Statistics

**Statistics** is the study of the collection, organization, analysis, interpretation and presentation of [data](http://en.wikipedia.org/wiki/Data) It deals with all aspects of this, including the planning of data collection in terms of the design of [surveys](http://en.wikipedia.org/wiki/Statistical_survey) and [experiments](http://en.wikipedia.org/wiki/Experimental_design).

A [statistician](http://en.wikipedia.org/wiki/Statistician) is someone who is particularly well-versed in the ways of thinking necessary for the successful application of statistical analysis. Such people have often gained experience through working in any of a [wide number of fields](http://en.wikipedia.org/wiki/List_of_fields_of_application_of_statistics). There is also a discipline called [*mathematical statistics*](http://en.wikipedia.org/wiki/Mathematical_statistics) that studies statistics mathematically.

The word *statistics*, when referring to the scientific discipline, is singular, as in "Statistics is an art.]This should not be confused with the word *statistic*, referring to a quantity (such as [mean](http://en.wikipedia.org/wiki/Mean) or [median](http://en.wikipedia.org/wiki/Median)) calculated from a set of data, whose plural is *statistics* ("this statistic seems wrong" or "these statistics are misleading").

## Scope

Some consider statistics to be a mathematical body of science pertaining to the collection, analysis, interpretation or explanation, and presentation of [data](http://en.wikipedia.org/wiki/Data), while others consider it a branch of [mathematics](http://en.wikipedia.org/wiki/Mathematics) concerned with collecting and interpreting data. Because of its empirical roots and its focus on applications, statistics is usually considered to be a distinct mathematical science rather than a branch of mathematics. Much of statistics is non-mathematical: ensuring that [data collection](http://en.wikipedia.org/wiki/Data_collection) is undertaken in a way that allows valid conclusions to be drawn; coding and archiving of data so that information is retained and made useful for international comparisons of [official statistics](http://en.wikipedia.org/wiki/Official_statistics); reporting of results and summarised data (tables and graphs) in ways that are comprehensible to those who need to make use of them; implementing procedures that ensure the [privacy of census information](http://en.wikipedia.org/wiki/Census#Privacy).

Statisticians improve the quality of data by coming up with a specific [design of experiments](http://en.wikipedia.org/wiki/Design_of_experiments) and [survey sampling](http://en.wikipedia.org/wiki/Survey_sampling). Statistics itself also provides tools for prediction and forecasting the use of data and [statistical models](http://en.wikipedia.org/wiki/Statistical_model). Statistics is applicable to a wide variety of [academic disciplines](http://en.wikipedia.org/wiki/Academic_discipline), including [natural](http://en.wikipedia.org/wiki/Natural) and [social sciences](http://en.wikipedia.org/wiki/Social_science), government, and business. [Statistical consultants](http://en.wikipedia.org/wiki/Statistical_consultant) are available to provide help for organizations and companies without direct access to expertise relevant to their particular questions.

Statistical methods can be used for summarizing or describing a collection of data; this is called [*descriptive statistics*](http://en.wikipedia.org/wiki/Descriptive_statistics). This is particularly useful in research, when communicating the results of experiments. In addition, patterns in the data may be [modeled](http://en.wikipedia.org/wiki/Mathematical_model) in a way that accounts for [randomness](http://en.wikipedia.org/wiki/Randomness) and uncertainty in the observations, and are then used for drawing inferences about the process or population being studied; this is called [*inferential statistics*](http://en.wikipedia.org/wiki/Inferential_statistics). Inference is a vital element of scientific advance, since it provides a means for drawing conclusions from data that are subject to random variation. To prove the propositions being investigated further, the conclusions are tested as well, as part of the [scientific method](http://en.wikipedia.org/wiki/Scientific_method). Descriptive statistics and analysis of the new data tend to provide more information as to the truth of the proposition.

Descriptive statistics and the application of inferential statistics (a.k.a., predictive statistics) together comprise *applied statistics*. *Theoretical statistics* concerns both the logical arguments underlying justification of approaches to [statistical inference](http://en.wikipedia.org/wiki/Statistical_inference), as well encompassing [*mathematical statistics*](http://en.wikipedia.org/wiki/Mathematical_statistics). Mathematical statistics includes not only the manipulation of [probability distributions](http://en.wikipedia.org/wiki/Probability_distribution) necessary for deriving results related to methods of estimation and inference, but also various aspects of [computational statistics](http://en.wikipedia.org/wiki/Computational_statistics) and the [design of experiments](http://en.wikipedia.org/wiki/Design_of_experiments).

Statistics is closely related to the [probability theory](http://en.wikipedia.org/wiki/Probability_theory), with which it is often grouped; the difference is roughly that in probability theory, one starts from the given parameters of a total population to [deduce](http://en.wikipedia.org/wiki/Deductive_reasoning) probabilities pertaining to samples, but statistical inference moves in the opposite direction, [inductive inference](http://en.wikipedia.org/wiki/Inductive_reasoning) from samples to the parameters of a larger or total population.

## History

The use of statistical methods dates back at least to the 5th century BC. The earliest writing on statistics was found in a 9th century book entitled: "Manuscript on Deciphering Cryptographic Messages", written by [Al-Kindi](http://en.wikipedia.org/wiki/Al-Kindi). In his book, he gave a detailed description of how to use statistics and [frequency analysis](http://en.wikipedia.org/wiki/Frequency_analysis) to decipher encrypted messages, this was the birth of both statistics and cryptanalysis, according to the Saudi engineer Ibrahim Al-Kadi.

The *[Nuova Cronica](http://en.wikipedia.org/wiki/Nuova_Cronica" \o "Nuova Cronica)*, a 14th century [history of Florence](http://en.wikipedia.org/wiki/History_of_Florence) by the Florentine banker and official [Giovanni Villani](http://en.wikipedia.org/wiki/Giovanni_Villani), includes much statistical information on population, ordinances, commerce and trade, education, and religious facilities and has been described as the first introduction of statistics as a positive element in history.

Some scholars pinpoint the origin of statistics to 1663, with the publication of *Natural and Political Observations upon the Bills of Mortality* by [John Graunt](http://en.wikipedia.org/wiki/John_Graunt). Early applications of statistical thinking revolved around the needs of states to base policy on demographic and economic data, hence its [*stat-* etymology](http://en.wikipedia.org/wiki/History_of_statistics#Etymology). The scope of the discipline of statistics broadened in the early 19th century to include the collection and analysis of data in general. Today, statistics is widely employed in government, business, and natural and social sciences.

Its mathematical foundations were laid in the 17th century with the development of the [probability theory](http://en.wikipedia.org/wiki/Probability_theory) by [Blaise Pascal](http://en.wikipedia.org/wiki/Blaise_Pascal" \o "Blaise Pascal) and [Pierre de Fermat](http://en.wikipedia.org/wiki/Pierre_de_Fermat). Probability theory arose from the study of games of chance. The [method of least squares](http://en.wikipedia.org/wiki/Method_of_least_squares) was first described by [Carl Friedrich Gauss](http://en.wikipedia.org/wiki/Carl_Friedrich_Gauss) around 1794. The use of modern [computers](http://en.wikipedia.org/wiki/Computer) has expedited large-scale statistical computation, and has also made possible new methods that are impractical to perform manually.

## Overview

In applying statistics to a scientific, industrial, or societal problem, it is necessary to begin with a [population](http://en.wikipedia.org/wiki/Statistical_population) or process to be studied. Populations can be diverse topics such as "all persons living in a country" or "every atom composing a crystal". A population can also be composed of observations of a process at various times, with the data from each observation serving as a different member of the overall group. Data collected about this kind of "population" constitutes what is called a [time series](http://en.wikipedia.org/wiki/Time_series).

For practical reasons, a chosen subset of the population called a [sample](http://en.wikipedia.org/wiki/Sampling_(statistics)) is studied — as opposed to compiling data about the entire group (an operation called[census](http://en.wikipedia.org/wiki/Census)). Once a sample that is representative of the population is determined, data is collected for the sample members in an observational or [experimental](http://en.wikipedia.org/wiki/Experiment) setting. This data can then be subjected to statistical analysis, serving two related purposes: description and inference.

* [Descriptive statistics](http://en.wikipedia.org/wiki/Descriptive_statistics) summarize the population data by describing what was observed in the sample numerically or graphically. Numerical descriptors include[mean](http://en.wikipedia.org/wiki/Mean) and [standard deviation](http://en.wikipedia.org/wiki/Standard_deviation) for [continuous data](http://en.wikipedia.org/wiki/Continuous_probability_distribution) types (like heights or weights), while frequency and percentage are more useful in terms of describing [categorical data](http://en.wikipedia.org/wiki/Categorical_data) (like race).
* [Inferential statistics](http://en.wikipedia.org/wiki/Inferential_statistics) uses patterns in the sample data to draw inferences about the population represented, accounting for randomness. These inferences may take the form of: answering yes/no questions about the data ([hypothesis testing](http://en.wikipedia.org/wiki/Hypothesis_testing)), estimating numerical characteristics of the data ([estimation](http://en.wikipedia.org/wiki/Estimation_theory)), describing [associations](http://en.wikipedia.org/wiki/Association_(statistics)" \o "Association (statistics))within the data ([correlation](http://en.wikipedia.org/wiki/Correlation_and_dependence)) and modeling relationships within the data (for example, using [regression analysis](http://en.wikipedia.org/wiki/Regression_analysis)). Inference can extend to [forecasting](http://en.wikipedia.org/wiki/Forecasting), [prediction](http://en.wikipedia.org/wiki/Prediction) and estimation of unobserved values either in or associated with the population being studied; it can include [extrapolation](http://en.wikipedia.org/wiki/Extrapolation) and [interpolation](http://en.wikipedia.org/wiki/Interpolation) of time series or [spatial data](http://en.wikipedia.org/wiki/Spatial_data_analysis), and can also include [data mining](http://en.wikipedia.org/wiki/Data_mining).

"... it is only the manipulation of uncertainty that interests us. We are not concerned with the matter that is uncertain. Thus we do not study the mechanism of rain; only whether it will rain."

The concept of [correlation](http://en.wikipedia.org/wiki/Correlation) is particularly noteworthy for the potential confusion it can cause. Statistical analysis of a [data set](http://en.wikipedia.org/wiki/Data_set) often reveals that two variables (properties) of the population under consideration tend to vary together, as if they were connected. For example, a study of annual income that also looks at age of death might find that poor people tend to have shorter lives than affluent people. The two variables are said to be correlated; however, they may or may not be the cause of one another. The correlation phenomena could be caused by a third, previously unconsidered phenomenon, called a lurking variable or [confounding variable](http://en.wikipedia.org/wiki/Confounding_variable). For this reason, there is no way to immediately infer the existence of a causal relationship between the two variables.

For a sample to be used as a guide to an entire population, it is important that it is truly a representative of that overall population. Representative sampling assures that the inferences and conclusions can be safely extended from the sample to the population as a whole. A major problem lies in determining the extent to which the sample chosen is actually representative. Statistics offers methods to estimate and correct for any random trending within the sample and data collection procedures. There are also methods of [experimental design](http://en.wikipedia.org/wiki/Experimental_design) for experiments that can lessen these issues at the outset of a study, strengthening its capability to discern truths about the population.

Randomness is studied using the [mathematical discipline](http://en.wikipedia.org/wiki/Mathematics) of [probability theory](http://en.wikipedia.org/wiki/Probability_theory). Probability is used in "[mathematical statistics](http://en.wikipedia.org/wiki/Mathematical_statistics)" (alternatively, "[statistical theory](http://en.wikipedia.org/wiki/Statistical_theory)") to study the [sampling distributions](http://en.wikipedia.org/wiki/Sampling_distribution) of [sample statistics](http://en.wikipedia.org/wiki/Sample_statistic) and, more generally, the properties of [statistical procedures](http://en.wikipedia.org/wiki/Statistical_decision_theory). The use of any statistical method is valid when the system or population under consideration satisfies the assumptions of the method.

[Misuse of statistics](http://en.wikipedia.org/wiki/Misuse_of_statistics) can produce subtle, but serious errors in description and interpretation — subtle in the sense that even experienced professionals make such errors, and serious in the sense that they can lead to devastating decision errors. For instance, social policy, medical practice, and the reliability of structures like bridges all rely on the proper use of statistics.

Even when statistical techniques are correctly applied, the results can be difficult to interpret for those lacking expertise. The [statistical significance](http://en.wikipedia.org/wiki/Statistical_significance) of a trend in the data — which measures the extent to which a trend could be caused by random variation in the sample — may or may not agree with an intuitive sense of its significance. The set of basic statistical skills (and skepticism) that people need to deal with information in their everyday lives properly is referred to as [statistical literacy](http://en.wikipedia.org/wiki/Statistical_literacy).

## Statistical methods

### Experimental and observational studies

A common goal for a statistical research project is to investigate [causality](http://en.wikipedia.org/wiki/Causality), and in particular to draw a conclusion on the effect of changes in the values of predictors or independent on [dependent variables](http://en.wikipedia.org/wiki/Dependent_variable) or response. There are two major types of causal statistical studies: [experimental studies](http://en.wikipedia.org/wiki/Controlled_experiment) and [observational studies](http://en.wikipedia.org/wiki/Observational_study). 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 lies 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 has 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 response are investigated.

#### Experiments

The basic steps of a statistical experiment are:

1. Planning the research, including finding the number of replicates of the study, using the following information: preliminary estimates regarding the size of treatment, [alternative hypotheses](http://en.wikipedia.org/wiki/Alternative_hypothesis), and the estimated [experimental variability](http://en.wikipedia.org/wiki/Experimental_error). Consideration of the selection of experimental subjects and the ethics of research is necessary. Statisticians recommend that experiments compare (at least) one new treatment with a standard treatment or control, to allow an unbiased estimate of the difference in treatment effects.
2. [Design of experiments](http://en.wikipedia.org/wiki/Design_of_experiments), using [blocking](http://en.wikipedia.org/wiki/Blocking_(statistics)) to reduce the influence of [confounding variables](http://en.wikipedia.org/wiki/Confounding_variable), and [randomized assignment](http://en.wikipedia.org/wiki/Randomized_assignment) of treatments to subjects to allow unbiased of treatment effects and experimental error. At this stage, the experimenters and statisticians write the [*experimental protocol*](http://en.wikipedia.org/wiki/Protocol_(natural_sciences)) that shall guide the performance of the experiment and that specifies the *primary analysis* of the experimental data.
3. Performing the experiment following the [experimental protocol](http://en.wikipedia.org/wiki/Protocol_(natural_sciences)) and [analyzing the data](http://en.wikipedia.org/wiki/Analysis_of_variance) following the experimental protocol.
4. Further examining the data set in secondary analyses, to suggest new hypotheses for future study.
5. Documenting and presenting the results of the study.

Experiments on human behavior have special concerns. The famous [Hawthorne study](http://en.wikipedia.org/wiki/Hawthorne_study) examined changes to the working environment at the Hawthorne plant of the Western. The researchers were interested in determining whether increased illumination would increase the productivity of the line workers. The researchers first measured the productivity in the plant, then modified the illumination in an area of the plant and checked if the changes in illumination affected productivity. It turned out that productivity indeed improved (under the experimental conditions). However, the study is heavily criticized today for errors in experimental procedures, specifically for the lack of a [control group](http://en.wikipedia.org/wiki/Control_group) and [blindness](http://en.wikipedia.org/wiki/Double-blind). The [Hawthorne effect](http://en.wikipedia.org/wiki/Hawthorne_effect) refers to finding that an outcome (in this case, worker productivity) changed due to observation itself. Those in the Hawthorne study became more productive not because the lighting was changed but because they were being observed.

#### Observational study

An example of an observational study is one that 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 performs statistical analysis. In this case, the researchers would collect observations of both smokers and non-smokers, perhaps through a [case-control study](http://en.wikipedia.org/wiki/Case-control_study), and then look for the number of cases of lung cancer in each group.

### Levels of measurement

There are four main [levels of measurement](http://en.wikipedia.org/wiki/Level_of_measurement) used in statistics: nominal, ordinal, interval, and ratio. Each of these has different degrees of usefulness in statistical research. Ratio measurements have both a meaningful zero value and the distances between different measurements defined; they provide the greatest flexibility in statistical methods that can be used for analyzing the data. Interval measurements have meaningful distances between measurements defined, but the zero value is arbitrary (as in the case with [longitude](http://en.wikipedia.org/wiki/Longitude) and temperature measurements in [Celsius](http://en.wikipedia.org/wiki/Celsius) or [Fahrenheit](http://en.wikipedia.org/wiki/Fahrenheit)). Ordinal measurements have imprecise differences between consecutive values, but have a meaningful order to those values. Nominal measurements have no meaningful rank order among values.

Because variables conforming only to nominal or ordinal measurements cannot be reasonably measured numerically, sometimes they are grouped together as categorical, whereas ratio and interval measurements are grouped together as [quantitative variables](http://en.wikipedia.org/wiki/Variable_(mathematics)#Applied_statistics), which can be either [discrete](http://en.wikipedia.org/wiki/Probability_distribution#Discrete_probability_distribution) or [continuous](http://en.wikipedia.org/wiki/Probability_distribution#Continuous_probability_distribution), due to their numerical nature.

### Key terms used in statistics

#### Null hypothesis

Interpretation of statistical information can often involve the development of a [null hypothesis](http://en.wikipedia.org/wiki/Null_hypothesis) in that the assumption is that whatever is proposed as a cause has no effect on the variable being measured.

The best illustration for a novice is the predicament encountered by a jury trial. The null hypothesis, H0, asserts that the defendant is innocent, whereas the alternative hypothesis, H1, asserts that the defendant is guilty. The indictment comes because of suspicion of the guilt. The H0 (status quo) stands in opposition to H1 and is maintained unless H1 is supported by evidence"beyond a reasonable doubt". However,"failure to reject H0" in this case does not imply innocence, but merely that the evidence was insufficient to convict. So the jury does not necessarily *accept* H0 but *fails to reject* H0. While one can not "prove" a null hypothesis one can test how close it is to being true with a [power test](http://en.wikipedia.org/wiki/Statistical_power), which tests for type II errors.

#### Error

Working from a [null hypothesis](http://en.wikipedia.org/wiki/Null_hypothesis) two basic forms of error are recognized:

* [Type I errors](http://en.wikipedia.org/wiki/Type_I_and_type_II_errors#Type_I_error) where the null hypothesis is falsely rejected giving a "false positive".
* [Type II errors](http://en.wikipedia.org/wiki/Type_I_and_type_II_errors#Type_II_error) where the null hypothesis fails to be rejected and an actual difference between populations is missed giving a "false negative".

Error also refers to the extent to which individual observations in a sample differ from a central value, such as the sample or population mean. Many statistical methods seek to minimize the mean-squared error, and these are called "[methods of least squares](http://en.wikipedia.org/wiki/Least_squares)."

Measurement processes that generate statistical data are also subject to error. Many of these errors are classified as [random](http://en.wikipedia.org/wiki/Random_error) (noise) or [systematic](http://en.wikipedia.org/wiki/Systematic_error) ([bias](http://en.wikipedia.org/wiki/Bias)), but other important types of errors (e.g., blunder, such as when an analyst reports incorrect units) can also be important.

#### Interval estimation

Most studies will only sample part of a population and so the results are not fully representative of the whole population. Any estimates obtained from the sample only approximate the population value. [Confidence intervals](http://en.wikipedia.org/wiki/Confidence_intervals) allow statisticians to express how closely the sample estimate matches the true value in the whole population. Often they are expressed as 95% confidence intervals. Formally, a 95% confidence interval for a value is a range where, if the sampling and analysis were repeated under the same conditions (yielding a different dataset), the interval would include the true (population) value 95% of the time. This does *not* imply that the probability that the true value is in the confidence interval is 95%. From the [frequentist](http://en.wikipedia.org/wiki/Frequentist_inference" \o "Frequentist inference) perspective, such a claim does not even make sense, as the true value is not a [random variable](http://en.wikipedia.org/wiki/Random_variable). Either the true value is or is not within the given interval. However, it is true that, before any data are sampled and given a plan for how the confidence interval will be constructed, the probability is 95% that the yet-to-be-calculated interval will cover the true value: at this point, the limits of the interval are yet-to-be-observed[random variables](http://en.wikipedia.org/wiki/Random_variable). One approach that does yield an interval that can be interpreted as having a given probability of containing the true value is to use a [credible interval](http://en.wikipedia.org/wiki/Credible_interval)from [Bayesian statistics](http://en.wikipedia.org/wiki/Bayesian_statistics): this approach depends on a different way of [interpreting what is meant by "probability"](http://en.wikipedia.org/wiki/Probability_interpretations), that is as a [Bayesian probability](http://en.wikipedia.org/wiki/Bayesian_probability).

#### Significance

Statistics rarely give a simple Yes/No type answer to the question asked of them. Interpretation often comes down to the level of statistical significance applied to the numbers and often refers to the probability of a value accurately rejecting the null hypothesis (sometimes referred to as the [p-value](http://en.wikipedia.org/wiki/P-value)).

Referring to statistical significance does not necessarily mean that the overall result is significant in real world terms. For example, in a large study of a drug it may be shown that the drug has a statistically significant but very small beneficial effect, such that the drug will be unlikely to help the patient in a noticeable way.

Criticisms arise because the hypothesis testing approach forces one hypothesis (the [null hypothesis](http://en.wikipedia.org/wiki/Null_hypothesis)) to be "favored," 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, but it is possible to properly formulate tests in account for this.

One response involves going beyond reporting only the [significance level](http://en.wikipedia.org/wiki/Significance_level) to include the [*p*-value](http://en.wikipedia.org/wiki/P-value) when reporting whether a hypothesis is rejected or accepted. The *p*-value, however, does not indicate the size of the effect. A better and increasingly common approach is to report [confidence intervals](http://en.wikipedia.org/wiki/Confidence_interval). Although these are produced from the same calculations as those of hypothesis tests or *p*-values, they describe both the size of the effect and the uncertainty surrounding it.

### Examples

Some well-known statistical [tests](http://en.wikipedia.org/wiki/Statistical_hypothesis_testing) and [procedures](http://en.wikipedia.org/wiki/Procedure_(term)) are:

* [Analysis of variance](http://en.wikipedia.org/wiki/Analysis_of_variance) (ANOVA)
* [Chi-squared test](http://en.wikipedia.org/wiki/Chi-squared_test)
* [Correlation](http://en.wikipedia.org/wiki/Correlation)
* [Factor analysis](http://en.wikipedia.org/wiki/Factor_analysis)
* [Mann–Whitney U](http://en.wikipedia.org/wiki/Mann%E2%80%93Whitney_U)
* [Mean square weighted deviation](http://en.wikipedia.org/wiki/Mean_square_weighted_deviation) (MSWD)
* [Pearson product-moment correlation coefficient](http://en.wikipedia.org/wiki/Pearson_product-moment_correlation_coefficient)
* [Regression analysis](http://en.wikipedia.org/wiki/Regression_analysis)
* [Spearman's rank correlation coefficient](http://en.wikipedia.org/wiki/Spearman%27s_rank_correlation_coefficient)
* [Student's t-test](http://en.wikipedia.org/wiki/Student%27s_t-test)
* [Time series analysis](http://en.wikipedia.org/wiki/Time_series_analysis)

## Specialized disciplines

*Main article:*[*List of fields of application of statistics*](http://en.wikipedia.org/wiki/List_of_fields_of_application_of_statistics)

Statistical techniques are used in a wide range of types of scientific and social research, including: [biostatistics](http://en.wikipedia.org/wiki/Biostatistics), [computational biology](http://en.wikipedia.org/wiki/Computational_biology), [computational sociology](http://en.wikipedia.org/wiki/Computational_sociology),[network biology](http://en.wikipedia.org/wiki/Network_biology), [social science](http://en.wikipedia.org/wiki/Social_science), [sociology](http://en.wikipedia.org/wiki/Sociology) and [social research](http://en.wikipedia.org/wiki/Social_research). Some fields of inquiry use applied statistics so extensively that they have [specialized terminology](http://en.wikipedia.org/wiki/Specialized_terminology). These disciplines include:

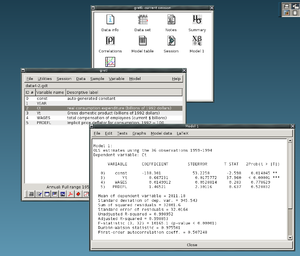
* [Actuarial science](http://en.wikipedia.org/wiki/Actuarial_science)
* [Applied information economics](http://en.wikipedia.org/wiki/Applied_information_economics)
* [Biostatistics](http://en.wikipedia.org/wiki/Biostatistics)
* [Business statistics](http://en.wikipedia.org/wiki/Business_statistics)
* [Chemo metrics](http://en.wikipedia.org/wiki/Chemometrics) (for analysis of data from [chemistry](http://en.wikipedia.org/wiki/Chemistry))
* [Data mining](http://en.wikipedia.org/wiki/Data_mining) (applying statistics and [pattern recognition](http://en.wikipedia.org/wiki/Pattern_recognition) to discover knowledge from data)
* [Demography](http://en.wikipedia.org/wiki/Demography)
* [Econometrics](http://en.wikipedia.org/wiki/Econometrics)
* [Energy statistics](http://en.wikipedia.org/wiki/Energy_statistics)
* [Engineering statistics](http://en.wikipedia.org/wiki/Engineering_statistics)
* [Epidemiology](http://en.wikipedia.org/wiki/Epidemiology)
* [Geography](http://en.wikipedia.org/wiki/Geography) and [Geographic Information Systems](http://en.wikipedia.org/wiki/Geographic_Information_Systems), specifically in [Spatial analysis](http://en.wikipedia.org/wiki/Spatial_analysis)
* [Image processing](http://en.wikipedia.org/wiki/Image_processing)
* [Psychological statistics](http://en.wikipedia.org/wiki/Psychological_statistics)
* [Reliability engineering](http://en.wikipedia.org/wiki/Reliability_engineering)
* [Social statistics](http://en.wikipedia.org/wiki/Social_statistics)

In addition, there are particular types of statistical analysis that have also developed their own specialised terminology and methodology:

* [Bootstrap](http://en.wikipedia.org/wiki/Bootstrapping_(statistics)) & [Jackknife Resampling](http://en.wikipedia.org/wiki/Resampling_(statistics))
* [Multivariate statistics](http://en.wikipedia.org/wiki/Multivariate_statistics)
* [Statistical classification](http://en.wikipedia.org/wiki/Statistical_classification)
* [Statistical surveys](http://en.wikipedia.org/wiki/Statistical_survey)
* [Structured data analysis (statistics)](http://en.wikipedia.org/wiki/Structured_data_analysis_(statistics))
* [Structural equation modeling](http://en.wikipedia.org/wiki/Structural_equation_modelling)
* [Survival analysis](http://en.wikipedia.org/wiki/Survival_analysis)
* Statistics in various sports, particularly [baseball](http://en.wikipedia.org/wiki/Baseball_statistics) and [cricket](http://en.wikipedia.org/wiki/Cricket_statistics)

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](http://en.wikipedia.org/wiki/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

[](http://en.wikipedia.org/wiki/File:Gretl_screenshot.png)

[http://bits.wikimedia.org/static-1.21wmf7/skins/common/images/magnify-clip.png](http://en.wikipedia.org/wiki/File:Gretl_screenshot.png)

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](http://en.wikipedia.org/wiki/Linear_model), but powerful computers, coupled with suitable numerical[algorithms](http://en.wikipedia.org/wiki/Algorithms), caused an increased interest in [nonlinear models](http://en.wikipedia.org/wiki/Nonlinear_regression) (such as [neural networks](http://en.wikipedia.org/wiki/Neural_networks)) as well as the creation of new types, such as [generalized linear models](http://en.wikipedia.org/wiki/Generalized_linear_model) and [multilevel models](http://en.wikipedia.org/wiki/Multilevel_model).

Increased computing power has also led to the growing popularity of computationally intensive methods based on [resampling](http://en.wikipedia.org/wiki/Resampling_(statistics)), such as permutation tests and the [bootstrap](http://en.wikipedia.org/wiki/Bootstrapping_(statistics)), while techniques such as [Gibbs sampling](http://en.wikipedia.org/wiki/Gibbs_sampling) have made use of Bayesian models more feasible. The computer revolution has implications for the future of statistics with new emphasis on "experimental" and "empirical" statistics. A large number of both general and special purpose [statistical software](http://en.wikipedia.org/wiki/List_of_statistical_packages) are now available.

## Misuse

There is a general perception that statistical knowledge is all-too-frequently intentionally [misused](http://en.wikipedia.org/wiki/Misuse_of_statistics) by finding ways to interpret only the data that are favorable to the presenter.[[17]](http://en.wikipedia.org/wiki/Statistics#cite_note-Huff-17) A mistrust and misunderstanding of statistics is associated with the quotation, "[There are three kinds of lies: lies, damned lies, and statistics](http://en.wikipedia.org/wiki/Lies,_damned_lies,_and_statistics)". Misuse of statistics can be both inadvertent and intentional, and the book [*How to Lie With Statistics*](http://en.wikipedia.org/wiki/How_to_Lie_With_Statistics)[[17]](http://en.wikipedia.org/wiki/Statistics#cite_note-Huff-17) outlines a range of considerations. In an attempt to shed light on the use and misuse of statistics, reviews of statistical techniques used in particular fields are conducted (e.g. Warne, Lazo, Ramos, and Ritter (2012)).[[18]](http://en.wikipedia.org/wiki/Statistics#cite_note-18)

## Ways to avoid misuse of statistics include using proper diagrams and avoiding bias

 "The misuse occurs when such conclusions are held to be representative of the universe by those who either deliberately or unconsciously overlook the sampling bias.  Bar graphs are arguably the easiest diagrams to use and understand, and they can be made either with simple computer programs or hand drawn. Unfortunately, most people do not look for bias or errors, so they do not see them. Thus, we believe something to be truth that is not well-represented. In order to make data gathered from statistics believable and accurate, the sample taken must be representative of the whole. As Huff's book states, “The dependability of a sample can be destroyed by [bias]… allow yourself some degree of skepticism."

## Statistics applied to mathematics or the arts

Traditionally, statistics was concerned with drawing inferences using a semi-standardized methodology that was "required learning" in most sciences. This has changed with use of statistics in non-inferential contexts. What was once considered a dry subject, taken in many fields as a degree-requirement, is now viewed enthusiastically. Initially derided by some mathematical purists, it is now considered essential methodology in certain areas.

* In [number theory](http://en.wikipedia.org/wiki/Number_theory), [scatter plots](http://en.wikipedia.org/wiki/Scatter_plot) of data generated by a distribution function may be transformed with familiar tools used in statistics to reveal underlying patterns, which may then lead to hypotheses.
* Methods of statistics including predictive methods in [forecasting](http://en.wikipedia.org/wiki/Forecasting), are combined with [chaos theory](http://en.wikipedia.org/wiki/Chaos_theory) and [fractal geometry](http://en.wikipedia.org/wiki/Fractal_geometry) to create video works that are considered to have great beauty.
* The [process art](http://en.wikipedia.org/wiki/Process_art) of [Jackson Pollock](http://en.wikipedia.org/wiki/Jackson_Pollock) relied on artistic experiments whereby underlying distributions in nature were artistically revealed. With the advent of computers, methods of statistics were applied to formalize such distribution driven natural processes, in order to make and analyze moving video art.
* Methods of statistics may be used predicatively in [performance art](http://en.wikipedia.org/wiki/Performance_art), as in a card trick based on a [Markov process](http://en.wikipedia.org/wiki/Markov_process) that only works some of the time, the occasion of which can be predicted using statistical methodology.
* Statistics can be used to predicatively create art, as in the statistical or [stochastic music](http://en.wikipedia.org/wiki/Stochastic_music) invented by [Iannis Xenakis](http://en.wikipedia.org/wiki/Iannis_Xenakis), where the music is performance-specific. Though this type of artistry does not always come out as expected, it does behave in ways that are predictable and tunable using statistics.