based upon one's ancestry, not unlike

the system that affirmative action aims to counter. Both the scheduling system and affirmative action mandate the hiring of citizens from within designated groups. However, without sufficient restrictions based upon the actual socio-economic standing of the recipients of the aid provided, these types of system can, and allegedly do, result in the unintentional institutionalization of a reversed form of the same systemic bias, which works against the goal of rendering institutional participation open to people with a wider range of backgrounds. It can therefore be argued that all human institutions can do is to minimize bias as much as possible, and utilize education to increase awareness of it wherever possible.

MAJOR CAUSES

The study of systemic bias as part of the field titled organizational behaviour in industrial organization economics is studied in several principle modalities in both non-profit and for-profit institutions. The issue of concern is that patterns of behaviour may develop within large institutions which become harmful to the productivity and viability of the larger institutions from which they develop. The three major categories of study for maladaptive organizational behaviour and systemic bias are counterproductive work behaviour, human resource mistreatment, and the amelioration of stress-inducing behaviour.

Counterproductive Work Behaviour

Counterproductive work behaviour, or CWB, consists of behaviour by employees that harms or intends to harm organizations and people in organizations.

Mistreatment of Human Resources

There are several types of mistreatment that employees endure in organizations.

- *Abusive supervision:* Abusive supervision is the extent to which a supervisor engages in a pattern of behaviour that harms subordinates.
- *Bullying:* Although definitions of bullying vary, it involves a repeated pattern of harmful behaviors directed towards an individual.
- *Incivility:* Incivility consists of low-intensity discourteous and rude behaviour with ambiguous intent to harm that violates norms for appropriate behaviour in the workplace.
- *Sexual harassment:* Sexual harassment is behaviour that denigrates or mistreats an individual due to his or her gender, creates an offensive workplace, and interferes with an individual being able to do their job.
- *Stress:* Occupational stress concerns the imbalance between the demands (aspects of the job that require mental or physical effort) and resources that help cope with these demands.

EXAMPLES

Financial Week reported May 5, 2008 (*emphasis added*):

- But we travel in a world with a systemic bias to optimism that typically chooses to avoid the topic of the impending bursting of investment bubbles. Collectively, this is done for career or business reasons. As discussed many times in the investment business, pessimism or realism in the face of probable trouble is just plain bad for business and bad for careers. What I am only slowly realizing, though, is how similar the career risk appears to be for the Fed. It doesn't want to move against bubbles because Congress and business do not like it and show their dislike in unmistakable terms. Even Federal reserve chairmen get bullied and have their faces slapped if they stick to their guns, which will, not surprisingly, be rare since everyone values his career or does not want to be replaced

à la Volcker. So, be as optimistic as possible, be nice to everyone, bail everyone out and hope for the best. If all goes well, after all, you will have a lot of grateful bailees who will happily hire you for $300,000 a pop.

VERSUS SYSTEMATIC BIAS

The difference between the words *systemic* and *systematic* is somewhat ambiguous. "Systemic bias" and the older, more common expression "systematic bias" are often used to refer to the same thing; some users seek to draw a distinction between them, suggesting that systemic bias is most frequently associated with human systems, and related to favoritism. In engineering and computational mechanics, the word *bias* is sometimes used as a synonym of systematic error. In this case, the bias is referred to the result of a measurement or computation, rather than to the measurement instrument or computational method. Some authors try to draw a distinction between systemic and systematic corresponding to that between unplanned and planned, or to that between arising from the characteristics of a system and from an individual flaw. In a less formal sense, *systemic* biases are sometimes said to arise from the nature of the interworkings of the system, whereas *systematic* biases stem from a concerted effort to favour certain outcomes. Consider the difference between affirmative action (systematic) compared to racism and caste (systemic).

VERIFICATION BIAS

In statistics, verification bias is a type of measurement bias in which the results of a diagnostic test affect whether the gold standard procedure is used to verify the test result. This type of bias is also known as "work-up bias" or "referral bias". In clinical practice, verification bias is more likely to occur when a preliminary diagnostic test is negative. Because many gold standard tests can be invasive, expensive, and carry a higher risk (*e.g.* angiography, biopsy, surgery), patients and physicians may be more reluctant to undergo further work-up if a preliminary test is negative. In cohort studies, obtaining a gold standard test on every patient may not always be ethical, practical, or cost effective. These studies can thus be subjected to verification bias. One method to limit verification bias in clinical studies is to perform gold standard testing in a random sample of study participants. In most situations, verification bias introduces a sensitivity estimate that is too high and a specificity that is too low.

WET BIAS

The term wet bias refers to the phenomenon whereby some weather forecasters (usually deliberately) report a higher probability of precipitation (in particular, of rain) than the probability they believe (and the probability borne out by empirical evidence), in order to increase the usefulness and actionability of their forecast. The Weather Channel has been empirically shown, and has also admitted, to having a wet bias in the case of low probability of precipitation (for instance, a 5 per cent probability may be reported as a 20 per cent probability) but not at high probabilities of precipitation (so a 60 per cent probability will be reported as a 60 per cent probability). Some local TV stations have been shown as having significantly greater wet bias, often reporting a 100 per cent probability of precipitation in cases where it rains only 70 per cent of the time.

DISCOVERY

In 2002, Eric Floehr, a computer science graduate of the Ohio State University, started collecting historical data of weather forecasts made by the National Weather Service, The Weather Channel, and AccuWeather for the United States, and collected the data on a web site called ForecastWatch.com. Floehr found that the commercial forecasts were biased: they consistently predicted a higher probability of precipitation than actually occurred.

The National Weather Service forecasts were unbiased, whereas those at The Weather Channel were biased for low probabilities of precipitation: when the Weather Channel predicted a 20 per cent probability of precipitation, it had historically rained only 5 per cent of the time, but a 70 per cent probability of precipitation could be taken at face value. Blogger Dan Allan noted that The Weather Channel is also biased at the upper end: a probability of 90 per cent or higher will be rounded up to 100 per cent. On the other hand, local TV stations tended to exaggerate the probability of precipitation throughout (except when they forecast a probability of 0 per cent, in which case it still rained about 10 per cent of the time). The findings on wet bias, though informally well-known within the weather forecasting community for some time, were first popularized outside the weather forecasting community in Nate Silver's 2012 book *The Signal and the Noise*. The term *wet bias* is used because this is a systematic bias in the direction of the weather being wetter than it actually is.

REASONS FOR WET BIAS

According to Silver, The Weather Channel has openly admitted to deliberately exaggerating the probability of precipitation when it is low. This is because of biased incentives: if the correct low probability of precipitation is given, viewers may interpret the forecast as if there were no probability of rain, and then be upset if it does rain. In other words, The Weather Channel compensates for the people that have greater loss aversion than they think they do, and therefore miscalculate their cost-loss ratio when it is low, by deliberately inflating probabilities. Silver quotes Dr. Rose of The Weather Channel as saying, "If the forecast was objective, if it has zero bias in precipitation, we are in trouble."

6

Confidence Interval, p-value and Null Hypothesis

CONFIDENCE INTERVAL

In statistics, a confidence interval (CI) is a type of interval estimate, computed from the statistics of the observed data, that might contain the true value of an unknown population parameter. The interval has an associated confidence level that, loosely speaking, quantifies the level of confidence that the parameter lies in the interval. More strictly speaking, the confidence level represents the frequency (*i.e.* the proportion) of possible confidence intervals that contain the true value of the unknown population parameter. In other words, if confidence intervals are constructed using a given confidence level from an infinite number of independent sample statistics, the proportion of those intervals that contain the true value of the parameter will be equal to the confidence level.

Confidence intervals consist of a range of potential values of the unknown population parameter. However, the interval computed from a particular sample does not necessarily include the true value of the parameter. Since the observed data are random samples from the true population, the confidence interval obtained from the data is also random.

The confidence level is designated prior to examining the data. Most commonly, the 95 per cent confidence level is used. However, other confidence levels can be used, for example, 90 per cent and 99 per cent. Factors affecting the width of the confidence interval include the size of the sample, the confidence level, and the variability in the sample. A larger sample will, all other things being equal, tend to produce a better estimate of the population parameter. Confidence intervals were introduced to statistics by Jerzy Neyman in a paper published in 1937.

CONCEPTUAL BASIS

Interval estimates can be contrasted with point estimates. A point estimate is a single value given as the estimate of a population parameter that is of interest, for example, the mean of some quantity. An interval estimate specifies instead a range within which the parameter is estimated to lie. Confidence intervals are commonly reported in tables or graphs along with point estimates of the same parameters, to show the reliability of the estimates. For example, a confidence interval can be used to describe how reliable survey results are. In a poll of election–voting intentions, the result might be that 40 per cent of respondents intend to vote for a certain party. A 99 per cent confidence interval for the proportion in the whole population having the same intention on the survey might be 30 per cent to 50 per cent. From the same data one may calculate a 90 per cent

confidence interval, which in this case might be 37 per cent to 43 per cent. A major factor determining the length of a confidence interval is the size of the sample used in the estimation procedure, for example, the number of people taking part in a survey.

Meaning and Interpretation

Various interpretations of a confidence interval can be given (taking the 90 per cent confidence interval as an example in the following).

- The confidence interval can be expressed in terms of samples (or repeated samples): "*Were this procedure to be repeated on numerous samples, the fraction of calculated confidence intervals (which would differ for each sample) that encompass the true population parameter would tend towards 90 per cent.*"
- The confidence interval can be expressed in terms of a single sample: "*There is a 90 per cent probability that the calculated confidence interval from some future experiment encompasses the true value of the population parameter.*" Note this is a probability statement about the confidence interval, not the population parameter. This considers the probability associated with a confidence interval from a pre-experiment point of view, in the same context in which arguments for the random allocation of treatments to study items are made. Here the experimenter sets out the way in which they intend to calculate a confidence interval and to know, before they do the actual experiment, that the interval they will end up calculating has a particular chance of covering the true but unknown value. This is very similar to the "repeated sample" interpretation above, except that it avoids relying on considering hypothetical repeats of a sampling procedure that may not be repeatable in any meaningful sense.
- The explanation of a confidence interval can amount to something like: "*The confidence interval represents values for the population parameter for which the difference between the parameter and the observed estimate is not statistically significant at the 10 per cent level*". In fact, this relates to one particular way in which a confidence interval may be constructed.

In each of the above, the following applies: If the true value of the parameter lies outside the 90 per cent confidence interval, then a sampling event has occurred (namely, obtaining a point estimate of the parameter at least this far from the true parameter value) which had a probability of 10 per cent (or less) of happening by chance.

Misunderstandings

Confidence intervals are frequently misunderstood, and published studies have shown that even professional scientists often misinterpret them.

- A 95 per cent confidence interval does not mean that for a given realized interval there is a 95 per cent probability that the population parameter lies within the interval (*i.e.*, a 95 per cent probability that the interval covers the population parameter). According to the frequentist interpretation, once an experiment is done and an interval calculated, this interval either covers the parameter value or it does not; it is no longer a matter of probability. The 95 per cent probability relates to the reliability of the estimation procedure, not to a specific calculated interval. Neyman himself (the original proponent of confidence intervals) made this point in his original paper:
- "It will be noticed that in the above description, the probability statements refer to the problems of estimation with which the statistician will be concerned in the future. In fact, I have repeatedly stated that the frequency of correct results will tend to α. Consider now the case when a sample is already drawn, and the calculations have given [particular limits]. Can we say that in this particular

case the probability of the true value [falling between these limits] is equal to α? The answer is obviously in the negative. The parameter is an unknown constant, and no probability statement concerning its value may be made..."

Deborah Mayo expands on this further as follows:

- "It must be stressed, however, that having seen the value [of the data], Neyman-Pearson theory never permits one to conclude that the specific confidence interval formed covers the true value of 0 with either (1 - α)100 per cent probability or (1 - α)100 per cent degree of confidence. Seidenfeld's remark seems rooted in a (not uncommon) desire for Neyman-Pearson confidence intervals to provide something which they cannot legitimately provide; namely, a measure of the degree of probability, belief, or support that an unknown parameter value lies in a specific interval. Following Savage (1962), the probability that a parameter lies in a specific interval may be referred to as a measure of final precision. While a measure of final precision may seem desirable, and while confidence levels are often (wrongly) interpreted as providing such a measure, no such interpretation is warranted. Admittedly, such a misinterpretation is encouraged by the word 'confidence'."
- A 95 per cent confidence interval does not mean that 95 per cent of the sample data lie within the interval.
- A confidence interval is not a definitive range of plausible values for the sample parameter, though it may be understood as an estimate of plausible values for the population parameter.
- A particular confidence interval of 95 per cent calculated from an experiment does not mean that there is a 95 per cent probability of a sample parameter from a repeat of the experiment falling within this interval.

Philosophical Issues

The principle behind confidence intervals was formulated to provide an answer to the question raised in statistical inference of how to deal with the uncertainty inherent in results derived from data that are themselves only a randomly selected subset of a population. There are other answers, notably that provided by Bayesian inference in the form of credible intervals. Confidence intervals correspond to a chosen rule for determining the confidence bounds, where this rule is essentially determined before any data are obtained, or before an experiment is done.

The rule is defined such that over all possible datasets that might be obtained, there is a high probability ("high" is specifically quantified) that the interval determined by the rule will include the true value of the quantity under consideration. The Bayesian approach appears to offer intervals that can, subject to acceptance of an interpretation of "probability" as Bayesian probability, be interpreted as meaning that the specific interval calculated from a given dataset has a particular probability of including the true value, conditional on the data and other information available. The confidence interval approach does not allow this since in this formulation and at this same stage, both the bounds of the interval and the true values are fixed values, and there is no randomness involved. On the other hand, the Bayesian approach is only as valid as the prior probability used in the computation, whereas the confidence interval does not depend on assumptions about the prior probability.

The questions concerning how an interval expressing uncertainty in an estimate might be formulated, and of how such intervals might be interpreted, are not strictly mathematical problems and are philosophically problematic. Mathematics can take over once the basic principles of an approach to 'inference' have been established, but it has only a limited role in saying why one approach should be preferred to another: For example, a confidence level of 95 per cent is often used in the biological sciences, but this is a matter of convention or arbitration. In the physical sciences, a much higher level may be used.

Relationship with other Statistical Topics

Statistical Hypothesis Testing

Confidence intervals are closely related to statistical significance testing. For example, if for some estimated parameter θ one wants to test the null hypothesis that $\theta = 0$ against the alternative that $\theta \neq 0$, then this test can be performed by determining whether the confidence interval for θ contains 0. More generally, given the availability of a hypothesis testing procedure that can test the null hypothesis $\theta = \theta_0$ against the alternative that $\theta \neq \theta_0$ for any value of θ_0, then a confidence interval with confidence level $\gamma = 1 - \alpha$ can be defined as containing any number θ_0 for which the corresponding null hypothesis is not rejected at significance level α. If the estimates of two parameters (for example, the mean values of a variable in two independent groups) have confidence intervals that do not overlap, then the difference between the two values is more significant than that indicated by the individual values of α. So, this "test" is too conservative and can lead to a result that is more significant than the individual values of α would indicate. If two confidence intervals overlap, the two means still may be significantly different. Accordingly, and consistent with the Mantel-Haenszel Chi-squared test, is a proposed fix whereby one reduces the error bounds for the two means by multiplying them by the square root of ½ (0.707107) before making the comparison.

While the formulations of the notions of confidence intervals and of statistical hypothesis testing are distinct, they are in some senses related and to some extent complementary. While not all confidence intervals are constructed in this way, one general purpose approach to constructing confidence intervals is to define a $100(1 - \alpha)$ per cent confidence interval to consist of all those values θ_0 for which a test of the hypothesis $\theta = \theta_0$ is not rejected at a significance level of 100α per cent. Such an approach may not always be available since it presupposes the practical availability of an appropriate significance test. Naturally, any assumptions required for the significance test would carry over to the confidence intervals.

It may be convenient to make the general correspondence that parameter values within a confidence interval are equivalent to those values that would not be rejected by a hypothesis test, but this would be dangerous. In many instances the confidence intervals that are quoted are only approximately valid, perhaps derived from "plus or minus twice the standard error," and the implications of this for the supposedly corresponding hypothesis tests are usually unknown. It is worth noting that the confidence interval for a parameter is not the same as the acceptance region of a test for this parameter, as is sometimes thought. The confidence interval is part of the parameter space, whereas the acceptance region is part of the sample space. For the same reason, the confidence level is not the same as the complementary probability of the level of significance.

Confidence Region

Confidence regions generalize the confidence interval concept to deal with multiple quantities. Such regions can indicate not only the extent of likely sampling errors but can also reveal whether (for example) it is the case that if the estimate for one quantity is unreliable, then the other is also likely to be unreliable.

Confidence Band

A confidence band is used in statistical analysis to represent the uncertainty in an estimate of a curve or function based on limited or noisy data. Similarly, a prediction band is used to represent the uncertainty about the value of a new data point on the curve, but subject to noise. Confidence and prediction bands are often used as part of the graphical presentation of results of a regression analysis. Confidence bands are closely related to confidence intervals, which represent the uncertainty in an estimate of a single numerical value.

"As confidence intervals, by construction, only refer to a single point, they are narrower (at this point) than a confidence band which is supposed to hold simultaneously at many points."

DESIRABLE PROPERTIES

When applying standard statistical procedures, there will often be standard ways of constructing confidence intervals. These will have been devised so as to meet certain desirable properties, which will hold given that the assumptions on which the procedure rely are true. These desirable properties may be described as: validity, optimality, and invariance. Of these "validity" is most important, followed closely by "optimality". "Invariance" may be considered as a property of the method of derivation of a confidence interval rather than of the rule for constructing the interval. In non-standard applications, the same desirable properties would be sought.

- *Validity.* This means that the nominal coverage probability (confidence level) of the confidence interval should hold, either exactly or to a good approximation.
- *Optimality.* This means that the rule for constructing the confidence interval should make as much use of the information in the data-set as possible. Recall that one could throw away half of a dataset and still be able to derive a valid confidence interval. One way of assessing optimality is by the length of the interval so that a rule for constructing a confidence interval is judged better than another if it leads to intervals whose lengths are typically shorter.
- *Invariance.* In many applications, the quantity being estimated might not be tightly defined as such. For example, a survey might result in an estimate of the median income in a population, but it might equally be considered as providing an estimate of the logarithm of the median income, given that this is a common scale for presenting graphical results. It would be desirable that the method used for constructing a confidence interval for the median income would give equivalent results when applied to constructing a confidence interval for the logarithm of the median income: specifically the values at the ends of the latter interval would be the logarithms of the values at the ends of former interval.

CONFIDENCE INTERVALS FOR PROPORTIONS AND RELATED QUANTITIES

An approximate confidence interval for a population mean can be constructed for random variables that are not normally distributed in the population, relying on the central limit theorem, if the sample sizes and counts are big enough. The formulae are identical to the case above (where the sample mean is actually normally distributed about the population mean). The approximation will be quite good with only a few dozen observations in the sample if the probability distribution of the random variable is not too different from the normal distribution (*e.g.* its cumulative distribution function does not have any discontinuities and its skewness is moderate).

One type of sample mean is the mean of an indicator variable, which takes on the value 1 for true and the value 0 for false. The mean of such a variable is equal to the proportion that has the variable equal to one (both in the population and in any sample). This is a useful property of indicator variables, especially for hypothesis testing. To apply the central limit theorem, one must use a large enough sample. A rough rule of thumb is that one should see at least 5 cases in which the indicator is 1 and at least 5 in which it is 0. Confidence intervals constructed using the above formulae may include negative numbers or numbers greater than 1, but proportions obviously cannot be negative or exceed 1. Additionally, sample proportions can only take on a finite number of values, so the central limit theorem and the normal distribution are not the best tools for building a confidence interval.

COUNTER-EXAMPLES

Since confidence interval theory was proposed, a number of counter-examples to the theory have been developed to show how the interpretation of confidence intervals can be problematic, at least if one interprets them naïvely.

P-VALUE

In statistical hypothesis testing, the *p*-value or probability value or asymptotic significance is the probability for a given statistical model that, when the null hypothesis is true, the statistical summary (such as the sample mean difference between two compared groups) would be greater or equal to the actual observed results. The use of *p*-values in statistical hypothesis testing is common in many fields of research such as physics, economics, finance, political science, psychology, biology, criminal justice, criminology, and sociology. Their misuse has been a matter of considerable controversy.

Italicisation, capitalisation and hyphenation of the term varies. For example, AMA style uses "*P* value," APA style uses "*p* value," and the American Statistical Association uses "*p*-value."

BASIC CONCEPTS

The *p*-value is used in the context of null hypothesis testing in order to quantify the idea of statistical significance of evidence. Null hypothesis testing is a reductio ad absurdum argument adapted to statistics. In essence, a claim is assumed valid if its counter-claim is improbable. As such, the only hypothesis that needs to be specified in this test and which embodies the counter-claim is referred to as the null hypothesis (that is, the hypothesis to be nullified). A result is said to be statistically significant if it allows us to reject the null hypothesis. That is, as per the reductio ad absurdum reasoning, the statistically significant result should be highly improbable if the null hypothesis is assumed to be true. The rejection of the null hypothesis implies that the correct hypothesis lies in the logical complement of the null hypothesis. However, unless there is a single alternative to the null hypothesis, the rejection of null hypothesis does not tell us which of the alternatives might be the correct one.

MISCONCEPTIONS

There is widespread agreement that *p*-values are often misused and misinterpreted. One practice that has been particularly criticized is accepting the alternative hypothesis for any *p*-value nominally less than.05 without other supporting evidence. Although *p*-values are helpful in assessing how incompatible the data are with a specified statistical model, contextual factors must also be considered, such as "the design of a study, the quality of the measurements, the external evidence for the phenomenon under study, and the validity of assumptions that underlie the data analysis". Another concern is that the *p*-value is often misunderstood as being the probability that the null hypothesis is true. Some statisticians have proposed replacing *p*-values with alternative measures of evidence, such as confidence intervals, likelihood ratios, or Bayes factors, but there is heated debate on the feasibility of these alternatives. Others have suggested to remove fixed significance thresholds and to interpret *p*-values as continuous indices of the strength of evidence against the null hypothesis.

USAGE

The *p*-value is widely used in statistical hypothesis testing, specifically in null hypothesis significance testing. In this method, as part of experimental design, before performing the experiment, one first chooses a model (the null hypothesis) and a threshold value for *p*, called the significance level of the test, traditionally 5 per cent or 1 per cent and denoted as α. If the *p*-value is less than the chosen significance level (α), that suggests that the

observed data is sufficiently inconsistent with the null hypothesis that the null hypothesis may be rejected. However, that does not prove that the tested hypothesis is true. When the *p*-value is calculated correctly, this test guarantees that the type I error rate is at most α. For typical analysis, using the standard $\alpha = 0.05$ cutoff, the null hypothesis is rejected when $p<.05$ and not rejected when $p>.05$. The *p*-value does not, in itself, support reasoning about the probabilities of hypotheses but is only a tool for deciding whether to reject the null hypothesis.

CALCULATION

Usually, X {\displaystyle X} X is a test statistic, rather than any of the actual observations. A test statistic is the output of a scalar function of all the observations. This statistic provides a single number, such as the average or the correlation coefficient, that summarizes the characteristics of the data, in a way relevant to a particular inquiry. As such, the test statistic follows a distribution determined by the function used to define that test statistic and the distribution of the input observational data. For the important case in which the data are hypothesized to follow the normal distribution, depending on the nature of the test statistic and thus the underlying hypothesis of the test statistic, different null hypothesis tests have been developed. Some such tests are z-test for normal distribution, t-test for Student's t-distribution, f-test for f-distribution. When the data do not follow a normal distribution, it can still be possible to approximate the distribution of these test statistics by a normal distribution by invoking the central limit theorem for large samples, as in the case of Pearson's chi-squared test. Thus computing a *p*-value requires a null hypothesis, a test statistic (together with deciding whether the researcher is performing a one-tailed test or a two-tailed test), and data. Even though computing the test statistic on given data may be easy, computing the sampling distribution under the null hypothesis, and then computing its cumulative distribution function (CDF) is often a difficult problem. Today, this computation is done using statistical software, often via numeric methods (rather than exact formulae), but, in the early and mid 20th century, this was instead done via tables of values, and one interpolated or extrapolated *p*-values from these discrete values. Rather than using a table of *p*-values, Fisher instead inverted the CDF, publishing a list of values of the test statistic for given fixed *p*-values; this corresponds to computing the quantile function (inverse CDF).

DISTRIBUTION

When the null hypothesis is true, if it takes the form H_0: $\theta = \theta_0$, and the underlying random variable is continuous, then the probability distribution of the *p*-value is uniform on the interval [0,1]. By contrast, if the alternative hypothesis is true, the distribution is dependent on sample size and the true value of the parameter being studied. The distribution of *p*-values for a group of studies is called a *p*-curve. The curve is affected by four factors: the proportion of studies that examined false null hypotheses, the power of the studies that investigated false null hypotheses, the alpha levels, and publication bias. A *p*-curve can be used to assess the reliability of scientific literature, such as by detecting publication bias or *p*-hacking.

EXAMPLES

Here a few simple examples follow, each illustrating a potential pitfall.

One Roll of a Pair of Dice

Suppose a researcher rolls a pair of dice once and assumes a null hypothesis that the dice are fair, not loaded or weighted towards any specific number/roll/result; uniform. The test statistic is "the sum of the rolled numbers" and is one-tailed. The researcher rolls the dice and observes that both dice show 6, yielding a test statistic of 12. The *p*-value of this outcome is 1/36 (because under the assumption of the null hypothesis, the test statistic is

uniformly distributed) or about 0.028 (the highest test statistic out of 6×6 = 36 possible outcomes). If the researcher assumed a significance level of 0.05, this result would be deemed significant and the hypothesis that the dice are fair would be rejected. In this case, a single roll provides a very weak basis (that is, insufficient data) to draw a meaningful conclusion about the dice. This illustrates the danger with blindly applying *p*-value without considering the experiment design.

Five Heads in a Row

Suppose a researcher flips a coin five times in a row and assumes a null hypothesis that the coin is fair. The test statistic of "total number of heads" can be one-tailed or two-tailed: a one-tailed test corresponds to seeing if the coin is biased towards heads, but a two-tailed test corresponds to seeing if the coin is biased either way. The researcher flips the coin five times and observes heads each time (HHHHH), yielding a test statistic of 5. In a one-tailed test, this is the upper extreme of all possible outcomes, and yields a *p*-value of $(1/2) = 1/32 \approx 0.03$. If the researcher assumed a significance level of 0.05, this result would be deemed significant and the hypothesis that the coin is fair would be rejected. In a two-tailed test, a test statistic of zero heads (TTTTT) is just as extreme and thus the data of HHHHH would yield a *p*-value of $2\times(1/2) = 1/16 \approx 0.06$, which is not significant at the 0.05 level. This demonstrates that specifying a direction (on a symmetric test statistic) halves the *p*-value (increases the significance) and can mean the difference between data being considered significant or not.

Sample Size Dependence

Suppose a researcher flips a coin some arbitrary number of times (n) and assumes a null hypothesis that the coin is fair. The test statistic is the total number of heads and is a two-tailed test. Suppose the researcher observes heads for each flip, yielding a test statistic of n and a *p*-value of 2/2. If the coin was flipped only 5 times, the *p*-value would be $2/32 = 0.0625$, which is not significant at the 0.05 level. But if the coin was flipped 10 times, the *p*-value would be $2/1024 \approx 0.002$, which is significant at the 0.05 level. In both cases the data suggest that the null hypothesis is false (that is, the coin is not fair somehow), but changing the sample size changes the *p*-value. In the first case, the sample size is not large enough to allow the null hypothesis to be rejected at the 0.05 level (in fact, the *p*-value can never be below 0.05 for the coin example). This demonstrates that in interpreting *p*-values, one must also know the sample size, which complicates the analysis.

Alternating Coin Flips

Suppose a researcher flips a coin ten times and assumes a null hypothesis that the coin is fair. The test statistic is the total number of heads and is two-tailed. Suppose the researcher observes alternating heads and tails with every flip (HTHTHTHTHT). This yields a test statistic of 5 and a *p*-value of 1 (completely unexceptional), as that is the expected number of heads.

Suppose instead that the test statistic for this experiment was the "number of alternations" (that is, the number of times when H followed T or T followed H), which is one-tailed. That would yield a test statistic of 9, which is extreme and has a *p*-value of $2/2^9 = 1/256 \approx 0.0039$. That would be considered extremely significant, well beyond the 0.05 level. These data indicate that, in terms of one test statistic, the data set is extremely unlikely to have occurred by chance, but it does not suggest that the coin is biased towards heads or tails. By the first test statistic, the data yield a high *p*-value, suggesting that the number of heads observed is not unlikely. By the second test statistic, the data yield a low *p*-value, suggesting that the pattern of flips observed is very, very unlikely. There is no "alternative hypothesis" (so only rejection of the null hypothesis is possible) and such data could have many causes. The data may instead be forged, or the coin may be flipped by a magician who intentionally alternated outcomes.

This example demonstrates that the *p*-value depends completely on the test statistic used and illustrates that *p*-values can only help researchers to reject a null hypothesis, not consider other hypotheses.

Coin Flipping

As an example of a statistical test, an experiment is performed to determine whether a coin flip is fair (equal chance of landing heads or tails) or unfairly biased (one outcome being more likely than the other). Suppose that the experimental results show the coin turning up heads 14 times out of 20 total flips. The null hypothesis is that the coin is fair, and the test statistic is the number of heads. If a right-tailed test is considered, the *p*-value of this result is the chance of a fair coin landing on heads *at least* 14 times out of 20 flips. That probability can be computed from binomial coefficients as

$$\begin{aligned}&\mathrm{Prob}\,(14\text{ heads})+\mathrm{Prob}(15\text{ heads})+\cdots+\mathrm{Prob}(20\text{ heads})\\&=\frac{1}{2^{20}}\binom{20}{14}+\binom{20}{15}+\cdots+\binom{20}{20}=\frac{60,460}{1,048,576}\approx 0.058\end{aligned}$$

This probability is the *p*-value, considering only extreme results that favour heads. This is called a one-tailed test. However, the deviation can be in either direction, favoring either heads or tails. The two-tailed *p*-value, which considers deviations favoring either heads or tails, may instead be calculated. As the binomial distribution is symmetrical for a fair coin, the two-sided *p*-value is simply twice the above calculated single-sided *p*-value: the two-sided *p*-value is 0.115.

In the above example:

- *Null hypothesis (H_0):* The coin is fair, with Prob(heads) = 0.5
- *Test statistic:* Number of heads
- *Level of significance:* 0.05
- *Observation O:* 14 heads out of 20 flips; and
- Two-tailed *p*-value of observation O given H_0 = 2*min(Prob(no. of heads ≥ 14 heads), Prob(no. of heads ≤ 14 heads))= 2*min(0.058, 0.978) = 2*0.058 = 0.115.

Note that the Prob(no. of heads ≤ 14 heads) = 1 - Prob(no. of heads ≥ 14 heads) + Prob(no. of head = 14) = 1 - 0.058 + 0.036 = 0.978; however, symmetry of the binomial distribution makes it an unnecessary computation to find the smaller of the two probabilities. Here, the calculated *p*-value exceeds 0.05, so the observation is consistent with the null hypothesis, as it falls within the range of what would happen 95 per cent of the time were the coin in fact fair.

Hence, the null hypothesis at the 5 per cent level is not rejected. Although the coin did not fall evenly, the deviation from the expected outcome is small enough to be consistent with chance. However, had one more head been obtained, the resulting *p*-value (two-tailed) would have been 0.0414 (4.14 per cent). The null hypothesis is rejected when a 5 per cent cut-off is used.

HISTORY

Computations of *p*-values date back to the 1700s, where they were computed for the human sex ratio at birth, and used to compute statistical significance compared to the null hypothesis of equal probability of male and female births. John Arbuthnot studied this question in 1710, and examined birth records in London for each of the 82 years from 1629 to 1710. In every year, the number of males born in London exceeded the number of females. Considering more male or more female births as equally likely, the probability of the observed outcome is 0.5, or about 1 in 4,836,000,000,000,000,000,000,000; in modern terms, the *p*-value. This is vanishingly small, leading Arbuthnot that this was not due to chance, but to divine providence: "From whence it follows, that it is Art, not Chance, that governs." In modern terms, he rejected the null hypothesis of equally likely male

and female births at the $p = 1/2$ significance level. This is and other work by Arbuthnot is credited as "… the first use of significance tests …" the first example of reasoning about statistical significance, and "… perhaps the first published report of a nonparametric test …", specifically the sign test.

The same question was later addressed by Pierre-Simon Laplace, who instead used a parametric test, modeling the number of male births with a binomial distribution:

- In the 1770s Laplace considered the statistics of almost half a million births. The statistics showed an excess of boys compared to girls. He concluded by calculation of a *p*-value that the excess was a real, but unexplained, effect.

The *p*-value was first formally introduced by Karl Pearson, in his Pearson's chi-squared test, using the chi-squared distribution and notated as capital P. The *p*-values for the chi-squared distribution (for various values of χ and degrees of freedom), now notated as *P,* was calculated in (Elderton 1902), collected in. The use of the *p*-value in statistics was popularized by Ronald Fisher, and it plays a central role in his approach to the subject. In his influential book *Statistical Methods for Research Workers* (1925), Fisher proposed the level $p = 0.05$, or a 1 in 20 chance of being exceeded by chance, as a limit for statistical significance, and applied this to a normal distribution (as a two-tailed test), thus yielding the rule of two standard deviations (on a normal distribution) for statistical significance. He then computed a table of values, similar to Elderton but, importantly, reversed the roles of χ and *p*.

That is, rather than computing *p* for different values of χ (and degrees of freedom *n*), he computed values of χ that yield specified *p*-values, specifically 0.99, 0.98, 0.95, 0,90, 0.80, 0.70, 0.50, 0.30, 0.20, 0.10, 0.05, 0.02, and 0.01. That allowed computed values of χ to be compared against cutoffs and encouraged the use of *p*-values (especially 0.05, 0.02, and 0.01) as cutoffs, instead of computing and reporting *p*-values themselves. The same type of tables were then compiled in (Fisher and Yates 1938), which cemented the approach. As an illustration of the application of *p*-values to the design and interpretation of experiments, in his following book *The Design of Experiments* (1935), Fisher presented the lady tasting tea experiment, which is the archetypal example of the *p*-value.

In later editions, Fisher explicitly contrasted the use of the *p*-value for statistical inference in science with the Neyman–Pearson method, which he terms "Acceptance Procedures". Fisher emphasizes that while fixed levels such as 5 per cent, 2 per cent, and 1 per cent are convenient, the exact *p*-value can be used, and the strength of evidence can and will be revised with further experimentation. In contrast, decision procedures require a clear-cut decision, yielding an irreversible action, and the procedure is based on costs of error, which, he argues, are inapplicable to scientific research.

RELATED QUANTITIES

A closely related concept is the E-value, which is the expected number of times in multiple testing that one expects to obtain a test statistic at least as extreme as the one that was actually observed if one assumes that the null hypothesis is true. The E-value is the product of the number of tests and the *p*-value.

NULL HYPOTHESIS

In inferential statistics, the null hypothesis is a general statement or default position that there is no relationship between two measured phenomena, or no association among groups. Testing (accepting, approving, rejecting, or disproving) the null hypothesis—and thus concluding that there are or are not grounds for believing that there *is* a relationship between two phenomena (*e.g.* that a potential treatment has a measurable effect)—is a central task in the modern practice of science; the field of statistics gives precise criteria for rejecting a null hypothesis. The null hypothesis is generally assumed to be true until evidence indicates otherwise. In statistics, it is often denoted H_0 (read "H-nought", "H-null", "H-oh", or "H-zero").

The concept of a null hypothesis is used differently in two approaches to statistical inference. In the significance testing approach of Ronald Fisher, a null hypothesis is rejected if the observed data are significantly unlikely to have occurred if the null hypothesis were true. In this case the null hypothesis is rejected and an alternative hypothesis is accepted in its place. If the data are consistent with the null hypothesis, then the null hypothesis is not rejected. In neither case is the null hypothesis or its alternative proven; the null hypothesis is tested with data and a decision is made based on how likely or unlikely the data are. This is analogous to the legal principle of presumption of innocence, in which a suspect or defendant is assumed to be innocent (null is not rejected) until proven guilty (null is rejected) beyond a reasonable doubt (to a statistically significant degree). In the hypothesis testing approach of Jerzy Neyman and Egon Pearson, a null hypothesis is contrasted with an alternative hypothesis and the two hypotheses are distinguished on the basis of data, with certain error rates.

Statistical inference can be done without a null hypothesis, by specifying a statistical model corresponding to each candidate hypothesis and using model selection techniques to choose the most appropriate model. (The most common selection techniques are based on either Akaike information criterion or Bayes factor.)

PRINCIPLE

Hypothesis testing requires constructing a statistical model of what the data would look like, given that chance or random processes alone were responsible for the results. The hypothesis that chance alone is responsible for the results is called the *null hypothesis*. The model of the result of the random process is called the *distribution under the null hypothesis*. The obtained results are then compared with the distribution under the null hypothesis, and the likelihood of finding the obtained results is thereby determined.

Hypothesis testing works by collecting data and measuring how likely the particular set of data is, assuming the null hypothesis is true, when the study is on a randomly selected representative sample. The null hypothesis assumes no relationship between variables in the population from which the sample is selected. If the data-set of a randomly selected representative sample is very unlikely relative to the null hypothesis (defined as being part of a class of sets of data that only rarely will be observed), the experimenter rejects the null hypothesis concluding it (probably) is false. This class of data-sets is usually specified via a test statistic which is designed to measure the extent of apparent departure from the null hypothesis. The procedure works by assessing whether the observed departure measured by the test statistic is larger than a value defined so that the probability of occurrence of a more extreme value is small under the null hypothesis (usually in less than either 5 per cent or 1 per cent of similar data-sets in which the null hypothesis does hold). If the data do not contradict the null hypothesis, then only a weak conclusion can be made: namely, that the observed data set provides no strong evidence against the null hypothesis. In this case, because the null hypothesis could be true or false, in some contexts this is interpreted as meaning that the data give insufficient evidence to make any conclusion; in other contexts it is interpreted as meaning that there is no evidence to support changing from a currently useful regime to a different one. For instance, a certain drug may reduce the chance of having a heart attack. Possible null hypotheses are "this drug does not reduce the chances of having a heart attack" or "this drug has no effect on the chances of having a heart attack". The test of the hypothesis consists of administering the drug to half of the people in a study group as a controlled experiment. If the data show a statistically significant change in the people receiving the drug, the null hypothesis is rejected.

BASIC DEFINITIONS

The *null hypothesis* and the *alternate hypothesis* are types of conjectures used in statistical tests, which are formal methods of reaching conclusions or making decisions on the basis of data. The hypotheses are conjectures about a statistical model of the population, which are based on a sample of the population. The tests are core elements of statistical inference, heavily used in the interpretation of scientific experimental data, to separate

scientific claims from statistical noise. "The statement being tested in a test of statistical significance is called the null hypothesis. The test of significance is designed to assess the strength of the evidence against the null hypothesis. Usually, the null hypothesis is a statement of 'no effect' or 'no difference'." It is often symbolized as H_0. The statement that is being tested against the null hypothesis is the alternative hypothesis. Symbols include H_1 and H_a. Statistical significance test: "Very roughly, the procedure for deciding goes like this: Take a random sample from the population. If the sample data are consistent with the null hypothesis, then do not reject the null hypothesis; if the sample data are inconsistent with the null hypothesis, then reject the null hypothesis and conclude that the alternative hypothesis is true."

The following sections add context and nuance to the basic definitions.

EXAMPLE

Given the test scores of two random samples, one of men and one of women, does one group differ from the other? A possible null hypothesis is that the mean male score is the same as the mean female score:

H_0: $\mu_1 = \mu_2$

where

H_0 = the null hypothesis,

μ_1 = the mean of population 1, and

μ_2 = the mean of population 2.

A stronger null hypothesis is that the two samples are drawn from the same population, such that the variances and shapes of the distributions are also equal.

TERMINOLOGY

- *Simple hypothesis:* Any hypothesis which specifies the population distribution completely. For such a hypothesis the sampling distribution of any statistic is a function of the sample size alone.
- *Composite hypothesis:* Any hypothesis which does *not* specify the population distribution completely. Example: A hypothesis specifying a normal distribution with a specified mean and an unspecified variance.

The simple/composite distinction was made by Neyman and Pearson.

- *Exact hypothesis:* Any hypothesis that specifies an exact parameter value. Example: $\mu = 100$. Synonym: point hypothesis.
- *Inexact hypothesis :* Those specifying a parameter range or interval. Examples: $\mu \leq 100$; $95 \leq \mu \leq 105$.

Fisher required an exact null hypothesis for testing.

A one-tailed hypothesis (tested using a one-sided test) is an inexact hypothesis in which the value of a parameter is specified as being either:

- Above or equal to a certain value, or
- Below or equal to a certain value.

A one-tailed hypothesis is said to have directionality. Fisher's original (lady tasting tea) example was a one-tailed test. The null hypothesis was asymmetric. The probability of guessing all cups correctly was the same as guessing all cups incorrectly, but Fisher noted that only guessing correctly was compatible with the lady's claim.

GOALS OF NULL HYPOTHESIS TESTS

There are many types of significance tests for one, two or more samples, for means, variances and proportions, paired or unpaired data, for different distributions, for large and small samples; all have null hypotheses. There are also at least four goals of null hypotheses for significance tests:

- Technical null hypotheses are used to verify statistical assumptions. For example, the residuals between the data and a statistical model cannot be distinguished from random noise. If true, there is no justification for complicating the model.
- Scientific null assumptions are used to directly advance a theory. For example, the angular momentum of the universe is zero. If not true, the theory of the early universe may need revision.
- Null hypotheses of homogeneity are used to verify that multiple experiments are producing consistent results. For example, the effect of a medication on the elderly is consistent with that of the general adult population. If true, this strengthens the general effectiveness conclusion and simplifies recommendations for use.
- Null hypotheses that assert the equality of effect of two or more alternative treatments, for example, a drug and a placebo, are used to reduce scientific claims based on statistical noise. This is the most popular null hypothesis; It is so popular that many statements about significant testing assume such null hypotheses.

Rejection of the null hypothesis is *not necessarily* the real goal of a significance tester. An adequate statistical model may be associated with a failure to reject the null; the model is adjusted until the null is not rejected. The numerous uses of significance testing were well known to Fisher who discussed many in his book written a decade before defining the null hypothesis. A statistical significance test shares much mathematics with a confidence interval. They are mutually illuminating. A result is often significant when there is confidence in the sign of a relationship (the interval does not include 0). Whenever the sign of a relationship is important, statistical significance is a worthy goal. This also reveals weaknesses of significance testing: A result can be significant without a good estimate of the strength of a relationship; significance can be a modest goal. A weak relationship can also achieve significance with enough data. Reporting both significance and confidence intervals is commonly recommended. The varied uses of significance tests reduce the number of generalizations that can be made about all applications.

CHOICE OF THE NULL HYPOTHESIS

The choice of the null hypothesis is associated with sparse and inconsistent advice. Fisher mentioned few constraints on the choice and stated that many null hypotheses should be considered and that many tests are possible for each. The variety of applications and the diversity of goals suggests that the choice can be complicated. In many applications the formulation of the test is traditional. A familiarity with the range of tests available may suggest a particular null hypothesis and test. Formulating the null hypothesis is not automated (though the calculations of significance testing usually are). Sir David Cox has said, "How [the] translation from subject-matter problem to statistical model is done is often the most critical part of an analysis".

A statistical significance test is intended to test a hypothesis. If the hypothesis summarizes a set of data, there is no value in testing the hypothesis on that set of data. Example: If a study of last year's weather reports indicates that rain in a region falls primarily on weekends, it is only valid to test that null hypothesis on weather reports from any *other* year. Testing hypotheses suggested by the data is circular reasoning that proves nothing; It is a special limitation on the choice of the null hypothesis.

A routine procedure is as follows: Start from the scientific hypothesis. Translate this to a statistical alternative hypothesis and proceed: "Because H_a expresses the effect that we wish to find evidence for, we often begin with H_a and then set up H_0 as the statement that the hoped-for effect is not present." This advice is *reversed* for modeling applications where we hope not to find evidence against the null. A complex case example is as follows: The gold standard in clinical research is the randomized placebo-controlled double-blind clinical trial. But testing a new drug against a (medically ineffective) placebo may be unethical for a serious illness. Testing a new drug against an older medically effective drug raises fundamental philosophical issues regarding the goal

of the test and the motivation of the experimenters. The standard "no difference" null hypothesis may reward the pharmaceutical company for gathering inadequate data. "Difference" is a better null hypothesis in this case, but statistical significance is not an adequate criterion for reaching a nuanced conclusion which requires a good numeric estimate of the drug's effectiveness. A "minor" or "simple" proposed change in the null hypothesis ((new vs old) rather than (new vs placebo)) can have a dramatic effect on the utility of a test for complex non-statistical reasons.

Directionality

The choice of null hypothesis (H_0) and consideration of directionality is critical.

Tailedness of the null-hypothesis test

Consider the question of whether a tossed coin is fair (*i.e.* that on average it lands heads up 50 per cent of the time) and an experiment where you toss the coin 5 times. A possible result of the experiment that we consider here is 5 heads. Let outcomes be considered unlikely with respect to an assumed distribution if their probability is lower than a significance threshold of 0.05. A potential null hypothesis implying a one-tail test is "this coin is not biased towards heads". Beware that, in this context, the word "tail" takes two meanings: either as outcome of a single toss, or as region of extremal values in a probability distribution. Indeed, with a fair coin the probability of this experiment outcome is 1/2=0.031, which would be even lower if the coin were biased in favour of tails. Therefore, the observations are not likely enough for the null hypothesis to hold, and the test refutes it. Since the coin is ostensibly neither fair nor biased towards tails, the conclusion of the experiment is that the coin is biased towards heads. Alternatively, a null hypothesis implying a two-tailed test is "this coin is fair". This one null hypothesis could be examined by looking out for either too many tails or too many heads in the experiments. The outcomes that would tend to refuse this null hypothesis are those with a large number of heads or a large number of tails, and our experiment with 5 heads would seem to belong to this class. However, the probability of 5 tosses of the same kind, irrespective of whether these are head or tails, is twice as much as that of the 5-head occurrence singly considered. Hence, under this two-tailed null hypothesis, the observation receives a probability value of 0.063. Hence again, with the same significance threshold used for the one-tailed test (0.05), the same outcome is not statistically significant. Therefore, the two-tailed null hypothesis will be preserved in this case, not supporting the conclusion reached with the single-tailed null hypothesis, that the coin is biased towards heads. This example illustrates that the conclusion reached from a statistical test may depend on the precise formulation of the null and alternative hypotheses.

Discussion

Fisher said, "the null hypothesis must be exact, that is free of vagueness and ambiguity, because it must supply the basis of the 'problem of distribution,' of which the test of significance is the solution", implying a more restrictive domain for H_0. According to this view, the null hypothesis must be numerically exact—it must state that a particular quantity or difference is equal to a particular number. In classical science, it is most typically the statement that there is *no effect* of a particular treatment; in observations, it is typically that there is *no difference* between the value of a particular measured variable and that of a prediction. Most statisticians believe that it is valid to state direction as a part of null hypothesis, or as part of a null hypothesis/alternative hypothesis pair. However, the results are not a full description of all the results of an experiment, merely a single result tailored to one particular purpose. For example, consider an H_0 that claims the population mean for a new treatment is an improvement on a well-established treatment with population mean = 10 (known from long experience), with the one-tailed alternative being that the new treatment's mean > 10. If the sample evidence

obtained through *x*-bar equals -200 and the corresponding t-test statistic equals -50, the conclusion from the test would be that there is no evidence that the new treatment is better than the existing one: it would not report that it is markedly worse, but that is not what this particular test is looking for. To overcome any possible ambiguity in reporting the result of the test of a null hypothesis, it is best to indicate whether the test was two-sided and, if one-sided, to include the direction of the effect being tested. The statistical theory required to deal with the simple cases of directionality dealt with here, and more complicated ones, makes use of the concept of an unbiased test. The directionality of hypotheses is not always obvious. The explicit null hypothesis of Fisher's Lady tasting tea example was that the Lady had no such ability, which led to a symmetric probability distribution. The one-tailed nature of the test resulted from the one-tailed alternate hypothesis (a term not used by Fisher). The null hypothesis became implicitly one-tailed. The logical negation of the Lady's one-tailed claim was also one-tailed. (Claim: Ability > 0; Stated null: Ability = 0; Implicit null: Ability ≤ 0). Pure arguments over the use of one-tailed tests are complicated by the variety of tests. Some tests (for instance the χ goodness of fit test) are inherently one-tailed. Some probability distributions are asymmetric. The traditional tests of 3 or more groups are two-tailed. Advice concerning the use of one-tailed hypotheses has been inconsistent and accepted practice varies among fields. The greatest objection to one-tailed hypotheses is their potential subjectivity. A non-significant result can sometimes be converted to a significant result by the use of a one-tailed hypothesis (as the fair coin test, at the whim of the analyst). The flip side of the argument: One-sided tests are less likely to ignore a real effect. One-tailed tests can suppress the publication of data that differs in sign from predictions. Objectivity was a goal of the developers of statistical tests. It is a common practice to use a one-tailed hypotheses by default. However, "If you do not have a specific direction firmly in mind in advance, use a two-sided alternative. Moreover, some users of statistics argue that we should *always* work with the two-sided alternative."

One alternative to this advice is to use three-outcome tests. It eliminates the issues surrounding directionality of hypotheses by testing twice, once in each direction and combining the results to produce three possible outcomes. Variations on this approach have a history, being suggested perhaps 10 times since 1950.

Disagreements over one-tailed tests flow from the philosophy of science. While Fisher was willing to ignore the unlikely case of the Lady guessing all cups of tea incorrectly (which may have been appropriate for the circumstances), medicine believes that a proposed treatment that kills patients is significant in every sense and should be reported and perhaps explained. Poor statistical reporting practices have contributed to disagreements over one-tailed tests. Statistical significance resulting from two-tailed tests is insensitive to the sign of the relationship; Reporting significance alone is inadequate. "The treatment has an effect" is the uninformative result of a two-tailed test. "The treatment has a beneficial effect" is the more informative result of a one-tailed test. "The treatment has an effect, reducing the average length of hospitalization by 1.5 days" is the most informative report, combining a two-tailed significance test result with a numeric estimate of the relationship between treatment and effect. Explicitly reporting a numeric result eliminates a philosophical advantage of a one-tailed test. An underlying issue is the appropriate form of an experimental science without numeric predictive theories: A model of numeric results is more informative than a model of effect signs (positive, negative or unknown) which is more informative than a model of simple significance (non-zero or unknown); in the absence of numeric theory signs may suffice.

HISTORY OF STATISTICAL TESTS

The history of the null and alternative hypotheses is embedded in the history of statistical tests.

- Before 1925: There are occasional transient traces of statistical tests for centuries in the past, which provide early examples of null hypotheses. In the late 19th century statistical significance was defined. In the early 20th century important probability distributions were defined. Gossett and Pearson worked on specific cases of significance testing.

- *1925:* Fisher published the first edition of Statistical Methods for Research Workers which defined the statistical significance test and made it a mainstream method of analysis for much of experimental science. The text was devoid of proofs and weak on explanations, but it was filled with real examples. It placed statistical practice in the sciences well in advance of published statistical theory.
- 1933: In a series of papers (published over a decade starting in 1928) Neyman and Pearson defined the statistical hypothesis test as a proposed improvement on Fisher's test. The papers provided much of the terminology for statistical tests including *alternative hypothesis* and H_0 as a hypothesis to be tested using observational data (with H_1, H_2... as alternatives). Neyman did not use the term null hypothesis in later writings about his method.
- 1935: Fisher published the first edition of the book "The Design of Experiments" which introduced the null hypothesis (by example rather than by definition) and carefully explained the rationale for significance tests in the context of the interpretation of experimental results.
- Following: Fisher and Neyman quarreled over the relative merits of their competing formulations until Fisher's death in 1962. Career changes and World War II ended the partnership of Neyman and Pearson. The formulations were merged by relatively anonymous textbook writers, experimenters (journal editors) and mathematical statisticians without input from the principals. The subject today combines much of the terminology and explanatory power of Neyman and Pearson with the scientific philosophy and calculations provided by Fisher. Whether statistical testing is properly one subject or two remains a source of disagreement. Sample of two: One text refers to the subject as hypothesis testing (with no mention of significance testing in the index) while another says significance testing (with a section on inference as a decision). Fisher developed significance testing as a flexible tool for researchers to weigh their evidence. Instead testing has become institutionalized. Statistical significance has become a rigidly defined and enforced criterion for the publication of experimental results in many scientific journals. In some fields significance testing has become the dominant and nearly exclusive form of statistical analysis. As a consequence the limitations of the tests have been exhaustively studied. Books have been filled with the collected criticism of significance testing.

7

Regression and Power

REGRESSION ANALYSIS

In statistical modeling, regression analysis is a set of statistical processes for estimating the relationships among variables. It includes many techniques for modeling and analyzing several variables, when the focus is on the relationship between a dependent variable and one or more independent variables (or 'predictors'). More specifically, regression analysis helps one understand how the typical value of the dependent variable (or 'criterion variable') changes when any one of the independent variables is varied, while the other independent variables are held fixed. Most commonly, regression analysis estimates the conditional expectation of the dependent variable given the independent variables – that is, the average value of the dependent variable when the independent variables are fixed. Less commonly, the focus is on a quantile, or other location parameter of the conditional distribution of the dependent variable given the independent variables. In all cases, a function of the independent variables called the regression function is to be estimated. In regression analysis, it is also of interest to characterize the variation of the dependent variable around the prediction of the regression function using a probability distribution. A related but distinct approach is Necessary Condition Analysis (NCA), which estimates the maximum (rather than average) value of the dependent variable for a given value of the independent variable (ceiling line rather than central line) in order to identify what value of the independent variable is necessary but not sufficient for a given value of the dependent variable.

Regression analysis is widely used for prediction and forecasting, where its use has substantial overlap with the field of machine learning. Regression analysis is also used to understand which among the independent variables are related to the dependent variable, and to explore the forms of these relationships. In restricted circumstances, regression analysis can be used to infer causal relationships between the independent and dependent variables. However this can lead to illusions or false relationships, so caution is advisable. Many techniques for carrying out regression analysis have been developed. Familiar methods such as linear regression and ordinary least squares regression are parametric, in that the regression function is defined in terms of a finite number of unknown parameters that are estimated from the data. Nonparametric regression refers to techniques that allow the regression function to lie in a specified set of functions, which may be infinite-dimensional.

The performance of regression analysis methods in practice depends on the form of the data generating process, and how it relates to the regression approach being used. Since the true form of the data-generating process is generally not known, regression analysis often depends to some extent on making assumptions about this process. These assumptions are sometimes testable if a sufficient quantity of data is available. Regression models for prediction are often useful even when the assumptions are moderately violated, although they may not perform optimally. However, in many applications, especially with small effects or questions of causality based on observational data, regression methods can give misleading results.

In a narrower sense, regression may refer specifically to the estimation of continuous response (dependent) variables, as opposed to the discrete response variables used in classification. The case of a continuous dependent variable may be more specifically referred to as *metric regression* to distinguish it from related problems.

HISTORY

The earliest form of regression was the method of least squares, which was published by Legendre in 1805, and by Gauss in 1809. Legendre and Gauss both applied the method to the problem of determining, from astronomical observations, the orbits of bodies about the Sun (mostly comets, but also later the then newly discovered minor planets). Gauss published a further development of the theory of least squares in 1821, including a version of the Gauss–Markov theorem. The term "regression" was coined by Francis Galton in the nineteenth century to describe a biological phenomenon. The phenomenon was that the heights of descendants of tall ancestors tend to regress down towards a normal average (a phenomenon also known as regression towards the mean). For Galton, regression had only this biological meaning, but his work was later extended by Udny Yule and Karl Pearson to a more general statistical context. In the work of Yule and Pearson, the joint distribution of the response and explanatory variables is assumed to be Gaussian. This assumption was weakened by R.A. Fisher in his works of 1922 and 1925. Fisher assumed that the conditional distribution of the response variable is Gaussian, but the joint distribution need not be. In this respect, Fisher's assumption is closer to Gauss's formulation of 1821. In the 1950s and 1960s, economists used electromechanical desk "calculators" to calculate regressions. Before 1970, it sometimes took up to 24 hours to receive the result from one regression. Regression methods continue to be an area of active research. In recent decades, new methods have been developed for robust regression, regression involving correlated responses such as time series and growth curves, regression in which the predictor (independent variable) or response variables are curves, images, graphs, or other complex data objects, regression methods accommodating various types of missing data, nonparametric regression, Bayesian methods for regression, regression in which the predictor variables are measured with error, regression with more predictor variables than observations, and causal inference with regression.

UNDERLYING ASSUMPTIONS

Classical assumptions for regression analysis include:

- The sample is representative of the population for the inference prediction.
- The error is a random variable with a mean of zero conditional on the explanatory variables.
- The independent variables are measured with no error. (Note: If this is not so, modeling may be done instead using errors-in-variables model techniques).
- The independent variables (predictors) are linearly independent, *i.e.* it is not possible to express any predictor as a linear combination of the others.
- The errors are uncorrelated, that is, the variance–covariance matrix of the errors is diagonal and each non-zero element is the variance of the error.
- The variance of the error is constant across observations (homoscedasticity). If not, weighted least squares or other methods might instead be used.

These are sufficient conditions for the least-squares estimator to possess desirable properties; in particular, these assumptions imply that the parameter estimates will be unbiased, consistent, and efficient in the class of linear unbiased estimators. It is important to note that actual data rarely satisfies the assumptions. That is, the method is used even though the assumptions are not true. Variation from the assumptions can sometimes be used as a measure of how far the model is from being useful. Many of these assumptions may be relaxed in more advanced treatments. Reports of statistical analyses usually include analyses of tests on the sample data and methodology for the fit and usefulness of the model.

Independent and dependent variables often refer to values measured at point locations. There may be spatial trends and spatial autocorrelation in the variables that violate statistical assumptions of regression. Geographic weighted regression is one technique to deal with such data. Also, variables may include values aggregated by areas. With aggregated data the modifiable areal unit problem can cause extreme variation in regression parameters. When analyzing data aggregated by political boundaries, postal codes or census areas results may be very distinct with a different choice of units.

DIAGNOSTICS

Once a regression model has been constructed, it may be important to confirm the goodness of fit of the model and the statistical significance of the estimated parameters. Commonly used checks of goodness of fit include the R-squared, analyses of the pattern of residuals and hypothesis testing. Statistical significance can be checked by an F-test of the overall fit, followed by t-tests of individual parameters. Interpretations of these diagnostic tests rest heavily on the model assumptions. Although examination of the residuals can be used to invalidate a model, the results of a t-test or F-test are sometimes more difficult to interpret if the model's assumptions are violated. For example, if the error term does not have a normal distribution, in small samples the estimated parameters will not follow normal distributions and complicate inference. With relatively large samples, however, a central limit theorem can be invoked such that hypothesis testing may proceed using asymptotic approximations.

LIMITED DEPENDENT VARIABLES

Limited dependent variables, which are response variables that are categorical variables or are variables constrained to fall only in a certain range, often arise in econometrics. The response variable may be non-continuous ("limited" to lie on some subset of the real line). For binary (zero or one) variables, if analysis proceeds with least-squares linear regression, the model is called the linear probability model. Nonlinear models for binary dependent variables include the probit and logit model. The multivariate probit model is a standard method of estimating a joint relationship between several binary dependent variables and some independent variables. For categorical variables with more than two values there is the multinomial logit. For ordinal variables with more than two values, there are the ordered logit and ordered probit models. Censored regression models may be used when the dependent variable is only sometimes observed, and Heckman correction type models may be used when the sample is not randomly selected from the population of interest. An alternative to such procedures is linear regression based on polychoric correlation (or polyserial correlations) between the categorical variables. Such procedures differ in the assumptions made about the distribution of the variables in the population. If the variable is positive with low values and represents the repetition of the occurrence of an event, then count models like the Poisson regression or the negative binomial model may be used.

NONLINEAR REGRESSION

When the model function is not linear in the parameters, the sum of squares must be minimized by an iterative procedure. This introduces many complications which are summarized in Differences between linear and non-linear least squares.

INTERPOLATION AND EXTRAPOLATION

Regression models predict a value of the Y variable given known values of the X variables. Prediction *within* the range of values in the dataset used for model-fitting is known informally as interpolation. Prediction *outside* this range of the data is known as extrapolation. Performing extrapolation relies strongly on the regression

assumptions. The further the extrapolation goes outside the data, the more room there is for the model to fail due to differences between the assumptions and the sample data or the true values. It is generally advised that when performing extrapolation, one should accompany the estimated value of the dependent variable with a prediction interval that represents the uncertainty. Such intervals tend to expand rapidly as the values of the independent variable(s) moved outside the range covered by the observed data. For such reasons and others, some tend to say that it might be unwise to undertake extrapolation.

However, this does not cover the full set of modeling errors that may be made: in particular, the assumption of a particular form for the relation between *Y* and *X*. A properly conducted regression analysis will include an assessment of how well the assumed form is matched by the observed data, but it can only do so within the range of values of the independent variables actually available. This means that any extrapolation is particularly reliant on the assumptions being made about the structural form of the regression relationship. Best-practice advice here is that a linear-in-variables and linear-in-parameters relationship should not be chosen simply for computational convenience, but that all available knowledge should be deployed in constructing a regression model. If this knowledge includes the fact that the dependent variable cannot go outside a certain range of values, this can be made use of in selecting the model – even if the observed dataset has no values particularly near such bounds. The implications of this step of choosing an appropriate functional form for the regression can be great when extrapolation is considered. At a minimum, it can ensure that any extrapolation arising from a fitted model is "realistic" (or in accord with what is known).

OTHER METHODS

Although the parameters of a regression model are usually estimated using the method of least squares, other methods which have been used include:

- Bayesian methods, *e.g.* Bayesian linear regression
- Percentage regression, for situations where reducing *percentage* errors is deemed more appropriate.
- Least absolute deviations, which is more robust in the presence of outliers, leading to quantile regression
- Nonparametric regression, requires a large number of observations and is computationally intensive
- Distance metric learning, which is learned by the search of a meaningful distance metric in a given input space.

SOFTWARE

All major statistical software packages perform least squares regression analysis and inference. Simple linear regression and multiple regression using least squares can be done in some spreadsheet applications and on some calculators. While many statistical software packages can perform various types of nonparametric and robust regression, these methods are less standardized; different software packages implement different methods, and a method with a given name may be implemented differently in different packages. Specialized regression software has been developed for use in fields such as survey analysis and neuroimaging.

ORDINARY LEAST SQUARES

In statistics, ordinary least squares (OLS) is a type of linear least squares method for estimating the unknown parameters in a linear regression model. OLS chooses the parameters of a linear function of a set of explanatory variables by the principle of least squares: minimizing the sum of the squares of the differences between the observed dependent variable (values of the variable being predicted) in the given dataset and those predicted by the linear function.

Geometrically, this is seen as the sum of the squared distances, parallel to the axis of the dependent variable, between each data point in the set and the corresponding point on the regression surface – the smaller the differences, the better the model fits the data. The resulting estimator can be expressed by a simple formula, especially in the case of a simple linear regression, in which there is a single regressor on the right side of the regression equation. The OLS estimator is consistent when the regressors are exogenous, and optimal in the class of linear unbiased estimators when the errors are homoscedastic and serially uncorrelated. Under these conditions, the method of OLS provides minimum-variance mean-unbiased estimation when the errors have finite variances. Under the additional assumption that the errors are normally distributed, OLS is the maximum likelihood estimator.

OLS is used in fields as diverse as economics (econometrics), political science, psychology and engineering (control theory and signal processing).

ASSUMPTIONS

There are several different frameworks in which the linear regression model can be cast in order to make the OLS technique applicable. Each of these settings produces the same formulas and same results. The only difference is the interpretation and the assumptions which have to be imposed in order for the method to give meaningful results. The choice of the applicable framework depends mostly on the nature of data in hand, and on the inference task which has to be performed. One of the lines of difference in interpretation is whether to treat the regressors as random variables, or as predefined constants. In the first case (random design) the regressors x_i are random and sampled together with the y_i's from some population, as in an observational study. This approach allows for more natural study of the asymptotic properties of the estimators. In the other interpretation (fixed design), the regressors X are treated as known constants set by a design, and y is sampled conditionally on the values of X as in an experiment. For practical purposes, this distinction is often unimportant, since estimation and inference is carried out while conditioning on X. All results stated in this article are within the random design framework.

HYPOTHESIS TESTING

Two hypothesis tests are particularly widely used. First, one wants to know if the estimated regression equation is any better than simply predicting that all values of the response variable equal its sample mean (if not, it is said to have no explanatory power). The null hypothesis of no explanatory value of the estimated regression is tested using an F-test. If the calculated F-value is found to be large enough to exceed its critical value for the pre-chosen level of significance, the null hypothesis is rejected and the alternative hypothesis, that the regression has explanatory power, is accepted. Otherwise, the null hypothesis of no explanatory power is accepted.

Second, for each explanatory variable of interest, one wants to know whether its estimated coefficient differs significantly from zero—that is, whether this particular explanatory variable in fact has explanatory power in predicting the response variable. Here the null hypothesis is that the true coefficient is zero. This hypothesis is tested by computing the coefficient's t-statistic, as the ratio of the coefficient estimate to its standard error. If the t-statistic is larger than a predetermined value, the null hypothesis is rejected and the variable is found to have explanatory power, with its coefficient significantly different from zero. Otherwise, the null hypothesis of a zero value of the true coefficient is accepted.

In addition, the Chow test is used to test whether two subsamples both have the same underlying true coefficient values. The sum of squared residuals of regressions on each of the subsets and on the combined data set are compared by computing an F-statistic; if this exceeds a critical value, the null hypothesis of no difference between the two subsets is rejected; otherwise, it is accepted.

PARTIAL LEAST SQUARES REGRESSION

Partial least squares regression (PLS regression) is a statistical method that bears some relation to principal components regression; instead of finding hyperplanes of maximum variance between the response and independent variables, it finds a linear regression model by projecting the predicted variables and the observable variables to a new space. Because both the *X* and *Y* data are projected to new spaces, the PLS family of methods are known as bilinear factor models. Partial least squares Discriminant Analysis (PLS-DA) is a variant used when the Y is categorical.

PLS is used to find the fundamental relations between two matrices (*X* and *Y*), *i.e.* a latent variable approach to modeling the covariance structures in these two spaces. A PLS model will try to find the multidimensional direction in the *X* space that explains the maximum multidimensional variance direction in the *Y* space. PLS regression is particularly suited when the matrix of predictors has more variables than observations, and when there is multicollinearity among *X* values. By contrast, standard regression will fail in these cases (unless it is regularized). Partial least squares was introduced by the Swedish statistician Herman O. A. Wold, who then developed it with his son, Svante Wold. An alternative term for PLS (and more correct according to Svante Wold) is *projection to latent structures*, but the term *partial least squares* is still dominant in many areas. Although the original applications were in the social sciences, PLS regression is today most widely used in chemometrics and related areas. It is also used in bioinformatics, sensometrics, neuroscience and anthropology.

Extensions

In 2002 a new method was published called orthogonal projections to latent structures (OPLS). In OPLS, continuous variable data is separated into predictive and uncorrelated information. This leads to improved diagnostics, as well as more easily interpreted visualization. However, these changes only improve the interpretability, not the predictivity, of the PLS models. L-PLS extends PLS regression to 3 connected data blocks. Similarly, OPLS-DA (Discriminant Analysis) may be applied when working with discrete variables, as in classification and biomarker studies. In 2015 partial least squares was related to a procedure called the three-pass regression filter (3PRF). Supposing the number of observations and variables are large, the 3PRF (and hence PLS) is asymptotically normal for the "best" forecast implied by a linear latent factor model. In stock market data, PLS has been shown to provide accurate out-of-sample forecasts of returns and cash-flow growth.

TOTAL LEAST SQUARES

In applied statistics, total least squares is a type of errors-in-variables regression, a least squares data modeling technique in which observational errors on both dependent and independent variables are taken into account. It is a generalization of Deming regression and also of orthogonal regression, and can be applied to both linear and non-linear models. The total least squares approximation of the data is generically equivalent to the best, in the Frobenius norm, low-rank approximation of the data matrix.

Geometrical Interpretation

When the independent variable is error-free a residual represents the "vertical" distance between the observed data point and the fitted curve (or surface). In total least squares a residual represents the distance between a data point and the fitted curve measured along some direction. In fact, if both variables are measured in the same units and the errors on both variables are the same, then the residual represents the shortest distance between the data point and the fitted curve, that is, the residual vector is perpendicular to the tangent of the curve. For this reason, this type of regression is sometimes called *two dimensional Euclidean regression* (Stein, 1983) or *orthogonal regression*.

Scale Invariant Methods

A serious difficulty arises if the variables are not measured in the same units. First consider measuring distance between a data point and the line, as in the diagram – what are the measurement units for this distance? If we consider measuring distance based on Pythagoras' Theorem then it is clear that we shall be adding quantities measured in different units, which is meaningless. Secondly, if we rescale one of the variables *e.g.*, measure in grams rather than kilograms, then we shall end up with different results (a different line). To avoid these problems it is sometimes suggested that we convert to dimensionless variables—this may be called normalization or standardization. However there are various ways of doing this, and these lead to fitted models which are not equivalent to each other. One approach is to normalize by known (or estimated) measurement precision thereby minimizing the Mahalanobis distance from the points to the line, providing a maximum-likelihood solution; the unknown precisions could be found via analysis of variance.

In short, total least squares does not have the property of units-invariance—*i.e.* it is not scale invariant. For a meaningful model we require this property to hold. A way forward is to realise that residuals (distances) measured in different units can be combined if multiplication is used instead of addition. Consider fitting a line: for each data point the product of the vertical and horizontal residuals equals twice the area of the triangle formed by the residual lines and the fitted line. We choose the line which minimizes the sum of these areas. Nobel laureate Paul Samuelson proved in 1942 that, in two dimensions, it is the only line expressible solely in terms of the ratios of standard deviations and the correlation coefficient which

- Fits the correct equation when the observations fall on a straight line,
- Exhibits scale invariance, and
- Exhibits invariance under interchange of variables. This solution has been rediscovered in different disciplines and is variously known as standardised major axis (Ricker 1975, Warton et al., 2006), the reduced major axis, the geometric mean functional relationship (Draper and Smith, 1998), least products regression, diagonal regression, line of organic correlation, and the least areas line (Tofallis, 2002). Tofallis (2015) has extended this approach to deal with multiple variables.

REGRESSION AS A STATISTICAL MODEL

Linear regression

In statistics, linear regression is a linear approach to modelling the relationship between a scalar response (or dependent variable) and one or more explanatory variables (or independent variables). The case of one explanatory variable is called simple linear regression. For more than one explanatory variable, the process is called multiple linear regression. This term is distinct from multivariate linear regression, where multiple correlated dependent variables are predicted, rather than a single scalar variable. In linear regression, the relationships are modeled using linear predictor functions whose unknown model parameters are estimated from the data. Such models are called linear models. Most commonly, the conditional mean of the response given the values of the explanatory variables (or predictors) is assumed to be an affine function of those values; less commonly, the conditional median or some other quantile is used. Like all forms of regression analysis, linear regression focuses on the conditional probability distribution of the response given the values of the predictors, rather than on the joint probability distribution of all of these variables, which is the domain of multivariate analysis. Linear regression was the first type of regression analysis to be studied rigorously, and to be used extensively in practical applications. This is because models which depend linearly on their unknown parameters are easier to fit than models which are non-linearly related to their parameters and because the statistical properties of the resulting estimators are easier to determine.

Linear regression has many practical uses. Most applications fall into one of the following two broad categories:

- If the goal is prediction, or forecasting, or error reduction, linear regression can be used to fit a predictive model to an observed data set of values of the response and explanatory variables. After developing such a model, if additional values of the explanatory variables are collected without an accompanying response value, the fitted model can be used to make a prediction of the response.
- If the goal is to explain variation in the response variable that can be attributed to variation in the explanatory variables, linear regression analysis can be applied to quantify the strength of the relationship between the response and the explanatory variables, and in particular to determine whether some explanatory variables may have no linear relationship with the response at all, or to identify which subsets of explanatory variables may contain redundant information about the response.

Linear regression models are often fitted using the least squares approach, but they may also be fitted in other ways, such as by minimizing the "lack of fit" in some other norm (as with least absolute deviations regression), or by minimizing a penalized version of the least squares cost function as in ridge regression (*L*-norm penalty) and lasso (*L*-norm penalty). Conversely, the least squares approach can be used to fit models that are not linear models. Thus, although the terms "least squares" and "linear model" are closely linked, they are not synonymous.

Assumptions

Standard linear regression models with standard estimation techniques make a number of assumptions about the predictor variables, the response variables and their relationship. Numerous extensions have been developed that allow each of these assumptions to be relaxed (*i.e.* reduced to a weaker form), and in some cases eliminated entirely. Generally these extensions make the estimation procedure more complex and time-consuming, and may also require more data in order to produce an equally precise model.

The following are the major assumptions made by standard linear regression models with standard estimation techniques (e.g. ordinary least squares):

- Weak exogeneity. This essentially means that the predictor variables x can be treated as fixed values, rather than random variables. This means, for example, that the predictor variables are assumed to be error-free—that is, not contaminated with measurement errors. Although this assumption is not realistic in many settings, dropping it leads to significantly more difficult errors-in-variables models.
- Linearity. This means that the mean of the response variable is a linear combination of the parameters (regression coefficients) and the predictor variables. Note that this assumption is much less restrictive than it may at first seem. Because the predictor variables are treated as fixed values, linearity is really only a restriction on the parameters. The predictor variables themselves can be arbitrarily transformed, and in fact multiple copies of the same underlying predictor variable can be added, each one transformed differently. This trick is used, for example, in polynomial regression, which uses linear regression to fit the response variable as an arbitrary polynomial function (up to a given rank) of a predictor variable. This makes linear regression an extremely powerful inference method. In fact, models such as polynomial regression are often "too powerful", in that they tend to overfit the data. As a result, some kind of regularization must typically be used to prevent unreasonable solutions coming out of the estimation process. Common examples are ridge regression and lasso regression. Bayesian linear regression can also be used, which by its nature is more or less immune to the problem of overfitting. (In fact, ridge regression and lasso regression can both be viewed as special cases of Bayesian linear regression, with particular types of prior distributions placed on the regression coefficients.)

- Constant variance (a.k.a. homoscedasticity). This means that different values of the response variable have the same variance in their errors, regardless of the values of the predictor variables. In practice this assumption is invalid (*i.e.* the errors are heteroscedastic) if the response variable can vary over a wide scale. In order to check for heterogeneous error variance, or when a pattern of residuals violates model assumptions of homoscedasticity (error is equally variable around the 'best-fitting line' for all points of x), it is prudent to look for a "fanning effect" between residual error and predicted values. This is to say there will be a systematic change in the absolute or squared residuals when plotted against the predictive variables. Errors will not be evenly distributed across the regression line. Heteroscedasticity will result in the averaging over of distinguishable variances around the points to get a single variance that is inaccurately representing all the variances of the line. In effect, residuals appear clustered and spread apart on their predicted plots for larger and smaller values for points along the linear regression line, and the mean squared error for the model will be wrong. Typically, for example, a response variable whose mean is large will have a greater variance than one whose mean is small. For example, a given person whose income is predicted to be $100,000 may easily have an actual income of $80,000 or $120,000 (a standard deviation of around $20,000), while another person with a predicted income of $10,000 is unlikely to have the same $20,000 standard deviation, which would imply their actual income would vary anywhere between -$10,000 and $30,000. (In fact, as this shows, in many cases—often the same cases where the assumption of normally distributed errors fails—the variance or standard deviation should be predicted to be proportional to the mean, rather than constant.) Simple linear regression estimation methods give less precise parameter estimates and misleading inferential quantities such as standard errors when substantial heteroscedasticity is present. However, various estimation techniques (*e.g.* weighted least squares and heteroscedasticity-consistent standard errors) can handle heteroscedasticity in a quite general way. Bayesian linear regression techniques can also be used when the variance is assumed to be a function of the mean. It is also possible in some cases to fix the problem by applying a transformation to the response variable (*e.g.* fit the logarithm of the response variable using a linear regression model, which implies that the response variable has a log-normal distribution rather than a normal distribution).
- Independence of errors. This assumes that the errors of the response variables are uncorrelated with each other. (Actual statistical independence is a stronger condition than mere lack of correlation and is often not needed, although it can be exploited if it is known to hold.) Some methods (*e.g.* generalized least squares) are capable of handling correlated errors, although they typically require significantly more data unless some sort of regularization is used to bias the model towards assuming uncorrelated errors. Bayesian linear regression is a general way of handling this issue.

Beyond these assumptions, several other statistical properties of the data strongly influence the performance of different estimation methods:

- The statistical relationship between the error terms and the regressors plays an important role in determining whether an estimation procedure has desirable sampling properties such as being unbiased and consistent.
- The arrangement, or probability distribution of the predictor variables x has a major influence on the precision of estimates of β. Sampling and design of experiments are highly developed subfields of statistics that provide guidance for collecting data in such a way to achieve a precise estimate of β.

Interpretation

A fitted linear regression model can be used to identify the relationship between a single predictor variable x_j and the response variable y when all the other predictor variables in the model are "held fixed". Specifically,

the interpretation of β_j is the expected change in y for a one-unit change in x_j when the other covariates are held fixed—that is, the expected value of the partial derivative of y with respect to x_j. This is sometimes called the *unique effect* of x_j on y. In contrast, the *marginal effect* of x_j on y can be assessed using a correlation coefficient or simple linear regression model relating only x_j to y; this effect is the total derivative of y with respect to x_j.

Care must be taken when interpreting regression results, as some of the regressors may not allow for marginal changes (such as dummy variables, or the intercept term), while others cannot be held fixed (recall the example from the introduction: it would be impossible to "hold t_i fixed" and at the same time change the value of t_i).

It is possible that the unique effect can be nearly zero even when the marginal effect is large. This may imply that some other covariate captures all the information in x_j, so that once that variable is in the model, there is no contribution of x_j to the variation in y. Conversely, the unique effect of x_j can be large while its marginal effect is nearly zero. This would happen if the other covariates explained a great deal of the variation of y, but they mainly explain variation in a way that is complementary to what is captured by x_j. In this case, including the other variables in the model reduces the part of the variability of y that is unrelated to x_j, thereby strengthening the apparent relationship with x_j. The meaning of the expression "held fixed" may depend on how the values of the predictor variables arise. If the experimenter directly sets the values of the predictor variables according to a study design, the comparisons of interest may literally correspond to comparisons among units whose predictor variables have been "held fixed" by the experimenter. Alternatively, the expression "held fixed" can refer to a selection that takes place in the context of data analysis. In this case, we "hold a variable fixed" by restricting our attention to the subsets of the data that happen to have a common value for the given predictor variable. This is the only interpretation of "held fixed" that can be used in an observational study. The notion of a "unique effect" is appealing when studying a complex system where multiple interrelated components influence the response variable. In some cases, it can literally be interpreted as the causal effect of an intervention that is linked to the value of a predictor variable. However, it has been argued that in many cases multiple regression analysis fails to clarify the relationships between the predictor variables and the response variable when the predictors are correlated with each other and are not assigned following a study design. A commonality analysis may be helpful in disentangling the shared and unique impacts of correlated independent variables.

Extensions

Numerous extensions of linear regression have been developed, which allow some or all of the assumptions underlying the basic model to be relaxed.

Simple and Multiple Linear Regression

The very simplest case of a single scalar predictor variable x and a single scalar response variable y is known as *simple linear regression*. The extension to multiple and/or vector-valued predictor variables (denoted with a capital X) is known as *multiple linear regression*, also known as *multivariable linear regression*. Nearly all real-world regression models involve multiple predictors, and basic descriptions of linear regression are often phrased in terms of the multiple regression model. Note, however, that in these cases the response variable y is still a scalar. Another term, *multivariate linear regression*, refers to cases where y is a vector, *i.e.*, the same as *general linear regression*.

Heteroscedastic Models

Various models have been created that allow for heteroscedasticity, *i.e.* the errors for different response variables may have different variances. For example, weighted least squares is a method for estimating linear

regression models when the response variables may have different error variances, possibly with correlated errors. Heteroscedasticity-consistent standard errors is an improved method for use with uncorrelated but potentially heteroscedastic errors.

Generalized Linear Models

Generalized linear models (GLMs) are a framework for modeling response variables that are bounded or discrete. This is used, for example:

- When modeling positive quantities (*e.g.* prices or populations) that vary over a large scale—which are better described using a skewed distribution such as the log-normal distribution or Poisson distribution (although GLMs are not used for log-normal data, instead the response variable is simply transformed using the logarithm function);
- When modeling categorical data, such as the choice of a given candidate in an election (which is better described using a Bernoulli distribution/binomial distribution for binary choices, or a categorical distribution/multinomial distribution for multi-way choices), where there are a fixed number of choices that cannot be meaningfully ordered;
- When modeling ordinal data, *e.g.* ratings on a scale from 0 to 5, where the different outcomes can be ordered but where the quantity itself may not have any absolute meaning (*e.g.* a rating of 4 may not be "twice as good" in any objective sense as a rating of 2, but simply indicates that it is better than 2 or 3 but not as good as 5).

Hierarchical Linear Models

Hierarchical linear models (or *multilevel regression*) organizes the data into a hierarchy of regressions, for example where A is regressed on B, and B is regressed on C. It is often used where the variables of interest have a natural hierarchical structure such as in educational statistics, where students are nested in classrooms, classrooms are nested in schools, and schools are nested in some administrative grouping, such as a school district. The response variable might be a measure of student achievement such as a test score, and different covariates would be collected at the classroom, school, and school district levels.

Errors-in-variables

Errors-in-variables models (or "measurement error models") extend the traditional linear regression model to allow the predictor variables X to be observed with error. This error causes standard estimators of β to become biased. Generally, the form of bias is an attenuation, meaning that the effects are biased towards zero.

Others

- In Dempster–Shafer theory, or a linear belief function in particular, a linear regression model may be represented as a partially swept matrix, which can be combined with similar matrices representing observations and other assumed normal distributions and state equations. The combination of swept or unswept matrices provides an alternative method for estimating linear regression models.

Estimation Methods

A large number of procedures have been developed for parameter estimation and inference in linear regression. These methods differ in computational simplicity of algorithms, presence of a closed-form solution, robustness

with respect to heavy-tailed distributions, and theoretical assumptions needed to validate desirable statistical properties such as consistency and asymptotic efficiency. Some of the more common estimation techniques for linear regression are summarized below.

Least-squares Estimation and Related Techniques

Linear least squares methods include mainly:

- Ordinary least squares
- Weighted least squares
- Generalized least squares

Maximum-likelihood Estimation and Related Techniques

- Maximum likelihood estimation can be performed when the distribution of the error terms is known to belong to a certain parametric family f_ϑ of probability distributions. When f_ϑ is a normal distribution with zero mean and variance θ, the resulting estimate is identical to the OLS estimate. GLS estimates are maximum likelihood estimates when ε follows a multivariate normal distribution with a known covariance matrix.
- Ridge regression and other forms of penalized estimation, such as Lasso regression, deliberately introduce bias into the estimation of β in order to reduce the variability of the estimate. The resulting estimates generally have lower mean squared error than the OLS estimates, particularly when multicollinearity is present or when overfitting is a problem. They are generally used when the goal is to predict the value of the response variable y for values of the predictors x that have not yet been observed. These methods are not as commonly used when the goal is inference, since it is difficult to account for the bias.
- Least absolute deviation (LAD) regression is a robust estimation technique in that it is less sensitive to the presence of outliers than OLS (but is less efficient than OLS when no outliers are present). It is equivalent to maximum likelihood estimation under a Laplace distribution model for ε.

Other Estimation Techniques

- Bayesian linear regression applies the framework of Bayesian statistics to linear regression. In particular, the regression coefficients β are assumed to be random variables with a specified prior distribution. The prior distribution can bias the solutions for the regression coefficients, in a way similar to (but more general than) ridge regression or lasso regression. In addition, the Bayesian estimation process produces not a single point estimate for the "best" values of the regression coefficients but an entire posterior distribution, completely describing the uncertainty surrounding the quantity. This can be used to estimate the "best" coefficients using the mean, mode, median, any quantile or any other function of the posterior distribution.
- Quantile regression focuses on the conditional quantiles of y given X rather than the conditional mean of y given X. Linear quantile regression models a particular conditional quantile, for example the conditional median, as a linear function βx of the predictors.
- Mixed models are widely used to analyze linear regression relationships involving dependent data when the dependencies have a known structure. Common applications of mixed models include analysis of data involving repeated measurements, such as longitudinal data, or data obtained from cluster sampling. They are generally fit as parametric models, using maximum likelihood or Bayesian

estimation. In the case where the errors are modeled as normal random variables, there is a close connection between mixed models and generalized least squares. Fixed effects estimation is an alternative approach to analyzing this type of data.

- Principal component regression (PCR) is used when the number of predictor variables is large, or when strong correlations exist among the predictor variables. This two-stage procedure first reduces the predictor variables using principal component analysis then uses the reduced variables in an OLS regression fit. While it often works well in practice, there is no general theoretical reason that the most informative linear function of the predictor variables should lie among the dominant principal components of the multivariate distribution of the predictor variables. The partial least squares regression is the extension of the PCR method which does not suffer from the mentioned deficiency.
- Least-angle regression is an estimation procedure for linear regression models that was developed to handle high-dimensional covariate vectors, potentially with more covariates than observations.
- The Theil–Sen estimator is a simple robust estimation technique that chooses the slope of the fit line to be the median of the slopes of the lines through pairs of sample points. It has similar statistical efficiency properties to simple linear regression but is much less sensitive to outliers.
- Other robust estimation techniques, including the α-trimmed mean approach, and L-, M-, S-, and R-estimators have been introduced.

Applications

Linear regression is widely used in biological, behavioral and social sciences to describe possible relationships between variables. It ranks as one of the most important tools used in these disciplines.

Trend Line

A trend line represents a trend, the long-term movement in time series data after other components have been accounted for. It tells whether a particular data set (say GDP, oil prices or stock prices) have increased or decreased over the period of time. A trend line could simply be drawn by eye through a set of data points, but more properly their position and slope is calculated using statistical techniques like linear regression. Trend lines typically are straight lines, although some variations use higher degree polynomials depending on the degree of curvature desired in the line.

Trend lines are sometimes used in business analytics to show changes in data over time. This has the advantage of being simple. Trend lines are often used to argue that a particular action or event (such as training, or an advertising campaign) caused observed changes at a point in time. This is a simple technique, and does not require a control group, experimental design, or a sophisticated analysis technique. However, it suffers from a lack of scientific validity in cases where other potential changes can affect the data.

Epidemiology

Early evidence relating tobacco smoking to mortality and morbidity came from observational studies employing regression analysis. In order to reduce spurious correlations when analyzing observational data, researchers usually include several variables in their regression models in addition to the variable of primary interest. For example, in a regression model in which cigarette smoking is the independent variable of primary interest and the dependent variable is lifespan measured in years, researchers might include education and income as additional independent variables, to ensure that any observed effect of smoking on lifespan is not due to those other socio-economic factors. However, it is never possible to include all possible confounding variables in an empirical analysis. For example, a hypothetical gene might increase mortality and also cause people to

smoke more. For this reason, randomized controlled trials are often able to generate more compelling evidence of causal relationships than can be obtained using regression analyses of observational data. When controlled experiments are not feasible, variants of regression analysis such as instrumental variables regression may be used to attempt to estimate causal relationships from observational data.

Finance

The capital asset pricing model uses linear regression as well as the concept of beta for analyzing and quantifying the systematic risk of an investment. This comes directly from the beta coefficient of the linear regression model that relates the return on the investment to the return on all risky assets.

Economics

Linear regression is the predominant empirical tool in economics. For example, it is used to predict consumption spending, fixed investment spending, inventory investment, purchases of a country's exports, spending on imports, the demand to hold liquid assets, labour demand, and labour supply.

Environmental Science

Linear regression finds application in a wide range of environmental science applications. In Canada, the Environmental Effects Monitoring Programme uses statistical analyses on fish and benthic surveys to measure the effects of pulp mill or metal mine effluent on the aquatic ecosystem.

Machine Learning

Linear regression plays an important role in the field of artificial intelligence such as machine learning. The linear regression algorithm is one of the fundamental supervised machine-learning algorithms due to its relative simplicity and well-known properties.

Simple Linear Regression

In statistics, simple linear regression is a linear regression model with a single explanatory variable. That is, it concerns two-dimensional sample points with one independent variable and one dependent variable (conventionally, the x and y coordinates in a Cartesian coordinate system) and finds a linear function (a non-vertical straight line) that, as accurately as possible, predicts the dependent variable values as a function of the independent variables. The adjective *simple* refers to the fact that the outcome variable is related to a single predictor. It is common to make the additional stipulation that the ordinary least squares method should be used: the accuracy of each predicted value is measured by its squared *residual* (vertical distance between the point of the data set and the fitted line), and the goal is to make the sum of these squared deviations as small as possible. Other regression methods that can be used in place of ordinary least squares include least absolute deviations (minimizing the sum of absolute values of residuals) and the Theil–Sen estimator (which chooses a line whose slope is the median of the slopes determined by pairs of sample points). Deming regression (total least squares) also finds a line that fits a set of two-dimensional sample points, but (unlike ordinary least squares, least absolute deviations, and median slope regression) it is not really an instance of simple linear regression, because it does not separate the coordinates into one dependent and one independent variable and could potentially return a vertical line as its fit. The remainder of the article assumes an ordinary least squares regression. In this case, the slope of the fitted line is equal to the correlation between y and x corrected by the ratio of standard deviations of these variables. The intercept of the fitted line is such that the line passes through the center of mass (x, y) of the data points.

Model-based Properties

Description of the statistical properties of estimators from the simple linear regression estimates requires the use of a statistical model. The following is based on assuming the validity of a model under which the estimates are optimal. It is also possible to evaluate the properties under other assumptions, such as inhomogeneity, but this is discussed elsewhere.

Confidence Intervals

The formulas given in the previous section allow one to calculate the *point estimates* of α and β — that is, the coefficients of the regression line for the given set of data. However, those formulas don't tell us how precise the estimates are, *i.e.*, how much the estimators α and β vary from sample to sample for the specified sample size. Confidence intervals were devised to give a plausible set of values to the estimates one might have if one repeated the experiment a very large number of times.

The standard method of constructing confidence intervals for linear regression coefficients relies on the normality assumption, which is justified if either:

- The errors in the regression are normally distributed (the so-called *classic regression* assumption), or
- The number of observations n is sufficiently large, in which case the estimator is approximately normally distributed.

The latter case is justified by the central limit theorem.

Generalized Least Squares

In statistics, generalized least squares (GLS) is a technique for estimating the unknown parameters in a linear regression model when there is a certain degree of correlation between the residuals in a regression model. In these cases, ordinary least squares and weighted least squares can be statistically inefficient, or even give misleading inferences. GLS was first described by Alexander Aitken in 1934.

Weighted Least Squares

A special case of GLS called weighted least squares (WLS) occurs when all the off-diagonal entries of Ω are 0. This situation arises when the variances of the observed values are unequal (*i.e.* heteroscedasticity is present), but where no correlations exist among the observed variances. The weight for unit i is proportional to the reciprocal of the variance of the response for unit i.

General Linear Model

The general linear model or multivariate regression model is a statistical linear model. It may be written as

$$Y = X B + U,$$

where Y is a matrix with series of multivariate measurements (each column being a set of measurements on one of the dependent variables), X is a matrix of observations on independent variables that might be a design matrix (each column being a set of observations on one of the independent variables), B is a matrix containing parameters that are usually to be estimated and U is a matrix containing errors (noise). The errors are usually assumed to be uncorrelated across measurements, and follow a multivariate normal distribution. If the errors do not follow a multivariate normal distribution, generalized linear models may be used to relax assumptions about Y and U. The general linear model incorporates a number of different statistical models: ANOVA, ANCOVA, MANOVA, MANCOVA, ordinary linear regression, t-test and F-test. The general linear model is a

generalization of multiple linear regression model to the case of more than one dependent variable. If Y, B, and U were column vectors, the matrix equation above would represent multiple linear regression. Hypothesis tests with the general linear model can be made in two ways: multivariate or as several independent univariate tests. In multivariate tests the columns of Y are tested together, whereas in univariate tests the columns of Y are tested independently, *i.e.*, as multiple univariate tests with the same design matrix.

Applications

An application of the general linear model appears in the analysis of multiple brain scans in scientific experiments where Y contains data from brain scanners, X contains experimental design variables and confounds. It is usually tested in a univariate way (usually referred to a *mass-univariate* in this setting) and is often referred to as statistical parametric mapping.

PREDICTOR STRUCTURE

Polynomial Regression

In statistics, polynomial regression is a form of regression analysis in which the relationship between the independent variable x and the dependent variable y is modelled as an nth degree polynomial in x. Polynomial regression fits a nonlinear relationship between the value of x and the corresponding conditional mean of y, denoted $E(y \mid x)$, and has been used to describe nonlinear phenomena such as the growth rate of tissues, the distribution of carbon isotopes in lake sediments, and the progression of disease epidemics. Although *polynomial regression* fits a nonlinear model to the data, as a statistical estimation problem it is linear, in the sense that the regression function $E(y \mid x)$ is linear in the unknown parameters that are estimated from the data. For this reason, polynomial regression is considered to be a special case of multiple linear regression. The explanatory (independent) variables resulting from the polynomial expansion of the "baseline" variables are known as higher-degree terms. Such variables are also used in classification settings.

History

Polynomial regression models are usually fit using the method of least squares. The least-squares method minimizes the variance of the unbiased estimators of the coefficients, under the conditions of the Gauss–Markov theorem. The least-squares method was published in 1805 by Legendre and in 1809 by Gauss. The first design of an experiment for polynomial regression appeared in an 1815 paper of Gergonne. In the twentieth century, polynomial regression played an important role in the development of regression analysis, with a greater emphasis on issues of design and inference. More recently, the use of polynomial models has been complemented by other methods, with non-polynomial models having advantages for some classes of problems.

Interpretation

Although polynomial regression is technically a special case of multiple linear regression, the interpretation of a fitted polynomial regression model requires a somewhat different perspective. It is often difficult to interpret the individual coefficients in a polynomial regression fit, since the underlying monomials can be highly correlated. For example, x and x have correlation around 0.97 when x is uniformly distributed on the interval (0, 1). Although the correlation can be reduced by using orthogonal polynomials, it is generally more informative to consider the fitted regression function as a whole. Point-wise or simultaneous confidence bands can then be used to provide a sense of the uncertainty in the estimate of the regression function.

Segmented Regression

Segmented regression, also known as piecewise regression or "broken-stick regression", is a method in regression analysis in which the independent variable is partitioned into intervals and a separate line segment is fit to each interval. Segmented regression analysis can also be performed on multivariate data by partitioning the various independent variables. Segmented regression is useful when the independent variables, clustered into different groups, exhibit different relationships between the variables in these regions. The boundaries between the segments are *breakpoints*. Segmented linear regression is segmented regression whereby the relations in the intervals are obtained by linear regression.

No-effect Range

Segmented regression is often used to detect over which range an explanatory variable (X) has no effect on the dependent variable (Y), while beyond the reach there is a clear response, be it positive or negative. The reach of no effect may be found at the initial part of X domain or conversely at its last part. For the "no effect" analysis, application of the least squares method for the segmented regression analysis may not be the most appropriate technique because the aim is rather to find the longest stretch over which the Y-X relation can be considered to possess zero slope while beyond the reach the slope is significantly different from zero but knowledge about the best value of this slope is not material. The method to find the no-effect range is progressive partial regression over the range, extending the range with small steps until the regression coefficient gets significantly different from zero. In the next figure the break point is found at X=7.9 while for the same data, the least squares method yields a break point only at X=4.9. The latter value is lower, but the fit of the data beyond the break point is better. Hence, it will depend on the purpose of the analysis which method needs to be employed.

Local Regression

Local regression or local polynomial regression, also known as moving regression, is a generalization of moving average and polynomial regression. Its most common methods, initially developed for scatterplot smoothing, are LOESS (locally estimated scatterplot smoothing) and LOWESS (locally weighted scatterplot smoothing), both pronounced/loʊɛs/. They are two strongly related non-parametric regression methods that combine multiple regression models in a *k*-nearest-neighbour-based meta-model.

LOESS and LOWESS thus build on "classical" methods, such as linear and nonlinear least squares regression. They address situations in which the classical procedures do not perform well or cannot be effectively applied without undue labour. LOESS combines much of the simplicity of linear least squares regression with the flexibility of nonlinear regression. It does this by fitting simple models to localized subsets of the data to build up a function that describes the deterministic part of the variation in the data, point by point. In fact, one of the chief attractions of this method is that the data analyst is not required to specify a global function of any form to fit a model to the data, only to fit segments of the data. The trade-off for these features is increased computation. Because it is so computationally intensive, LOESS would have been practically impossible to use in the era when least squares regression was being developed. Most other modern methods for process modeling are similar to LOESS in this respect. These methods have been consciously designed to use our current computational ability to the fullest possible advantage to achieve goals not easily achieved by traditional approaches. A smooth curve through a set of data points obtained with this statistical technique is called a Loess Curve, particularly when each smoothed value is given by a weighted quadratic least squares regression over the span of values of the y-axis scattergram criterion variable. When each smoothed value is given by a weighted linear least squares regression over the span, this is known as a Lowess curve; however, some authorities treat Lowess and Loess as synonyms.

Model Definition

LOESS, originally proposed by Cleveland (1979) and further developed by Cleveland and Devlin (1988), specifically denotes a method that is also known as locally weighted polynomial regression. At each point in the range of the data set a low-degree polynomial is fitted to a subset of the data, with explanatory variable values near the point whose response is being estimated. The polynomial is fitted using weighted least squares, giving more weight to points near the point whose response is being estimated and less weight to points further away. The value of the regression function for the point is then obtained by evaluating the local polynomial using the explanatory variable values for that data point. The LOESS fit is complete after regression function values have been computed for each of the n data points. Many of the details of this method, such as the degree of the polynomial model and the weights, are flexible. The range of choices for each part of the method and typical defaults are briefly discussed next.

Degree of Local Polynomials

The local polynomials fit to each subset of the data are almost always of first or second degree; that is, either locally linear (in the straight line sense) or locally quadratic. Using a zero degree polynomial turns LOESS into a weighted moving average. Higher-degree polynomials would work in theory, but yield models that are not really in the spirit of LOESS. LOESS is based on the ideas that any function can be well approximated in a small neighborhood by a low-order polynomial and that simple models can be fit to data easily. High-degree polynomials would tend to overfit the data in each subset and are numerically unstable, making accurate computations difficult.

Advantages

As discussed above, the biggest advantage LOESS has over many other methods is the fact that it does not require the specification of a function to fit a model to all of the data in the sample. Instead the analyst only has to provide a smoothing parameter value and the degree of the local polynomial. In addition, LOESS is very flexible, making it ideal for modeling complex processes for which no theoretical models exist. These two advantages, combined with the simplicity of the method, make LOESS one of the most attractive of the modern regression methods for applications that fit the general framework of least squares regression but which have a complex deterministic structure. Although it is less obvious than for some of the other methods related to linear least squares regression, LOESS also accrues most of the benefits typically shared by those procedures. The most important of those is the theory for computing uncertainties for prediction and calibration. Many other tests and procedures used for validation of least squares models can also be extended to LOESS models.

Disadvantages

LOESS makes less efficient use of data than other least squares methods. It requires fairly large, densely sampled data sets in order to produce good models. This is because LOESS relies on the local data structure when performing the local fitting. Thus, LOESS provides less complex data analysis in exchange for greater experimental costs. Another disadvantage of LOESS is the fact that it does not produce a regression function that is easily represented by a mathematical formula. This can make it difficult to transfer the results of an analysis to other people. In order to transfer the regression function to another person, they would need the data set and software for LOESS calculations. In nonlinear regression, on the other hand, it is only necessary to write down a functional form in order to provide estimates of the unknown parameters and the estimated uncertainty. Depending on the application, this could be either a major or a minor drawback to using LOESS. In particular, the simple form of LOESS can not be used for mechanistic modelling where fitted parameters specify

particular physical properties of a system. Finally, as discussed above, LOESS is a computationally intensive method (with the exception of evenly spaced data, where the regression can then be phrased as a non-causal finite impulse response filter). LOESS is also prone to the effects of outliers in the data set, like other least squares methods. There is an iterative, robust version of LOESS [Cleveland (1979)] that can be used to reduce LOESS' sensitivity to outliers, but too many extreme outliers can still overcome even the robust method.

NON-STANDARD

Nonlinear Regression

In statistics, nonlinear regression is a form of regression analysis in which observational data are modeled by a function which is a nonlinear combination of the model parameters and depends on one or more independent variables. The data are fitted by a method of successive approximations.

Ordinary and Weighted Least Squares

The best-fit curve is often assumed to be that which minimizes the sum of squared residuals. This is the ordinary least squares (OLS) approach. However, in cases where the dependent variable does not have constant variance, a sum of weighted squared residuals may be minimized. Each weight should ideally be equal to the reciprocal of the variance of the observation, but weights may be recomputed on each iteration, in an iteratively weighted least squares algorithm.

Nonparametric Regression

Nonparametric regression is a category of regression analysis in which the predictor does not take a predetermined form but is constructed according to information derived from the data. Nonparametric regression requires larger sample sizes than regression based on parametric models because the data must supply the model structure as well as the model estimates.

Gaussian Process Regression or Kriging

In Gaussian process regression, also known as Kriging, a Gaussian prior is assumed for the regression curve. The errors are assumed to have a multivariate normal distribution and the regression curve is estimated by its posterior mode. The Gaussian prior may depend on unknown hyperparameters, which are usually estimated via empirical Bayes. Smoothing splines have an interpretation as the posterior mode of a Gaussian process regression.

Kernel Regression

Kernel regression estimates the continuous dependent variable from a limited set of data points by convolving the data points' locations with a kernel function—approximately speaking, the kernel function specifies how to "blur" the influence of the data points so that their values can be used to predict the value for nearby locations.

Nonparametric Multiplicative Regression

Nonparametric multiplicative regression (NPMR) is a form of nonparametric regression based on multiplicative kernel estimation. Like other regression methods, the goal is to estimate a response (dependent variable) based on one or more predictors (independent variables). NPMR can be a good choice for a regression method if the following are true:

- The shape of the response surface is unknown.
- The predictors are likely to interact in producing the response; in other words, the shape of the response to one predictor is likely to depend on other predictors.
- The response is either a quantitative or binary (0/1) variable.

This is a smoothing technique that can be cross-validated and applied in a predictive way.

NPMR Behaves Like an Organism

NPMR has been useful for modeling the response of an organism to its environment. Organismal response to environment tends to be nonlinear and have complex interactions among predictors. NPMR allows you to model automatically the complex interactions among predictors in much the same way that organisms integrate the numerous factors affecting their performance. A key biological feature of an NPMR model is that failure of an organism to tolerate any single dimension of the predictor space results in overall failure of the organism. For example, assume that a plant needs a certain range of moisture in a particular temperature range. If either temperature or moisture fall outside the tolerance of the organism, then the organism dies. If it is too hot, then no amount of moisture can compensate to result in survival of the plant. Mathematically this works with NPMR because the product of the weights for the target point is zero or near zero if any of the weights for individual predictors (moisture or temperature) are zero or near zero. Note further that in this simple example, the second condition listed above is probably true: the response of the plant to moisture probably depends on temperature and vice versa. Optimizing the selection of predictors and their smoothing parameters in a multiplicative model is computationally intensive. With a large pool of predictors, the computer must search through a huge number of potential models in search for the best model. The best model has the best fit, subject to overfitting constraints or penalties.

The Local Model

NPMR can be applied with several different kinds of local models. By "local model" we mean the way that data points near a target point in the predictor space are combined to produce an estimate for the target point. The most common choices for the local models are the local mean estimator, a local linear estimator, or a local logistic estimator. In each case the weights can be extended multiplicatively to multiple dimensions. In words, the estimate of the response is a local estimate (for example a local mean) of the observed values, each value weighted by its proximity to the target point in the predictor space, the weights being the product of weights for individual predictors. The model allows interactions, because weights for individual predictors are combined by multiplication rather than addition.

Overfitting Controls

Understanding and using these controls on overfitting is essential to effective modeling with nonparametric regression. Nonparametric regression models can become overfit either by including too many predictors or by using small smoothing parameters (also known as bandwidth or tolerance). This can make a big difference with special problems, such as small data sets or clumped distributions along predictor variables. The methods for controlling overfitting differ between NPMR and the generalized linear modeling (GLMs). The most popular overfitting controls for GLMs are the Akaike information criterion (AIC) and the Bayesian information criterion (BIC) for model selection. The AIC and BIC depend on the number of parameters in a model. Because NPMR models do not have explicit parameters as such, these are not directly applicable to NPMR models. Instead, one can control overfitting by setting a minimum average neighborhood size, minimum data:predictor ratio, and a minimum improvement required to add a predictor to a model. Nonparametric regression models sometimes use an AIC based on the "effective number of parameters". This penalizes a measure of fit by the trace of the smoothing matrix—essentially how much each data point contributes to estimating itself, summed across all

data points. If, however, you use leave-one-out cross validation in the model fitting phase, the trace of the smoothing matrix is always zero, corresponding to zero parameters for the AIC. Thus, NPMR with cross-validation in the model fitting phase already penalizes the measure of fit, such that the error rate of the training data set is expected to approximate the error rate in a validation data set. In other words, the training error rate approximates the prediction (extra-sample) error rate.

Related Techniques

NPMR is essentially a smoothing technique that can be cross-validated and applied in a predictive way. Many other smoothing techniques are well known, for example smoothing splines and wavelets. The optimal choice of a smoothing method depends on the specific application. Nonparametric regression models always fits for larger data.

Regression Trees

Decision tree learning algorithms can be applied to learn to predict a dependent variable from data. Although the original Classification And Regression Tree (CART) formulation applied only to predicting univariate data, the framework can be used to predict multivariate data, including time series.

Semiparametric Regression

In statistics, semiparametric regression includes regression models that combine parametric and nonparametric models. They are often used in situations where the fully nonparametric model may not perform well or when the researcher wants to use a parametric model but the functional form with respect to a subset of the regressors or the density of the errors is not known. Semiparametric regression models are a particular type of semiparametric modelling and, since semiparametric models contain a parametric component, they rely on parametric assumptions and may be misspecified and inconsistent, just like a fully parametric model.

Robust Regression

In robust statistics, robust regression is a form of regression analysis designed to overcome some limitations of traditional parametric and non-parametric methods. Regression analysis seeks to find the relationship between one or more independent variables and a dependent variable. Certain widely used methods of regression, such as ordinary least squares, have favourable properties if their underlying assumptions are true, but can give misleading results if those assumptions are not true; thus ordinary least squares is said to be not robust to violations of its assumptions. Robust regression methods are designed to be not overly affected by violations of assumptions by the underlying data-generating process. In particular, least squares estimates for regression models are highly sensitive to (*i.e.* not robust against) outliers. While there is no precise definition of an outlier, outliers are observations which do not follow the pattern of the other observations. This is not normally a problem if the outlier is simply an extreme observation drawn from the tail of a normal distribution, but if the outlier results from non-normal measurement error or some other violation of standard ordinary least squares assumptions, then it compromises the validity of the regression results if a non-robust regression technique is used.

Applications

Heteroscedastic Errors

One instance in which robust estimation should be considered is when there is a strong suspicion of heteroscedasticity. In the homoscedastic model, it is assumed that the variance of the error term is constant for

all values of x. Heteroscedasticity allows the variance to be dependent on x, which is more accurate for many real scenarios. For example, the variance of expenditure is often larger for individuals with higher income than for individuals with lower incomes. Software packages usually default to a homoscedastic model, even though such a model may be less accurate than a heteroscedastic model. One simple approach (Tofallis, 2008) is to apply least squares to percentage errors, as this reduces the influence of the larger values of the dependent variable compared to ordinary least squares.

Presence of Outliers

Another common situation in which robust estimation is used occurs when the data contain outliers. In the presence of outliers that do not come from the same data-generating process as the rest of the data, least squares estimation is inefficient and can be biased. Because the least squares predictions are dragged towards the outliers, and because the variance of the estimates is artificially inflated, the result is that outliers can be masked. (In many situations, including some areas of geostatistics and medical statistics, it is precisely the outliers that are of interest.)

Although it is sometimes claimed that least squares (or classical statistical methods in general) are robust, they are only robust in the sense that the type I error rate does not increase under violations of the model. In fact, the type I error rate tends to be lower than the nominal level when outliers are present, and there is often a dramatic increase in the type II error rate. The reduction of the type I error rate has been labelled as the *conservatism* of classical methods.

History and Unpopularity of Robust Regression

Despite their superior performance over least squares estimation in many situations, robust methods for regression are still not widely used. Several reasons may help explain their unpopularity (Hampel et al. 1986, 2005). One possible reason is that there are several competing methods and the field got off to many false starts. Also, computation of robust estimates is much more computationally intensive than least squares estimation; in recent years, however, this objection has become less relevant, as computing power has increased greatly. Another reason may be that some popular statistical software packages failed to implement the methods (Stromberg, 2004). The belief of many statisticians that classical methods are robust may be another reason.

Although uptake of robust methods has been slow, modern mainstream statistics text books often include discussion of these methods. Also, modern statistical software packages such as R, Statsmodels, Stata and S-PLUS include considerable functionality for robust estimation.

Methods for Robust Regression

Least Squares Alternatives

The simplest methods of estimating parameters in a regression model that are less sensitive to outliers than the least squares estimates, is to use least absolute deviations. Even then, gross outliers can still have a considerable impact on the model, motivating research into even more robust approaches.

In 1964, Huber introduced M-estimation for regression. The M in M-estimation stands for "maximum likelihood type". The method is robust to outliers in the response variable, but turned out not to be resistant to outliers in the explanatory variables (leverage points). In fact, when there are outliers in the explanatory variables, the method has no advantage over least squares. In the 1980s, several alternatives to M-estimation were proposed as attempts to overcome the lack of resistance. Least trimmed squares (LTS) is a viable alternative and is currently (2007) the preferred choice of Rousseeuw and Ryan (1997, 2008). The Theil–Sen estimator has a lower breakdown point than LTS but is statistically efficient and popular. Another proposed solution was S-

estimation. This method finds a line (plane or hyperplane) that minimizes a robust estimate of the scale (from which the method gets the S in its name) of the residuals. This method is highly resistant to leverage points and is robust to outliers in the response. However, this method was also found to be inefficient.

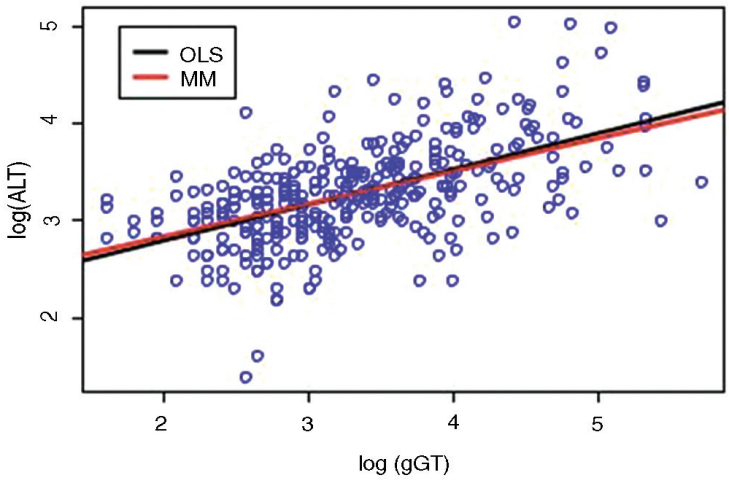
MM-estimation attempts to retain the robustness and resistance of S-estimation, whilst gaining the efficiency of M-estimation. The method proceeds by finding a highly robust and resistant S-estimate that minimizes an M-estimate of the scale of the residuals (the first M in the method's name). The estimated scale is then held constant whilst a close-by M-estimate of the parameters is located (the second M).

Unit Weights

Another robust method is the use of unit weights (Wainer and Thissen, 1976), a method that can be applied when there are multiple predictors of a single outcome. Ernest Burgess (1928) used unit weights to predict success on parole. He scored 21 positive factors as present (*e.g.*, "no prior arrest" = 1) or absent ("prior arrest" = 0), then summed to yield a predictor score, which was shown to be a useful predictor of parole success. Samuel S. Wilks (1938) showed that nearly all sets of regression weights sum to composites that are very highly correlated with one another, including unit weights, a result referred to as Wilk's theorem (Ree, Carretta, and Earles, 1998). Robyn Dawes (1979) examined decision making in applied settings, showing that simple models with unit weights often outperformed human experts. Bobko, Roth, and Buster (2007) reviewed the literature on unit weights and concluded that decades of empirical studies show that unit weights perform similar to ordinary regression weights on cross validation.

Example: BUPA Liver Data

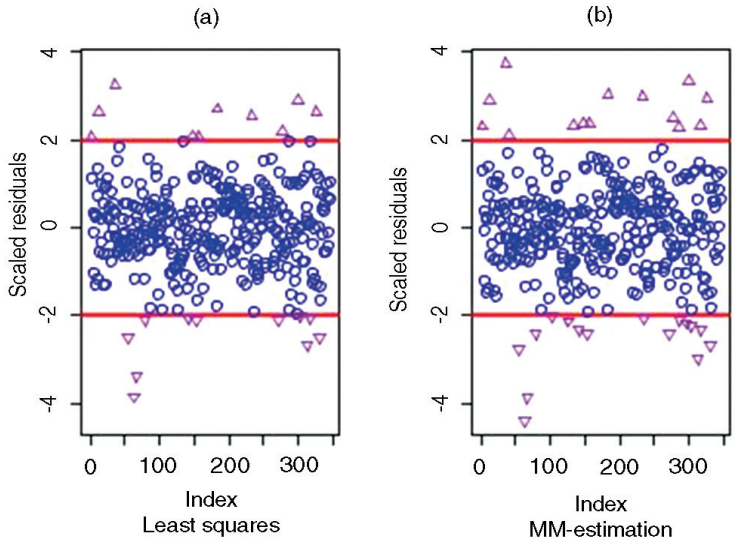
The BUPA liver data have been studied by various authors, including Breiman (2001). The data can be found at the classic data sets page, and there is some discussion in the article on the Box-Cox transformation. A plot of the logs of ALT versus the logs of γGT appears below. The two regression lines are those estimated by ordinary least squares (OLS) and by robust MM-estimation. The analysis was performed in R using software made available by Venables and Ripley (2002).



The two regression lines appear to be very similar (and this is not unusual in a data set of this size). However, the advantage of the robust approach comes to light when the estimates of residual scale are considered. For ordinary least squares, the estimate of scale is 0.420, compared to 0.373 for the robust method. Thus, the relative efficiency of ordinary least squares to MM-estimation in this example is 1.266. This inefficiency leads to loss of power in hypothesis tests and to unnecessarily wide confidence intervals on estimated parameters.

Outlier Detection

Another consequence of the inefficiency of the ordinary least squares fit is that several outliers are masked because the estimate of residual scale is inflated, the scaled residuals are pushed closer to zero than when a more appropriate estimate of scale is used. The plots of the scaled residuals from the two models appear below. The variable on the *x* axis is just the observation number as it appeared in the data set. Rousseeuw and Leroy (1986) contains many such plots.



The horizontal reference lines are at 2 and -2, so that any observed scaled residual beyond these boundaries can be considered to be an outlier. Clearly, the least squares method leads to many interesting observations being masked.

Whilst in one or two dimensions outlier detection using classical methods can be performed manually, with large data sets and in high dimensions the problem of masking can make identification of many outliers impossible. Robust methods automatically detect these observations, offering a serious advantage over classical methods when outliers are present.

Quantile Regression

Quantile regression is a type of regression analysis used in statistics and econometrics. Whereas the method of least squares results in estimates of the conditional *mean* of the response variable given certain values of the predictor variables, quantile regression aims at estimating either the conditional median or other quantiles of the response variable. Essentially, quantile regression is the extension of linear regression and we use it when the conditions of linear regression are not applicable.

Advantages and Applications

Quantile regression is desired if conditional quantile functions are of interest. One advantage of quantile regression, relative to the ordinary least squares regression, is that the quantile regression estimates are more robust against outliers in the response measurements. However, the main attraction of quantile regression goes beyond that. Different measures of central tendency and statistical dispersion can be useful to obtain a more comprehensive analysis of the relationship between variables.

In ecology, quantile regression has been proposed and used as a way to discover more useful predictive relationships between variables in cases where there is no relationship or only a weak relationship between the means of such variables. The need for and success of quantile regression in ecology has been attributed to the

complexity of interactions between different factors leading to data with unequal variation of one variable for different ranges of another variable. Another application of quantile regression is in the areas of growth charts, where percentile curves are commonly used to screen for abnormal growth.

Mathematics

The mathematical forms arising from quantile regression are distinct from those arising in the method of least squares. The method of least squares leads to a consideration of problems in an inner product space, involving projection onto subspaces, and thus the problem of minimizing the squared errors can be reduced to a problem in numerical linear algebra. Quantile regression does not have this structure, and instead leads to problems in linear programming that can be solved by the simplex method.

History

The idea of estimating a median regression slope, a major theorem about minimizing sum of the absolute deviances and a geometrical algorithm for constructing median regression was proposed in 1760 by Ruder Josip Boskoviæ, a Jesuit Catholic priest from Dubrovnik. He was interested in the ellipticity of the earth, building on Isaac Newton's suggestion that its rotation could cause it to bulge at the equator with a corresponding flattening at the poles. He finally produced the first geometric procedure for determining the equator of a rotating planet from three observations of a surface feature. More importantly for quantile regression, he was able to develop the first evidence of the least absolute criterion and preceded the least squares introduced by Legendre in 1805 by fifty years. Other thinkers began building upon Boskoviæ's idea such as Pierre-Simon Laplace, who developed the so-called "methode de situation." This led to Francis Edgeworth's plural median - a geometric approach to median regression - and is recognized as the precursor of the simplex method. The works of Boskoviæ, Laplace, and Edgeworth were recognized as a prelude to Roger Koenker's contributions to quantile regression.

Median regression computations for larger data sets are quite tedious compared to the least squares method, for which reason it has historically generated a lack of popularity among statisticians, until the widespread adoption of computers in the latter part of the 20th century.

Bayesian Methods for Quantile Regression

Because quantile regression does not normally assume a parametric likelihood for the conditional distributions of Y|X, the Bayesian methods work with a working likelihood. A convenient choice is the asymmetric Laplacian likelihood, because the mode of the resulting posterior under a flat prior is the usual quantile regression estimates. The posterior inference, however, must be interpreted with care. Yang, Wang and He provided a posterior variance adjustment for valid inference. In addition, Yang and He showed that one can have asymptotically valid posterior inference if the working likelihood is chosen to be the empirical likelihood.

Implementations

Numerous statistical software packages include implementations of quantile regression:

- Matlab function quantreg
- Eviews, since version 6.
- gretl has the quantreg command.
- R offers several packages that implement quantile regression, most notably quantreg by Roger Koenker, but also gbm, quantregForest and qrnn

- Python, via Scikit-garden
- SAS through proc quantreg (ver. 9.2) and proc quantselect (ver. 9.3).
- Stata, via the qreg command.
- Vowpal Wabbit, via —loss_function quantile.
- Statsmodels package for Python, via —QuantReg
- Mathematica package QuantileRegression.m[1] hosted at the MathematicaForPrediction project at GitHub.

Isotonic Regression

In statistics, isotonic regression or monotonic regression is the technique of fitting a free-form line to a sequence of observations under the following constraints: the fitted free-form line has to be non-decreasing everywhere, and it has to lie as close to the observations as possible.

Applications

Isotonic regression has applications in statistical inference. For example, one might use it to fit an isotonic curve to the means of some set of experimental results when an increase in those means according to some particular ordering is expected. A benefit of isotonic regression is that it is not constrained by any functional form, such as the linearity imposed by linear regression, as long as the function is monotonic increasing. Another application is nonmetric multidimensional scaling, where a low-dimensional embedding for data points is sought such that order of distances between points in the embedding matches order of dissimilarity between points. Isotonic regression is used iteratively to fit ideal distances to preserve relative dissimilarity order.

Software for computing isotone (monotonic) regression has been developed for the R statistical package, the Stata statistical package and the Python programming language.

NON-NORMAL ERRORS

Generalized Linear Model

In statistics, the generalized linear model (GLM) is a flexible generalization of ordinary linear regression that allows for response variables that have error distribution models other than a normal distribution. The GLM generalizes linear regression by allowing the linear model to be related to the response variable via a *link function* and by allowing the magnitude of the variance of each measurement to be a function of its predicted value.

Generalized linear models were formulated by John Nelder and Robert Wedderburn as a way of unifying various other statistical models, including linear regression, logistic regression and Poisson regression. They proposed an iteratively reweighted least squares method for maximum likelihood estimation of the model parameters. Maximum-likelihood estimation remains popular and is the default method on many statistical computing packages. Other approaches, including Bayesian approaches and least squares fits to variance stabilized responses, have been developed.

Intuition

Ordinary linear regression predicts the expected value of a given unknown quantity (the *response variable*, a random variable) as a linear combination of a set of observed values (*predictors*). This implies that a constant change in a predictor leads to a constant change in the response variable (*i.e.* a *linear-response model*). This is

appropriate when the response variable has a normal distribution (intuitively, when a response variable can vary essentially indefinitely in either direction with no fixed "zero value", or more generally for any quantity that only varies by a relatively small amount, *e.g.* human heights). However, these assumptions are inappropriate for some types of response variables. For example, in cases where the response variable is expected to be always positive and varying over a wide range, constant input changes lead to geometrically varying, rather than constantly varying, output changes. As an example, a prediction model might predict that 10 degree temperature decrease would lead to 1,000 fewer people visiting the beach is unlikely to generalize well over both small beaches (*e.g.* those where the expected attendance was 50 at a particular temperature) and large beaches (*e.g.* those where the expected attendance was 10,000 at a low temperature). The problem with this kind of prediction model would imply a temperature drop of 10 degrees would lead to 1,000 fewer people visiting the beach, a beach whose expected attendance was 50 at a higher temperature would now be predicted to have the impossible attendance value of -950. Logically, a more realistic model would instead predict a constant *rate* of increased beach attendance (*e.g.* an increase in 10 degrees leads to a doubling in beach attendance, and a drop in 10 degrees leads to a halving in attendance). Such a model is termed an *exponential-response model* (or *log-linear model*, since the logarithm of the response is predicted to vary linearly).

Similarly, a model that predicts a probability of making a yes/no choice (a Bernoulli variable) is even less suitable as a linear-response model, since probabilities are bounded on both ends (they must be between 0 and 1). Imagine, for example, a model that predicts the likelihood of a given person going to the beach as a function of temperature. A reasonable model might predict, for example, that a change in 10 degrees makes a person two times more or less likely to go to the beach. But what does "twice as likely" mean in terms of a probability? It cannot literally mean to double the probability value (*e.g.* 50 per cent becomes 100 per cent, 75 per cent becomes 150 per cent, etc.). Rather, it is the *odds* that are doubling: from 2:1 odds, to 4:1 odds, to 8:1 odds, etc. Such a model is a *log-odds or logistic model*.

Generalized linear models cover all these situations by allowing for response variables that have arbitrary distributions (rather than simply normal distributions), and for an arbitrary function of the response variable (the *link function*) to vary linearly with the predicted values (rather than assuming that the response itself must vary linearly). For example, the case above of predicted number of beach attendees would typically be modeled with a Poisson distribution and a log link, while the case of predicted probability of beach attendance would typically be modeled with a Bernoulli distribution (or binomial distribution, depending on exactly how the problem is phrased) and a log-odds (or *logit*) link function.

Correlated or Clustered Data

The standard GLM assumes that the observations are uncorrelated. Extensions have been developed to allow for correlation between observations, as occurs for example in longitudinal studies and clustered designs:

- Generalized estimating equations (GEEs) allow for the correlation between observations without the use of an explicit probability model for the origin of the correlations, so there is no explicit likelihood. They are suitable when the random effects and their variances are not of inherent interest, as they allow for the correlation without explaining its origin. The focus is on estimating the average response over the population ("population-averaged" effects) rather than the regression parameters that would enable prediction of the effect of changing one or more components of X on a given individual. GEEs are usually used in conjunction with Huber-White standard errors.
- Generalized linear mixed models (GLMMs) are an extension to GLMs that includes random effects in the linear predictor, giving an explicit probability model that explains the origin of the correlations. The resulting "subject-specific" parameter estimates are suitable when the focus is on estimating the

effect of changing one or more components of X on a given individual. GLMMs are also referred to as multilevel models and as mixed model. In general, fitting GLMMs is more computationally complex and intensive than fitting GEEs.

Confusion with General Linear Models

The term "generalized linear model", and especially its abbreviation GLM, are sometimes confused with general linear model. Co-originator John Nelder has expressed regret over this terminology.

Binomial Regression

In statistics, binomial regression is a technique in which the response (often referred to as *Y*) is the result of a series of Bernoulli trials, or a series of one of two possible disjoint outcomes (traditionally denoted "success" or 1, and "failure" or 0). In binomial regression, the probability of a success is related to explanatory variables: the corresponding concept in ordinary regression is to relate the mean value of the unobserved response to explanatory variables. Binomial regression models are essentially the same as binary choice models, one type of discrete choice model. The primary difference is in the theoretical motivation: Discrete choice models are motivated using utility theory so as to handle various types of correlated and uncorrelated choices, while binomial regression models are generally described in terms of the generalized linear model, an attempt to generalize various types of linear regression models. As a result, discrete choice models are usually described primarily with a latent variable indicating the "utility" of making a choice, and with randomness introduced through an error variable distributed according to a specific probability distribution. Note that the latent variable itself is not observed, only the actual choice, which is assumed to have been made if the net utility was greater than 0. Binary regression models, however, dispense with both the latent and error variable and assume that the choice itself is a random variable, with a link function that transforms the expected value of the choice variable into a value that is then predicted by the linear predictor. It can be shown that the two are equivalent, at least in the case of binary choice models: the link function corresponds to the quantile function of the distribution of the error variable, and the inverse link function to the cumulative distribution function (CDF) of the error variable. The latent variable has an equivalent if one imagines generating a uniformly distributed number between 0 and 1, subtracting from it the mean (in the form of the linear predictor transformed by the inverse link function), and inverting the sign. One then has a number whose probability of being greater than 0 is the same as the probability of success in the choice variable, and can be thought of as a latent variable indicating whether a 0 or 1 was chosen. In machine learning, binomial regression is considered a special case of probabilistic classification, and thus a generalization of binary classification.

Example Application

In one published example of an application of binomial regression, the details were as follows. The observed outcome variable was whether or not a fault occurred in an industrial process. There were two explanatory variables: the first was a simple two-case factor representing whether or not a modified version of the process was used and the second was an ordinary quantitative variable measuring the purity of the material being supplied for the process.

Poisson Regression

In statistics, Poisson regression is a generalized linear model form of regression analysis used to model count data and contingency tables. Poisson regression assumes the response variable *Y* has a Poisson distribution,

and assumes the logarithm of its expected value can be modeled by a linear combination of unknown parameters. A Poisson regression model is sometimes known as a log-linear model, especially when used to model contingency tables.

Negative binomial regression is a popular generalization of Poisson regression because it loosens the highly restrictive assumption that the variance is equal to the mean made by the Poisson model. The traditional negative binomial regression model, commonly known as NB2, is based on the Poisson-gamma mixture distribution. This model is popular because it models the Poisson heterogeneity with a gamma distribution.

Poisson regression models are generalized linear models with the logarithm as the (canonical) link function, and the Poisson distribution function as the assumed probability distribution of the response.

Logistic Regression

In statistics, the logistic model (or logit model) is a widely used statistical model that, in its basic form, uses a logistic function to model a binary dependent variable; many more complex extensions exist. In regression analysis, logistic regression (or logit regression) is estimating the parameters of a logistic model; it is a form of binomial regression. Mathematically, a binary logistic model has a dependent variable with two possible values, such as pass/fail, win/lose, alive/dead or healthy/sick; these are represented by an indicator variable, where the two values are labeled "0" and "1". In the logistic model, the log-odds (the logarithm of the odds) for the value labeled "1" is a linear combination of one or more independent variables ("predictors"); the independent variables can each be a binary variable (two classes, coded by an indicator variable) or a continuous variable (any real value). The corresponding probability of the value labeled "1" can vary between 0 (certainly the value "0") and 1 (certainly the value "1"), hence the labeling; the function that converts log-odds to probability is the logistic function, hence the name. The unit of measurement for the log-odds scale is called a *logit*, from *logistic unit*, hence the alternative names. Analogous models with a different sigmoid function instead of the logistic function can also be used, such as the probit model; the defining characteristic of the logistic model is that increasing one of the independent variables multiplicatively scales the odds of the given outcome at a *constant* rate, with each dependent variable having its own parameter; for a binary independent variable this generalizes the odds ratio.

Logistic regression was developed by statistician David Cox in 1958. The binary logistic regression model has extensions to more than two levels of the dependent variable: categorical outputs with more than two values are modelled by multinomial logistic regression, and if the multiple categories are ordered, by ordinal logistic regression, for example the proportional odds ordinal logistic model. The model itself simply models probability of output in terms of input, and does not perform statistical classification (it is not a classifier), though it can be used to make a classifier, for instance by choosing a cutoff value and classifying inputs with probability greater than the cutoff as one class, below the cutoff as the other; this is a common way to make a binary classifier. The coefficients are generally not computed by a closed-form expression, unlike linear least squares.

Applications

Logistic regression is used in various fields, including machine learning, most medical fields, and social sciences. For example, the Trauma and Injury Severity Score (TRISS), which is widely used to predict mortality in injured patients, was originally developed by Boyd et al. using logistic regression. Many other medical scales used to assess severity of a patient have been developed using logistic regression. Logistic regression may be used to predict the risk of developing a given disease (*e.g.* diabetes; coronary heart disease), based on observed characteristics of the patient (age, sex, body mass index, results of various blood tests, etc.). Another example might be to predict whether an Indian voter will vote BJP or Trinamool Congress or Left Front or Congress, based on age, income, sex, race, state of residence, votes in previous elections, etc. The technique

can also be used in engineering, especially for predicting the probability of failure of a given process, system or product. It is also used in marketing applications such as prediction of a customer's propensity to purchase a product or halt a subscription, etc. In economics it can be used to predict the likelihood of a person's choosing to be in the labour force, and a business application would be to predict the likelihood of a homeowner defaulting on a mortgage. Conditional random fields, an extension of logistic regression to sequential data, are used in natural language processing.

Logistic Regression vs. other Approaches

Logistic regression measures the relationship between the categorical dependent variable and one or more independent variables by estimating probabilities using a logistic function, which is the cumulative logistic distribution. Thus, it treats the same set of problems as probit regression using similar techniques, with the latter using a cumulative normal distribution curve instead. Equivalently, in the latent variable interpretations of these two methods, the first assumes a standard logistic distribution of errors and the second a standard normal distribution of errors. Logistic regression can be seen as a special case of the generalized linear model and thus analogous to linear regression. The model of logistic regression, however, is based on quite different assumptions (about the relationship between dependent and independent variables) from those of linear regression. In particular the key differences between these two models can be seen in the following two features of logistic regression. First, the conditional distribution y | x is a Bernoulli distribution rather than a Gaussian distribution, because the dependent variable is binary. Second, the predicted values are probabilities and are therefore restricted to (0,1) through the logistic distribution function because logistic regression predicts the probability of particular outcomes rather than the outcomes themselves. Logistic regression is an alternative to Fisher's 1936 method, linear discriminant analysis. If the assumptions of linear discriminant analysis hold, the conditioning can be reversed to produce logistic regression. The converse is not true, however, because logistic regression does not require the multivariate normal assumption of discriminant analysis.

Formal Mathematical Specification

There are various equivalent specifications of logistic regression, which fit into different types of more general models. These different specifications allow for different sorts of useful generalizations.

Setup

The basic setup of logistic regression is as follows. We are given a dataset containing N points. Each point i consists of a set of m input variables $x_{1,i}$... $x_{m,i}$ (also called independent variables, predictor variables, features, or attributes), and a binary outcome variable Y_i (also known as a dependent variable, response variable, output variable, or class), *i.e.* it can assume only the two possible values 0 (often meaning "no" or "failure") or 1 (often meaning "yes" or "success"). The goal of logistic regression is to use the dataset to create a predictive model of the outcome variable.

Some examples:

- The observed outcomes are the presence or absence of a given disease (*e.g.* diabetes) in a set of patients, and the explanatory variables might be characteristics of the patients thought to be pertinent (sex, race, age, blood pressure, body-mass index, etc.).
- The observed outcomes are the votes (*e.g.* Democratic or Republican) of a set of people in an election, and the explanatory variables are the demographic characteristics of each person (*e.g.* sex, race, age, income, etc.). In such a case, one of the two outcomes is arbitrarily coded as 1, and the other as 0.

As in linear regression, the outcome variables Y_i are assumed to depend on the explanatory variables $x_{1,i}$... $x_{m,i}$.

Explanatory Variables

As shown above in the above examples, the explanatory variables may be of any type: real-valued, binary, categorical, etc. The main distinction is between continuous variables (such as income, age and blood pressure) and discrete variables (such as sex or race). Discrete variables referring to more than two possible choices are typically coded using dummy variables (or indicator variables), that is, separate explanatory variables taking the value 0 or 1 are created for each possible value of the discrete variable, with a 1 meaning "variable does have the given value" and a 0 meaning "variable does not have that value". For example, a four-way discrete variable of blood type with the possible values "A, B, AB, O" can be converted to four separate two-way dummy variables, "is-A, is-B, is-AB, is-O", where only one of them has the value 1 and all the rest have the value 0.

This allows for separate regression coefficients to be matched for each possible value of the discrete variable. (In a case like this, only three of the four dummy variables are independent of each other, in the sense that once the values of three of the variables are known, the fourth is automatically determined. Thus, it is necessary to encode only three of the four possibilities as dummy variables. This also means that when all four possibilities are encoded, the overall model is not identifiable in the absence of additional constraints such as a regularization constraint. Theoretically, this could cause problems, but in reality almost all logistic regression models are fitted with regularization constraints.)

Bayesian

In a Bayesian statistics context, prior distributions are normally placed on the regression coefficients, usually in the form of Gaussian distributions. There is no conjugate prior of the likelihood function in logistic regression. When Bayesian inference was performed analytically, this made the posterior distribution difficult to calculate except in very low dimensions. Now, though, automatic software such as OpenBUGS, JAGS, PyMC3 or Stan allow these posteriors to be computed using simulation, so lack of conjugacy is not a concern. However, when the sample size or the number of parameters is large, full Bayesian simulation can be slow, and people often use approximate methods such as variational Bayes and expectation propagation.

Extensions

There are large numbers of extensions:

- Multinomial logistic regression (or multinomial logit) handles the case of a multi-way categorical dependent variable (with unordered values, also called "classification"). Note that the general case of having dependent variables with more than two values is termed *polytomous regression*.
- Ordered logistic regression (or ordered logit) handles ordinal dependent variables (ordered values).
- Mixed logit is an extension of multinomial logit that allows for correlations among the choices of the dependent variable.
- An extension of the logistic model to sets of interdependent variables is the conditional random field.
- Conditional logistic regression handles matched or stratified data when the strata are small. It is mostly used in the analysis of observational studies.

Software

Most statistical software can do binary logistic regression.

- SPSS
  (a) For basic logistic regression.

- Stata
- SAS
  (a) PROC LOGISTIC for basic logistic regression.
  (b) PROC CATMOD when all the variables are categorical.
  (c) PROC GLIMMIX for multilevel model logistic regression.
- R
  (a) glm in the stats package (using family = binomial)
  (b) lrm in the rms package
  (c) GLMNET package for an efficient implementation regularized logistic regression
  (d) lmer for mixed effects logistic regression
  (e) Rfast package command gm_logistic for fast and heavy calculations involving large scale data.
  (f) Arm package for bayesian logistic regression
- Python
  (a) Logit in the Statsmodels module.
  (b) LogisticRegression in the Scikit-learn module.
  (c) LogisticRegressor in the TensorFlow module.
  (d) Full example of logistic regression in the Theano tutorial [3]
  (e) Bayesian Logistic Regression with ARD prior code, tutorial
  (f) Variational Bayes Logistic Regression with ARD prior code, tutorial
  (g) Bayesian Logistic Regression code, tutorial
- NCSS
  (a) Logistic Regression in NCSS
- Matlab
  (a) mnrfit in the Statistics and Machine Learning Toolbox (with "incorrect" coded as 2 instead of 0)

Notably, Microsoft Excel's statistics extension package does not include it.

POWER

The power of a binary hypothesis test is the probability that the test correctly rejects the null hypothesis (H_0) when a specific alternative hypothesis (H_1) is true. The statistical power ranges from 0 to 1, and as statistical power increases, the probability of making a type II error (wrongly failing to reject the null) decreases. For a type II error probability of β, the corresponding statistical power is 1-β. For example, if experiment 1 has a statistical power of 0.7, and experiment 2 has a statistical power of 0.95, then there is a stronger probability that experiment 1 had a type II error than experiment 2, and experiment 2 is more reliable than experiment 1 due to the reduction in probability of a type II error. It can be equivalently thought of as the probability of accepting the alternative hypothesis (H_1) when it is true—that is, the ability of a test to detect a specific effect, if that specific effect actually exists. That is,

$$\text{power} = \Pr\left(\text{reject } H_0 \,\middle|\, H_1 \text{ is true}\right).$$

If H_1 is not an equality but rather simply the negation of H_0 (so for example with H_0: $\mu = 0$ for some unobserved population parameter μ, we have simply H_1: $\mu \neq 0$) then power cannot be calculated unless probabilities are known for all possible values of the parameter that violate the null hypothesis. Thus one generally refers to a test's power *against a specific alternative hypothesis*.

As the power increases, there is a decreasing probability of a type II error, also referred to as the false negative rate (β) since the power is equal to 1 - β. A similar concept is the type I error probability, also referred

to as the "false positive rate" or the level of a test under the null hypothesis. Power analysis can be used to calculate the minimum sample size required so that one can be reasonably likely to detect an effect of a given size. For example: "how many times do I need to toss a coin to conclude it is rigged by a certain amount?" Power analysis can also be used to calculate the minimum effect size that is likely to be detected in a study using a given sample size. In addition, the concept of power is used to make comparisons between different statistical testing procedures: for example, between a parametric test and a nonparametric test of the same hypothesis. In the context of binary classification, the power of a test is called its statistical sensitivity, its true positive rate, or its probability of detection.

BACKGROUND

Statistical tests use data from samples to assess, or make inferences about, a statistical population. In the concrete setting of a two-sample comparison, the goal is to assess whether the mean values of some attribute obtained for individuals in two sub-populations differ. For example, to test the null hypothesis that the mean scores of men and women on a test do not differ, samples of men and women are drawn, the test is administered to them, and the mean score of one group is compared to that of the other group using a statistical test such as the two-sample *z*-test. The power of the test is the probability that the test will find a statistically significant difference between men and women, as a function of the size of the true difference between those two populations.

FACTORS INFLUENCING POWER

Statistical power may depend on a number of factors. Some factors may be particular to a specific testing situation, but at a minimum, power nearly always depends on the following three factors:

- The statistical significance criterion used in the test
- The magnitude of the effect of interest in the population
- The sample size used to detect the effect

A significance criterion is a statement of how unlikely a positive result must be, if the null hypothesis of no effect is true, for the null hypothesis to be rejected. The most commonly used criteria are probabilities of 0.05 (5 per cent, 1 in 20), 0.01 (1 per cent, 1 in 100), and 0.001 (0.1 per cent, 1 in 1000). If the criterion is 0.05, the probability of the data implying an effect at least as large as the observed effect when the null hypothesis is true must be less than 0.05, for the null hypothesis of no effect to be rejected. One easy way to increase the power of a test is to carry out a less conservative test by using a larger significance criterion, for example 0.10 instead of 0.05.

This increases the chance of rejecting the null hypothesis (*i.e.* obtaining a statistically significant result) when the null hypothesis is false; that is, it reduces the risk of a type II error (false negative regarding whether an effect exists). But it also increases the risk of obtaining a statistically significant result (*i.e.* rejecting the null hypothesis) when the null hypothesis is not false; that is, it increases the risk of a type I error (false positive).

The magnitude of the effect of interest in the population can be quantified in terms of an effect size, where there is greater power to detect larger effects. An effect size can be a direct value of the quantity of interest, or it can be a standardized measure that also accounts for the variability in the population. For example, in an analysis comparing outcomes in a treated and control population, the difference of outcome means Y X would be a direct estimate of the effect size, whereas (Y X)/σ where σ is the common standard deviation of the outcomes in the treated and control groups, would be an estimated standardized effect size. If constructed appropriately, a standardized effect size, along with the sample size, will completely determine the power. An unstandardized (direct) effect size will rarely be sufficient to determine the power, as it does not contain information about the variability in the measurements..

The sample size determines the amount of sampling error inherent in a test result. Other things being equal, effects are harder to detect in smaller samples. Increasing sample size is often the easiest way to boost the statistical power of a test. How increased sample size translates to higher power is a measure of the efficiency of the test—for example, the sample size required for a given power. The precision with which the data are measured also influences statistical power. Consequently, power can often be improved by reducing the measurement error in the data. A related concept is to improve the "reliability" of the measure being assessed (as in psychometric reliability). The design of an experiment or observational study often influences the power. For example, in a two-sample testing situation with a given total sample size n, it is optimal to have equal numbers of observations from the two populations being compared (as long as the variances in the two populations are the same). In regression analysis and analysis of variance, there are extensive theories and practical strategies for improving the power based on optimally setting the values of the independent variables in the model.

INTERPRETATION

Although there are no formal standards for power (sometimes referred to as π), most researchers assess the power of their tests using $\pi = 0.80$ as a standard for adequacy. This convention implies a four-to-one trade off between β-risk and α-risk. (β is the probability of a Type II error, and α is the probability of a Type I error; 0.2 and 0.05 are conventional values for β and α). However, there will be times when this 4-to-1 weighting is inappropriate.

In medicine, for example, tests are often designed in such a way that no false negatives (Type II errors) will be produced. But this inevitably raises the risk of obtaining a false positive (a Type I error). The rationale is that it is better to tell a healthy patient "we may have found something—let's test further," than to tell a diseased patient "all is well."

Power analysis is appropriate when the concern is with the correct rejection of a false null hypothesis. In many contexts, the issue is less about determining if there is or is not a difference but rather with getting a more refined estimate of the population effect size. For example, if we were expecting a population correlation between intelligence and job performance of around 0.50, a sample size of 20 will give us approximately 80 per cent power (alpha = 0.05, two-tail) to reject the null hypothesis of zero correlation. However, in doing this study we are probably more interested in knowing whether the correlation is 0.30 or 0.60 or 0.50. In this context we would need a much larger sample size in order to reduce the confidence interval of our estimate to a range that is acceptable for our purposes. Techniques similar to those employed in a traditional power analysis can be used to determine the sample size required for the width of a confidence interval to be less than a given value.

Many statistical analyses involve the estimation of several unknown quantities. In simple cases, all but one of these quantities are nuisance parameters. In this setting, the only relevant power pertains to the single quantity that will undergo formal statistical inference. In some settings, particularly if the goals are more "exploratory", there may be a number of quantities of interest in the analysis. For example, in a multiple regression analysis we may include several covariates of potential interest. In situations such as this where several hypotheses are under consideration, it is common that the powers associated with the different hypotheses differ. For instance, in multiple regression analysis, the power for detecting an effect of a given size is related to the variance of the covariate. Since different covariates will have different variances, their powers will differ as well.

Any statistical analysis involving multiple hypotheses is subject to inflation of the type I error rate if appropriate measures are not taken. Such measures typically involve applying a higher threshold of stringency to reject a hypothesis in order to compensate for the multiple comparisons being made (*e.g.* as in the Bonferroni method). In this situation, the power analysis should reflect the multiple testing approach to be used. Thus, for example,

a given study may be well powered to detect a certain effect size when only one test is to be made, but the same effect size may have much lower power if several tests are to be performed. It is also important to consider the statistical power of a hypothesis test when interpreting its results. A test's power is the probability of correctly rejecting the null hypothesis when it is false; a test's power is influenced by the choice of significance level for the test, the size of the effect being measured, and the amount of data available. A hypothesis test may fail to reject the null, for example, if a true difference exists between two populations being compared by a t-test but the effect is small and the sample size is too small to distinguish the effect from random chance. Many clinical trials, for instance, have low statistical power to detect differences in adverse effects of treatments, since such effects may be rare and the number of affected patients small.

A PRIORI VS. POST HOC ANALYSIS

Power analysis can either be done before (*a priori* or prospective power analysis) or after (*post hoc* or retrospective power analysis) data are collected. *A priori* power analysis is conducted prior to the research study, and is typically used in estimating sufficient sample sizes to achieve adequate power. *Post-hoc* analysis of "observed power" is conducted after a study has been completed, and uses the obtained sample size and effect size to determine what the power was in the study, assuming the effect size in the sample is equal to the effect size in the population. Whereas the utility of prospective power analysis in experimental design is universally accepted, post hoc power analysis is fundamentally flawed. Falling for the temptation to use the statistical analysis of the collected data to estimate the power will result in uninformative and misleading values. In particular, it has been shown that *post-hoc* "observed power" is a one-to-one function of the *p*-value attained. This has been extended to show that all *post-hoc* power analyses suffer from what is called the "power approach paradox" (PAP), in which a study with a null result is thought to show *more* evidence that the null hypothesis is actually true when the *p*-value is smaller, since the apparent power to detect an actual effect would be higher. In fact, a smaller *p*-value is properly understood to make the null hypothesis *relatively* less likely to be true.

APPLICATION

Funding agencies, ethics boards and research review panels frequently request that a researcher perform a power analysis, for example to determine the minimum number of animal test subjects needed for an experiment to be informative. In frequentist statistics, an underpowered study is unlikely to allow one to choose between hypotheses at the desired significance level. In Bayesian statistics, hypothesis testing of the type used in classical power analysis is not done. In the Bayesian framework, one updates his or her prior beliefs using the data obtained in a given study. In principle, a study that would be deemed underpowered from the perspective of hypothesis testing could still be used in such an updating process. However, power remains a useful measure of how much a given experiment size can be expected to refine one's beliefs. A study with low power is unlikely to lead to a large change in beliefs.

EXTENSION

Bayesian Power

In the frequentist setting, parameters are assumed to have a specific value which is unlikely to be true. This issue can be addressed by assuming the parameter has a distribution. The resulting power is sometimes referred to as Bayesian power which is commonly used in clinical trial design.

Predictive Probability of Success

Both frequentist power and Bayesian power use statistical significance as the success criterion. However, statistical significance is often not enough to define success. To address this issue, the power concept can be extended to the concept of predictive probability of success (PPOS). The success criterion for PPOS is not restricted to statistical significance and is commonly used in clinical trial designs.

SOFTWARE FOR POWER AND SAMPLE SIZE CALCULATIONS

Numerous free and/or open source programmes are available for performing power and sample size calculations. These include:

- G*Power
- WebPower Free online statistical power analysis
- powerandsamplesize.com Free and open source online calculators
- PowerUp! provides convenient excel-based functions to determine minimum detectable effect size and minimum required sample size for various experimental and quasi-experimental designs.
- PowerUpR is R package version of PowerUp! and additionally includes functions to determine sample size for various multilevel randomized experiments with or without budgetary constraints.
- R package pwr
- R package WebPower

Bibliography

Agresti, A.: *Categorical Data Analysis*, New York: Wiley, 2000.

Ahlquist, D. A., D. B. McGill, S. Schwartz, and W. F. Taylor: *Fecal blood levelsin health and disease: A study using HemoQuant*, *New England Journal of Medicine,* 2005.

Anderson, J. W. et al.: *Oat bran cereal lowers serum cholesterol total and LDLcholesterol in hypercholesterolemic men. American Journal of Clinical Nutrition,* 2000.

Armitage, P.: *Statistical Methods in Medical Research*. New York: Wiley, 2007.

Arsenault, P. S.: *Maternal and antenatal factors in the risk of sudden infant deathsyndrome. American Journal of Epidemiology,* 2000.

Bamber, D.: *The area above the ordinal dominance graph and the area below thereceiver operating graph. Journal of Mathematical Psychology*, 2005.

Begg, C. B. and B. McNeil: *Assessment of radiologic tests: Control of bias andother design considerations. Radiology,* 2005.

Berkowitz, G. S.: *An epidemiologic study of pre-term delivery. American Journal of Epidemiology,* 2001.

Berry, G.: *The analysis of mortality by the subject–years method. Biometrics,* 2003.

Biracree, T.: (1984). *Your intelligence quotient, in How You Rate*. *Lung cancer after employment in shipyards during World WarII. New England Journal of Medicine*, New York: Dell Publishing. Blot, W. J. et al. 2009..

Breslow, N. E. and N. E. Day: *Statistical Methods in Cancer Research, Vol. I:The Analysis of Case–Control Studies*, Lyons, France: International Agency for Researchon Cancer, 2000.

Breslow, N.: *A generalized Kruskal–Wallis test for comparing K samples subjectto unequal patterns of censorship. Biometrika,* 2000.

Breslow, N.: *Covariance adjustment of relative-risk estimates in matched studies. Biometrics* 2002.

Brown, B. W., R. G. Miller, B. Efron, B. W. Brown, and L. E. Mose: *Prediction analyses for binary data. In Biostatistics Casebook* NewYork: Wiley, 1980.

Chin, T., W. Marine, E. Hall, C. Gravelle, and J. Speers: *The influence of Salkvaccination on the epidemic pattern and the spread of the virus in the community.American Journal of Hygiene,* 2001.

Cohen, J.: *A coecient of agreement for nominal scale. Educational and PsychologicalMeasuments,* 2000.

Coren, S.: *Left-handedness and accident-related injury risk. American Journal of Public Health,* 2009.

Cox, D. R. and D. Oakes: *Analysis of Survival Data*, New York: Chapman & Hall, 2004.

Cox, D. R. and E. J. Snell: *The Analysis of Binary Data, 2nd ed*., London:Chapman & Hall, 2009.

Cox, D. R.: *Regression models and life tables. Journal of the Royal StatisticalSociety*, 2002.

D'Angelo, L. J., J. C. Hierholzer, R. C. Holman, and J. D. Smith: *Epidemickeratoconjunctivitis caused by adenovirus type 8: Epidemiologic and laboratoryaspects of a large outbreak. American Journal of Epidemiology,* 2001.

Daniel, W. W.: *Biostatistics: A Foundation for Analysis in the Health Sciences*, New York: Wiley, 2007.

Dienstag, J. L. and D. M. Ryan: *Occupational exposure to hepatitis B virus inhospital personnel: Infection or immunization. American Journal of Epidemiology,* 2002.

Douglas, G.: *Drug therapy. New England Journal of Medicine. Influence of age on secretion of cholesterol and synthesis ofbile acids by the liver. New England Journal of Medicine,* 2000.

Engs, R. C. and D. J. Hanson: *University students' drinking patterns and problems:Examining the e¤ects of raising the purchase age. Public Health Reports,* 2008.

Fiskens, E. J. M. and D. Kronshout: *Cardiovascular risk factors and the 25 yearincidence of diabetes mellitus in middle-aged men. American Journal of Epidemiology,* 2009.

Fowkes, F. G. R. et al".: *Smoking, lipids, glucose intolerance, and blood pressureas risk factors for peripheral atherosclerosis compared with ischemic heart diseasein the Edinburgh Artery Study. American Journal of Epidemiology*, 2002.

Fox, A. J. and P. F. Collier: *Low mortality rates in industrial cohort studies dueto selection for work and survival in the industry. British Journal of Preventive and Social Medicine,* 2006.

Freeman, D. H.: *Applied Categorical Data Analysis*, New York: Marcel Dekker, 2000.

Freireich, E. J. et al.: *The e¤ect of 6-mercaptopurine on the duration of steroidinducedremissions in acute leukemia: A model for evaluation of other potentiallyuseful therapy. Blood*, 2003.

Frerichs, R. R. et al.: *Prevalence of depression in Los Angeles County. AmericanJournal of Epidemiology,* 2001.

Frome, E. L. and H. Checkoway: *Use of Poisson regression models in estimatingrates and ratios. American Journal of Epidemiology,* 2005.

Frome, E. L.: *The analysis of rates using Poisson regression models. Biometrics,* 2003.

Fulwood, R. et al.: *Total serum cholesterol levels of adults 20–74 years of age:United States, 1976–1980. Vital and Health Statistics*, 2006.

Gehan, E. A.: *A generalized two-sample Wilcoxon test for doubly censoreddata. Biometrika,* 2005.

Gehan, E. A.: *A generalized Wilcoxon test for comparing arbitrarily singlycensoredsamples. Biometrika*, 2005.

Grady, W. R. et al.: *Contraceptives failure in the United States: Estimates fromthe 1982 National Survey of Family Growth. Family Planning Perspectives,* 2006.

Graham, S. et al.: *Dietary epidemiology of cancer of the colon in western New York. American Journal of Epidemiology,* 2008.

Greenwood, M.: *The natural duration of cancer. Reports on Public Health andMedical Subjects, Her Majesty's Stationary*, 2006.

Gwirtsman, H. E. et al.: *Decreased caloric intake in normal weight patients withbulimia: Comparison with female volunteers. American Journal of Clinical Nutrition*, 2009.

Hanley, J. A. and B. J. McNeil: *Method for comparing the area under the ROC curves derived from the same cases. Radiology,* 2003.

Hanley, J. A. and B. J. McNeil: *The meaning and use of the area under a receiver operating characteristic (ROC) curve. Radiology,* 2002.

Helsing, K. J. and M. Szklo: *Mortality after bereavement. American Journal of Epidemiology,* 2001.

Herbst, A. L., H. Ulfelder, and D. C. Poskanzer.: *Adenocarcinoma of the vagina. New England Journal of Medicine,* 2002.

Hiller, R. and A. H. Kahn: *Blindness from glaucoma. British Journal of Ophthalmology,* 2006.

Hlatky, M. A., D. B. Pryor, F. E. Harrell, R. M. Cali¤, D. B. Mark, and R. A. Rosati: *Factors aoecting sensitivity and specificity of exercise electrocardiography: Multivariate analysis. American Journal of Medicine,* 2004.

Hollows, F. C. and P. A. Graham: *Intraocular pressure, glaucoma, and glaucoma suspects in a defined population. British Journal of Ophthalmology,* 2006.

Hosmer, D. W., Jr. and S. Lemeshow: *Applied Logistic Regression*, New York: Wiley, 2009.

Hsieh, F. Y.: *Sample size tables for logistic regression. Statistics in Medicine 8*, 2009.

Jackson, R. et al.: *Does recent alcohol consumption reduce the risk of acute myocardial infarction and coronary death in regular drinkers? American Journal of Epidemiology,* 2002.

Kaplan, E. L. and P. Meier: *Nonparametric estimation from incomplete observations. Journal of the American Statistical Association,* 2008.

Kay, R. and S. Little: *Transformations of the explanatory variables in the logistic regression model for binary data. Biometrika,* 2007.

Kelsey, J. L., V. A. Livolsi, T. R. Holford, D. B. Fischer, E. D. Mostow, P. E. Schartz, T. O'Connor, and C. White: *A case–control study of cancer of the endometrium. American Journal of Epidemiology,* 2002.

Khabbaz, R. et al.: *Epidemiologic assessment of screening tests for antibody to human T lymphotropic virus type I. American Journal of Public Health,* 2000.

Kleinbaum, D. G., L. L. Kupper, and K. E. Muller: *Applied Regression Analysis and Other Multivariate Methods*, Boston: PWS-Kent, 2008.

Kleinman, J. C. and A. Kopstein: *Who is being screened for cervical cancer? American Journal of Public Health,* 2001.

Klinhamer, P. J. J. M. et al.: *Intraobserver and interobserver variability in the quality assessment of cervical smears. Acta Cytologica,* 2009.

Knowler, W. C. et al.: *Diabetes incidence in Pima Indians: Contributions of obesity and parental diabetes. American Journal of Epidemiology,* 2001.

Koenig, J. Q. et al.: *Prior exposure to ozone potentiates subsequent response to sulfur dioxide in adolescent asthmatic subjects. American Review of Respiratory Disease,* 2000.

Kono, S. et al.: *Prevalence of gallstone disease in relation to smoking, alcohol use, obesity, and glucose tolerance: A study of self-defense ocials in Japan. American Journal of Epidemiology,* 2002.

Kushi, L. H. et al.: *The association of dietary fat with serum cholesterol in vegetarians: The eoects of dietary assessment on the correlation coecient. American Journal of Epidemiology,* 2008.

Lachin, J. M.: *Intoduction to sample size determination and power analysis for clinical trials. Controlled Clinical Trials 2*: *Applied Survival Analysis*, New York: Wiley, 2007.

Le, C. T. and B. R. Lindgren: *Computational implementation of the conditional logistic regression model in the analysis of epidemiologic matched studies. Computers and Biomedical Research,* 2008.

Le, C. T.: *Evaluation of confounding eoects in ROC studies. Biometrics,* 2007.

Lee, M.: *Improving patient comprehension of literature on smoking. American Journal of Public Health. Prior condom use and the risk of tubal pregnancy. American Journal of Public Health,* 2009.

Mack, T. M. et al.: *Estrogens and endometrial cancer in a retirement community. New England Journal of Medicine*, 2006.

Makuc, D. et al.: *National trends in the use of preventive health care by women. American Journal of Public Health*, 2009.

Mantel, N. and W. Haenszel: *Statistical aspects of the analysis of data from retrospective studies of disease. Journal of the National Cancer Institute,* 2009.

Matinez, F. D. et al.: *Maternal age as a risk factor for wheezing lower respiratory illness in the first year of life. American Journal of Epidemiology,* 2002.

May, D.: *Error rates in cervical cytological screening tests. British Journal of Cancer,* 2004.

McCusker, J. et al.: *Association of electronic fetal monitoring during labor with caesarean section rate with neonatal morbidity and mortality. American Journal of Public Health,* 2008.

Negri, E. et al.: *Risk factors for breast cancer: Pooled results from three Italian case–control studies. American Journal of Epidemiology,* 2008.

Nischan, P. et al.: *Smoking and invasive cervical cancer risk: Results from a case–control study. American Journal of Epidemiology,* 2008.

Nurminen, M. et al.: *Quantitated e¤ects of carbon disulfide exposure, elevated blood pressure and aging on coronary mortality. American Journal of Epidemiology,* 2002.

Ockene, J.: *The relationship of smoking cessation to coronary heart disease and lung cancer in the Multiple Risk Factor Intervention Trial. American Journal of Public Health,* 2000.

Padian, N. S.: *Sexual histories of heterosexual couples with one HIV-infected partner. American Journal of Public Health,* 2009.

Palta, M. et al.: *Comparison of self-reported and measured height and weight. American Journal of Epidemiology,* 2002.

Pappas, G. et al.: *Hypertension prevalence and the status of awareness, treatment, and control in the Hispanic health and nutrition examination survey (HHANES). American Journal of Public Health,* 2000.

Peto, R.: *Contribution to discussion of paper by D. R. Cox. Journal of the Royal Statistical Society, Evaluation of the ecacy of simulation games in trac safety education of kindergarten children. American Journal of Public Health*, 2002.

Renes, R. et al.: *Transmission of multiple drug-resistant tuberculosis: Report of school and community outbreaks. American Journal of Epidemiology*, 2001.

Rosenberg, L. et al.: *Case–control studies on the acute eoects of cocee upon the risk of myocardial infarction: Problems in the selection of a hospital control series. American Journal of Epidemiology,* 2001.

Rossignol, A. M.: *Tea and premenstrual syndrome in the People's Republic of China. American Journal of Public Health,* 2009.

Rousch, G. C. et al.: *Scrotal carcinoma in Connecticut metal workers: Sequel to a study of sinonasal cancer. American Journal of Epidemiology,* 2002.

Salem-Schatz, S. et al.: *Influence of clinical knowledge, organization context and practice style on transfusion decision making. Journal of the American Medical Association,* 2000.

Sandek, C. D. et al.: *A preliminary trial of the programmable implantable medication system for insulin delivery. New England Journal of Medicine,* 2009.

Schwarts, B. et al.: *Olfactory function in chemical workers exposed to acrylate and methacrylate vapors. American Journal of Public Health,* 2009.

Selby, J. V. et al.: *Precursors of essential hypertension: The role of body fat distribution. American Journal of Epidemiology,* 2009.

Shapiro, S. et al.: *Oral contraceptive use in relation to myocardial infarction*, 2009.

Strader, C. H. et al.: *Vasectomy and the incidence of testicular cancer. American Journal of Epidemiology,* 2008.

Index

E

F

G

H

I

L

M

N

O

P

R

S