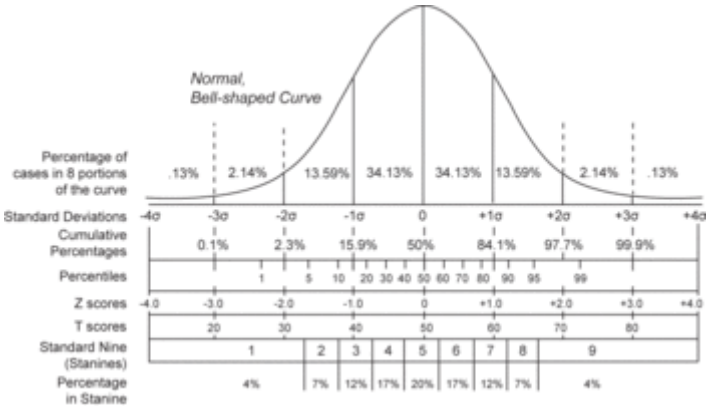
Statistics



Statistics is a mathematical science pertaining to the collection, analysis, interpretation or explanation, and presentation of data. It is applicable to a wide variety of academic disciplines, from the physical and social sciences to the humanities. Statistics are also used for making informed decisions.

Statistical methods can be used to summarize or describe a collection of data; this is called **descriptive statistics**. In addition, patterns in the data may be modeled in a way that accounts for randomness and uncertainty in the observations, and then used to draw inferences about the process or population being studied; this is called **inferential statistics**. Both descriptive and inferential statistics comprise **applied statistics**. There is also a discipline called **mathematical statistics**, which is concerned with the theoretical basis of the subject.

The word ***statistics*** is also the plural of ***statistic*** (singular), which refers to the result of applying a statistical algorithm to a set of data, as in economic statistics, crime statistics, etc.

History

Etymology

The word *statistics* ultimately derives from the New Latin term *statisticum collegium* ("council of state") and the Italian word *statista* ("statesman" or "politician"). The German *Statistik*, first introduced by Gottfried Achenwall (1749), originally designated the analysis of data about the state, signifying the "science of state" (then called *political arithmetic* in English). It acquired the meaning of the collection and classification of data generally in the early 19th century. It was introduced into English by Sir John Sinclair.

Thus, the original principal purpose of *Statistik* was data to be used by governmental and (often centralized) administrative bodies. The collection of data about states and localities continues, largely through national and international statistical services. In particular, censuses provide regular information about the population.

Origins in probability

The mathematical methods of statistics emerged from probability theory, which can be dated to the correspondence of Pierre de Fermat and Blaise Pascal (1654). Christiaan Huygens (1657) gave the earliest known scientific treatment of the subject. Jakob Bernoulli's *Ars Conjectandi* (posthumous, 1713) and Abraham de Moivre's *Doctrine of Chances* (1718) treated the subject as a branch of mathematics. In the modern era, the work of Kolmogorov has been instrumental in formulating the fundamental model of Probability Theory, which is used throughout statistics.

The theory of errors may be traced back to Roger Cotes' *Opera Miscellanea* (posthumous, 1722), but a memoir prepared by Thomas Simpson in 1755 (printed 1756) first applied the theory to the discussion of errors of observation. The reprint (1757) of this memoir lays down the axioms that positive and negative errors are equally probable, and that there are certain assignable limits within which all errors may be supposed to fall; continuous errors are discussed and a probability curve is given.

Pierre-Simon Laplace (1774) made the first attempt to deduce a rule for the combination of observations from the principles of the theory of probabilities. He represented the law of probability of errors by a curve. He deduced a formula for the mean of three observations. He also gave (1781) a formula for the law of facility of error (a term due to Lagrange, 1774), but one which led to unmanageable equations. Daniel Bernoulli (1778) introduced the principle of the maximum product of the probabilities of a system of concurrent errors.

The method of least squares, which was used to minimize errors in data measurement, was published independently by Adrien-Marie Legendre (1805), Robert Adrain (1808), and Carl Friedrich Gauss (1809). Gauss had used the method in his famous 1801 prediction of the location of the dwarf planet Ceres. Further proofs were given by Laplace (1810, 1812), Gauss (1823), James Ivory (1825, 1826), Hagen (1837), Friedrich Bessel (1838), W. F. Donkin (1844, 1856), John Herschel (1850), and Morgan Crofton (1870).

Other contributors were Ellis (1844), De Morgan (1864), Glaisher (1872), and Giovanni Schiaparelli (1875). Peters's (1856) formula for *r*, the probable error of a single observation, is well known.

In the nineteenth century authors on the general theory included Laplace, Sylvestre Lacroix (1816), Littrow (1833), Richard Dedekind (1860), Helmert (1872), Hermann Laurent (1873), Liagre, Didion, and Karl Pearson. Augustus De Morgan and George Boole improved the exposition of the theory.

Adolphe Quetelet (1796-1874), another important founder of statistics, introduced the notion of the "average man" (*l'homme moyen*) as a means of understanding complex social phenomena such as crime rates, marriage rates, or suicide rates.

Statistics today

During the 20th century, the creation of precise instruments for public health concerns (epidemiology, biostatistics, etc.) and economic and social purposes (unemployment rate, econometry, etc.) necessitated substantial advances in statistical practices: the Western welfare states developed after World War I had to possess specific knowledge of the "population".

Today the use of statistics has broadened far beyond its origins as a service to a state or government. Individuals and organizations use statistics to understand data and make informed decisions throughout the natural and social sciences, medicine, business, and other areas.

Statistics is generally regarded not as a subfield of mathematics but rather as a distinct, albeit allied, field. Many universities maintain separate mathematics and statistics departments. Statistics is also taught in departments as diverse as psychology, education, and public health.

Important contributors to statistics

See also: List of statisticians

- Thomas Bayes
- Pafnuty Chebyshev
- Sir David Cox
- Gertrude Cox
- George Dantzig
- W. Edwards Deming
- Bruno de Finetti
- Sir Ronald Fisher
- Sir Francis Galton
- Carl Friedrich Gauss
- William Sealey Gosset ("Student")
- Andrey Kolmogorov
- Aleksandr Lyapunov
- Abraham De Moivre
- Isaac Newton
- Florence Nightingale
- Blaise Pascal
- Karl Pearson
- Adolphe Quetelet
- Walter A. Shewhart
- Charles Spearman
- John Tukey
- C. R. Rao
- Rene Descartes
- George E. P. Box

Conceptual overview

In applying statistics to a scientific, industrial, or societal problem, one begins with a process or population to be studied. This might be a population of people in a country, of crystal grains in a rock, or of goods manufactured by a particular factory during a given period. It may instead be a process observed at various times; data collected about this kind of "population" constitute what is called a time series.

For practical reasons, rather than compiling data about an entire population, one usually instead studies a chosen subset of the population, called a sample. Data are collected about the sample in an observational or experimental setting. The data are then subjected to statistical analysis, which serves two related purposes: description and inference.

- Descriptive statistics can be used to summarize the data, either numerically or graphically, to describe the sample. Basic examples of numerical descriptors include the mean and standard deviation. Graphical summarizations include various kinds of charts and graphs.
- Inferential statistics is used to model patterns in the data, accounting for randomness and drawing inferences about the larger population. These inferences may take the form of answers to yes/no questions (hypothesis testing), estimates of numerical characteristics (estimation), descriptions of association (correlation), or modeling of relationships (regression). Other modeling techniques include ANOVA, time series, and data mining.

The concept of correlation is particularly noteworthy. Statistical analysis of a data set may reveal that two variables (that is, two properties of the population under consideration) tend to vary together, as if they are connected. For example, a study of annual income and age of death among people might find that poor people tend to have shorter lives than affluent people. The two variables are said to be correlated. However, one cannot immediately infer the existence of a causal relationship between the two variables; see correlation does not imply causation. The correlated phenomena could be caused by a third, previously unconsidered phenomenon, called a lurking variable.

If the sample is representative of the population, then inferences and conclusions made from the sample can be extended to the population as a whole. A major problem lies in determining the extent to which the chosen sample is representative. Statistics offers methods to estimate and correct for randomness in the sample and in the data collection procedure, as well as methods for designing robust experiments in the first place; see experimental design.

The fundamental mathematical concept employed in understanding such randomness is probability. Mathematical statistics (also called statistical theory) is the branch of applied mathematics that uses probability theory and analysis to examine the theoretical basis of statistics.

The use of any statistical method is valid only when the system or population under consideration satisfies the basic mathematical assumptions of the method. Misuse of statistics can produce subtle but serious errors in description and interpretation — subtle in that even experienced professionals sometimes make such errors, and serious in that they may affect social policy, medical practice and the reliability of structures such as bridges and nuclear power plants.

Even when statistics is correctly applied, the results can be difficult to interpret for a non-expert. For example, the statistical significance of a trend in the data — which measures the extent to which the trend could be caused by random variation in the sample — may not agree with one's intuitive sense of its significance. The set of basic statistical skills (and skepticism) needed by people to deal with information in their everyday lives is referred to as statistical literacy.

Statistical methods

Experimental and observational studies

A common goal for a statistical research project is to investigate causality, and in particular to draw a conclusion on the effect of changes in the values of predictors or independent variables on response or dependent variables. There are two major types of causal statistical studies, experimental studies and observational studies. In both types of studies, the effect of differences of an independent variable (or variables) on the behavior of the dependent variable are observed. The difference between the two types is in how the study is actually conducted. Each can be very effective.

An experimental study involves taking measurements of the system under study, manipulating the system, and then taking additional measurements using the same procedure to determine if the manipulation may have modified the values of the measurements. In contrast, an observational study does not involve experimental manipulation. Instead data are gathered and correlations between predictors and the response are investigated.

An example of an experimental study is the famous Hawthorne studies which attempted to test changes to the working environment at the Hawthorne plant of the Western Electric Company. The researchers were interested in whether increased illumination would increase the productivity of the assembly line workers. The researchers first measured productivity in the plant then modified the illumination in an area of the plant to see if changes in illumination would affect productivity. As it turns out, productivity improved under all the experimental conditions (see Hawthorne effect). However, the study is today heavily criticized for errors in experimental procedures, specifically the lack of a control group and blindedness.

An example of an observational study is a study which explores the correlation between smoking and lung cancer. This type of study typically uses a survey to collect observations about the area of interest and then perform statistical analysis. In this case, the researchers would collect observations of both smokers and non-smokers and then look at the number of cases of lung cancer in each group.

The basic steps for an experiment are to:

1. plan the research including determining information sources, research subject selection, and ethical considerations for the proposed research and method,
2. design the experiment concentrating on the system model and the interaction of independent and dependent variables,
3. summarize a collection of observations to feature their commonality by suppressing details (descriptive statistics),
4. reach consensus about what the observations tell us about the world we observe (statistical inference),
5. document and present the results of the study.

Levels of measurement

> *See: Stanley Stevens' "Scales of measurement" (1946): nominal, ordinal, interval, ratio*

There are four types of measurements or measurement scales used in statistics. The four types or levels of measurement (nominal, ordinal, interval, and ratio) have different degrees of usefulness in statistical research. Ratio measurements, where both a zero value and distances between different measurements are defined, provide the greatest flexibility in statistical methods that can be used for analysing the data. Interval measurements have meaningful distances between measurements but no meaningful zero value (such as IQ measurements or temperature measurements in Kelvin). Ordinal measurements have imprecise differences between consecutive values but a meaningful order to those values. Nominal measurements have no meaningful rank order among values.

Statistical techniques

Some well known statistical tests and procedures for research observations are:

- Student's t-test
- chi-square test
- Analysis of variance (ANOVA)
- Mann-Whitney U
- Regression analysis
- Factor Analysis
- Correlation
- Pearson product-moment correlation coefficient
- Spearman's rank correlation coefficient

Specialized disciplines

Some fields of inquiry use applied statistics so extensively that they have specialized terminology. These disciplines include:

- Actuarial science
- Applied Information Economics
- Biostatistics
- Business statistics
- Data mining (applying statistics and pattern recognition to discover knowledge from data)
- Economic statistics (Econometrics)
- Energy statistics
- Engineering statistics
- Epidemiology
- Geography and Geographic Information Systems, more specifically in Spatial analysis

- Demography
- Psychological statistics
- Quality
- Social statistics (for all the *social* sciences)
- Statistical literacy
- Statistical surveys
- Process analysis and chemometrics (for analysis of data from analytical chemistry and chemical engineering)
- Reliability engineering
- Image processing
- Statistics in various sports, particularly baseball and cricket

Statistics form a key basis tool in business and manufacturing as well. It is used to understand measurement systems variability, control processes (as in statistical process control or SPC), for summarizing data, and to make data-driven decisions. In these roles it is a key tool, and perhaps the only reliable tool.

Statistical computing

The rapid and sustained increases in computing power starting from the second half of the 20th century have had a substantial impact on the practice of statistical science. Early statistical models were almost always from the class of linear models, but powerful computers, coupled with suitable numerical algorithms, caused a resurgence of interest in nonlinear models (especially neural networks and decision trees) and the creation of new types, such as generalised linear models and multilevel models.

Increased computing power has also led to the growing popularity of computationally-intensive methods based on resampling, such as permutation tests and the bootstrap, while techniques such as Gibbs sampling have made Bayesian methods more feasible. The computer revolution has implications for the future of statistics, with a new emphasis on "experimental" and "empirical" statistics. A large number of both general and special purpose statistical packages are now available to practitioners.

Misuse

There is a general perception that statistical knowledge is all-too-frequently intentionally misused, by finding ways to interpret the data that are favorable to the presenter. A famous quote, variously attributed, but thought to be from Benjamin Disraeli is, "There are three types of lies - lies, damn lies, and statistics." The well-known book *How to Lie with Statistics* by Darrell Huff discusses many cases of deceptive uses of statistics, focusing on misleading graphs. By choosing (or rejecting, or modifying) a certain sample, results can be manipulated; throwing out outliers is one means of doing so. This may be the result of outright fraud or of subtle and unintentional bias on the part of the researcher. Thus, Harvard President Lawrence Lowell wrote in 1909 that statistics, "like

veal pies, are good if you know the person that made them, and are sure of the ingredients."

As further studies contradict previously announced results, people may become wary of trusting such studies. One might read a study that says (for example) "doing X reduces high blood pressure", followed by a study that says "doing X does not affect high blood pressure", followed by a study that says "doing X actually worsens high blood pressure". Often the studies were conducted on different groups with different protocols, or a small-sample study that promised intriguing results has not held up to further scrutiny in a large-sample study. However, many readers may not have noticed these distinctions, or the media may have oversimplified this vital contextual information, and the public's distrust of statistics is thereby increased.

However, deeper criticisms come from the fact that the hypothesis testing approach, widely used and in many cases required by law or regulation, forces one hypothesis to be 'favored' (the null hypothesis), and can also seem to exaggerate the importance of minor differences in large studies. A difference that is highly statistically significant can still be of no practical significance..

In the fields of psychology and medicine, especially with regard to the approval of new drug treatments by the Food and Drug Administration, criticism of the hypothesis testing approach has increased in recent years. One response has been a greater emphasis on the *p*-value over simply reporting whether a hypothesis was rejected at the given level of significance α. Here again, however, this summarises the evidence for an effect but not the size of the effect. One increasingly common approach is to report confidence intervals instead, since these indicate both the size of the effect and the uncertainty surrounding it. This aids in interpreting the results, as the confidence interval for a given α simultaneously indicates both statistical significance and effect size.

Note that both the *p*-value and confidence interval approaches are based on the same fundamental calculations as those entering into the corresponding hypothesis test. The results are stated in a more detailed format, rather than the yes-or-no finality of the hypothesis test, but use the same underlying statistical methodology.

A truly different approach is to use Bayesian methods. This approach has been criticized as well, however. The strong desire to see good drugs approved and harmful or useless drugs restricted remain conflicting tensions (type I and type II errors in the language of hypothesis testing).

In his book *Statistics As Principled Argument*, Robert P. Abelson articulates the position that statistics serves as a standardized means of settling disputes between scientists who could otherwise each argue the merits of their own positions *ad infinitum*. From this point of view, statistics is principally a form of rhetoric. This can be taken as a positive or a negative, but as with any means of settling a dispute, statistical methods can succeed only as long as both sides accept the approach and agree on the method to be used.