statistics

statistics, the science of collecting, analyzing, presenting, and interpreting data. Governmental needs for census data as well as information about a variety of economic activities provided much of the early impetus for the field of statistics. Currently the need to turn the large amounts of data available in many applied fields into useful information has stimulated both theoretical and practical developments in statistics.

TABLE OF CONTENTS

Data are the facts and figures that are collected, analyzed, and summarized for presentation and interpretation. Data may be classified as either quantitative or qualitative. Quantitative data measure either how much or how many of something, and qualitative data provide labels, or names, for categories of like items. For example, suppose that a particular study is interested in characteristics such as age, gender, marital status, and annual income for a sample of 100 individuals. These characteristics would be called the variables of the study, and data values for each of the variables would be associated with each individual. Thus, the data values of 28, male, single, and $30,000 would be recorded for a 28-year-old single male with an annual income of $30,000. With 100 individuals and 4 variables, the data set would have $100 \times 4 = 400$ items. In this example, age and annual income are quantitative variables; the corresponding data values indicate how many years and how much money for each individual. Gender and marital

status are qualitative variables. The labels male and female provide the qualitative data for gender, and the labels single, married, divorced, and widowed indicate marital status.

Sample survey methods are used to collect data from observational studies, and experimental design methods are used to collect data from experimental studies. The area of descriptive statistics is concerned primarily with methods of presenting and interpreting data using graphs, tables, and numerical summaries. Whenever statisticians use data from a sample—i.e., a subset of the population—to make statements about a population, they are performing statistical inference. Estimation and hypothesis testing are procedures used to make statistical inferences. Fields such as health care, biology, chemistry, physics, education, engineering, business, and economics make extensive use of statistical inference.

Methods of probability were developed initially for the analysis of gambling games. Probability plays a key role in statistical inference; it is used to provide measures of the quality and precision of the inferences. Many of the methods of statistical inference are described in this article. Some of these methods are used primarily for single-variable studies, while others, such as regression and correlation analysis, are used to make inferences about relationships among two or more variables.

Descriptive statistics

Descriptive statistics are tabular, graphical, and numerical summaries of data. The purpose of descriptive statistics is to facilitate the presentation and interpretation of data. Most of the statistical presentations appearing in newspapers and magazines are descriptive in nature. Univariate methods of descriptive statistics use data to enhance the understanding of

a single variable; multivariate methods focus on using statistics to understand the relationships among two or more variables. To illustrate methods of descriptive statistics, the previous example in which data were collected on the age, gender, marital status, and annual income of 100 individuals will be examined.

Tabular methods

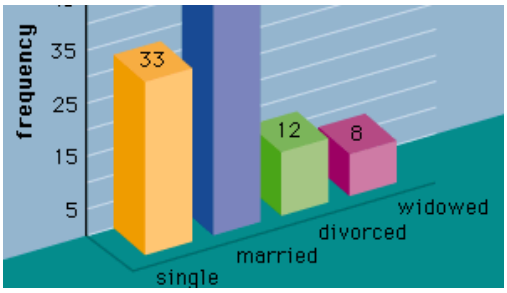
The most commonly used tabular summary of data for a single variable is a frequency distribution. A frequency distribution shows the number of data values in each of several nonoverlapping classes. Another tabular summary, called a relative frequency distribution, shows the fraction, or percentage, of data values in each class. The most common tabular summary of data for two variables is a cross tabulation, a two-variable analogue of a frequency distribution.

For a qualitative variable, a frequency distribution shows the number of data values in each qualitative category. For instance, the variable gender has two categories: male and female. Thus, a frequency distribution for gender would have two nonoverlapping classes to show the number of males and females. A relative frequency distribution for this variable would show the fraction of individuals that are male and the fraction of individuals that are female.

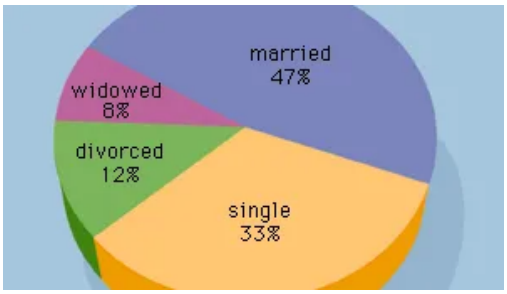
Constructing a frequency distribution for a quantitative variable requires more care in defining the classes and the division points between adjacent classes. For instance, if the age data of the example above ranged from 22 to 78 years, the following six nonoverlapping classes could be used: 20–29, 30–39, 40–49, 50–59, 60–69, and 70–79. A frequency distribution would show the number of data values in each of

these classes, and a relative frequency distribution would show the fraction of data values in each.

A cross tabulation is a two-way table with the rows of the table representing the classes of one variable and the columns of the table representing the classes of another variable. To construct a cross tabulation using the variables gender and age, gender could be shown with two rows, male and female, and age could be shown with six columns corresponding to the age classes 20–29, 30–39, 40–49, 50–59, 60–69, and 70–79. The entry in each cell of the table would specify the number of data values with the gender given by the row heading and the age given by the column heading. Such a cross tabulation could be helpful in understanding the relationship between gender and age.



bar graph



pie chart

Graphical methods

A number of graphical methods are available for describing data. A bar graph is a graphical device for depicting qualitative data that have been summarized in a frequency distribution. Labels for the categories of the qualitative variable are shown on the horizontal axis of the graph. A bar above each label is constructed such that the height of each bar is proportional to the number of data values in the category. A bar graph of the marital status for the 100 individuals in the above example is shown in Figure 1. There are 4 bars in the graph, one for each class. A pie chart is another graphical device for summarizing qualitative data. The size of each slice of the pie is proportional to the number of data values in the

corresponding class. A pie chart for the marital status of the 100 individuals is shown in Figure 2.

A histogram is the most common graphical presentation of quantitative data that have been summarized in a frequency distribution. The values of the quantitative variable are shown on the horizontal axis. A rectangle is drawn above each class such that the base of the rectangle is equal to the width of the class interval and its height is proportional to the number of data values in the class.

Numerical measures

A variety of numerical measures are used to summarize data. The proportion, or percentage, of data values in each category is the primary numerical measure for qualitative data. The mean, median, mode, percentiles, range, variance, and standard deviation are the most commonly used numerical measures for quantitative data. The mean, often called the average, is computed by adding all the data values for a variable and dividing the sum by the number of data values. The mean is a measure of the central location for the data. The median is another measure of central location that, unlike the mean, is not affected by extremely large or extremely small data values. When determining the median, the data values are first ranked in order from the smallest value to the largest value. If there is an odd number of data values, the median is the middle value; if there is an even number of data values, the median is the average of the two middle values. The third measure of central tendency is the mode, the data value that occurs with greatest frequency.

Percentiles provide an indication of how the data values are spread over the interval from the smallest value to the largest value. Approximately *p*

percent of the data values fall below the *p*th percentile, and roughly 100 − *p* percent of the data values are above the *p*th percentile. Percentiles are reported, for example, on most standardized tests. Quartiles divide the data values into four parts; the first quartile is the 25th percentile, the second quartile is the 50th percentile (also the median), and the third quartile is the 75th percentile.

The range, the difference between the largest value and the smallest value, is the simplest measure of variability in the data. The range is determined by only the two extreme data values. The variance (s^2) and the standard deviation (s), on the other hand, are measures of variability that are based on all the data and are more commonly used. Equation 1 shows the formula for computing the variance of a sample consisting of n items. In applying equation 1, the deviation (difference) of each data value from the sample mean is computed and squared. The squared deviations are then summed and divided by $n - 1$ to provide the sample variance.

$$s^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n - 1} \qquad (1)$$

The standard deviation is the square root of the variance. Because the unit of measure for the standard deviation is the same as the unit of measure for the data, many individuals prefer to use the standard deviation as the descriptive measure of variability.

Outliers

Sometimes data for a variable will include one or more values that appear unusually large or small and out of place when compared with the other data values. These values are known as outliers and often have been erroneously included in the data set. Experienced statisticians take steps to identify outliers and then review each one carefully for accuracy

and the appropriateness of its inclusion in the data set. If an error has been made, corrective action, such as rejecting the data value in question, can be taken. The mean and standard deviation are used to identify outliers. A z-score can be computed for each data value. With x representing the data value, $\bar{x}$ the sample mean, and s the sample standard deviation, the z-score is given by $z = (x - \bar{x})/s$. The z-score represents the relative position of the data value by indicating the number of standard deviations it is from the mean. A rule of thumb is that any value with a z-score less than −3 or greater than +3 should be considered an outlier.

Exploratory data analysis

Exploratory data analysis provides a variety of tools for quickly summarizing and gaining insight about a set of data. Two such methods are the five-number summary and the box plot. A five-number summary simply consists of the smallest data value, the first quartile, the median, the third quartile, and the largest data value. A box plot is a graphical device based on a five-number summary. A rectangle (i.e., the box) is drawn with the ends of the rectangle located at the first and third quartiles. The rectangle represents the middle 50 percent of the data. A vertical line is drawn in the rectangle to locate the median. Finally lines, called whiskers, extend from one end of the rectangle to the smallest data value and from the other end of the rectangle to the largest data value. If outliers are present, the whiskers generally extend only to the smallest and largest data values that are not outliers. Dots, or asterisks, are then placed outside the whiskers to denote the presence of outliers.

Probability

Probability is a subject that deals with uncertainty. In everyday terminology, probability can be thought of as a numerical measure of the likelihood that a particular event will occur. Probability values are assigned on a scale from 0 to 1, with values near 0 indicating that an event is unlikely to occur and those near 1 indicating that an event is likely to take place. A probability of 0.50 means that an event is equally likely to occur as not to occur.

Events and their probabilities

Oftentimes probabilities need to be computed for related events. For instance, advertisements are developed for the purpose of increasing sales of a product. If seeing the advertisement increases the probability of a person buying the product, the events “seeing the advertisement” and “buying the product” are said to be dependent. If two events are independent, the occurrence of one event does not affect the probability of the other event taking place. When two or more events are independent, the probability of their joint occurrence is the product of their individual probabilities. Two events are said to be mutually exclusive if the occurrence of one event means that the other event cannot occur; in this case, when one event takes place, the probability of the other event occurring is zero.

Random variables and probability distributions

A random variable is a numerical description of the outcome of a statistical experiment. A random variable that may assume only a finite number or an infinite sequence of values is said to be discrete; one that may assume any value in some interval on the real number line is said to be continuous. For instance, a random variable representing the number of automobiles sold at a particular dealership on one day would be

discrete, while a random variable representing the weight of a person in kilograms (or pounds) would be continuous.

The probability distribution for a random variable describes how the probabilities are distributed over the values of the random variable. For a discrete random variable, x, the probability distribution is defined by a probability mass function, denoted by $f(x)$. This function provides the probability for each value of the random variable. In the development of the probability function for a discrete random variable, two conditions must be satisfied: (1) $f(x)$ must be nonnegative for each value of the random variable, and (2) the sum of the probabilities for each value of the random variable must equal one.

A continuous random variable may assume any value in an interval on the real number line or in a collection of intervals. Since there is an infinite number of values in any interval, it is not meaningful to talk about the probability that the random variable will take on a specific value; instead, the probability that a continuous random variable will lie within a given interval is considered.

In the continuous case, the counterpart of the probability mass function is the probability density function, also denoted by $f(x)$. For a continuous random variable, the probability density function provides the height or value of the function at any particular value of x; it does not directly give the probability of the random variable taking on a specific value. However, the area under the graph of $f(x)$ corresponding to some interval, obtained by computing the integral of $f(x)$ over that interval, provides the probability that the variable will take on a value within that interval. A probability density function must satisfy two requirements:

(1) $f(x)$ must be nonnegative for each value of the random variable, and (2) the integral over all values of the random variable must equal one.

The expected value, or mean, of a random variable—denoted by $E(x)$ or μ—is a weighted average of the values the random variable may assume. In the discrete case the weights are given by the probability mass function, and in the continuous case the weights are given by the probability density function. The formulas for computing the expected values of discrete and continuous random variables are given by equations 2 and 3, respectively.

$$E(x) = \Sigma x f(x) \quad (2)$$

$$E(x) = \int x f(x) dx \quad (3)$$

The variance of a random variable, denoted by $\mathrm{Var}(x)$ or σ^2, is a weighted average of the squared deviations from the mean. In the discrete case the weights are given by the probability mass function, and in the continuous case the weights are given by the probability density function. The formulas for computing the variances of discrete and continuous random variables are given by equations 4 and 5, respectively. The standard deviation, denoted σ, is the positive square root of the variance. Since the standard deviation is measured in the same units as the random variable and the variance is measured in squared units, the standard deviation is often the preferred measure.

$$\mathrm{Var}(x) = \sigma^2 = \Sigma (x - \mu)^2 f(x) \quad (4)$$

$$\mathrm{Var}(x) = \sigma^2 = \int (x - \mu)^2 f(x) dx \quad (5)$$

Special probability distributions

The binomial distribution

Two of the most widely used discrete probability distributions are the binomial and Poisson. The binomial probability mass function (equation 6) provides the probability that x successes will occur in n trials of a binomial experiment. $f(x) = \binom{n}{x} p^x (1-p)^{(n-x)}$ (6)

A binomial experiment has four properties: (1) it consists of a sequence of n identical trials; (2) two outcomes, success or failure, are possible on each trial; (3) the probability of success on any trial, denoted p, does not change from trial to trial; and (4) the trials are independent. For instance, suppose that it is known that 10 percent of the owners of two-year old automobiles have had problems with their automobile's electrical system. To compute the probability of finding exactly 2 owners that have had electrical system problems out of a group of 10 owners, the binomial probability mass function can be used by setting $n = 10$, $x = 2$, and $p = 0.1$ in equation 6; for this case, the probability is 0.1937.

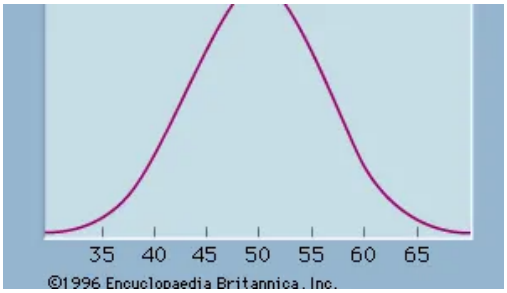
The Poisson distribution

The Poisson probability distribution is often used as a model of the number of arrivals at a facility within a given period of time. For instance, a random variable might be defined as the number of telephone calls coming into an airline reservation system during a period of 15 minutes. If the mean number of arrivals during a 15-minute interval is known, the Poisson probability mass function given by equation 7 can be used to compute the probability of x arrivals. $f(x) = \frac{\mu^x e^{-\mu}}{x!}$ (7)

For example, suppose that the mean number of calls arriving in a 15-minute period is 10. To compute the probability that 5 calls come in

within the next 15 minutes, μ = 10 and x = 5 are substituted in equation 7, giving a probability of 0.0378.

The normal distribution



normal probability distribution

The most widely used continuous probability distribution in statistics is the normal probability distribution. The graph corresponding to a normal probability density function with a mean of μ = 50 and a standard deviation of σ = 5 is shown in Figure 3. Like all normal distribution graphs, it is a bell-shaped curve. Probabilities for the normal probability distribution can be computed using statistical tables for the standard normal probability distribution, which is a normal probability distribution with a mean of zero and a standard deviation of one. A simple mathematical formula is used to convert any value from a normal probability distribution with mean μ and a standard deviation σ into a corresponding value for a standard normal distribution. The tables for the standard normal distribution are then used to compute the appropriate probabilities.

There are many other discrete and continuous probability distributions. Other widely used discrete distributions include the geometric, the hypergeometric, and the negative binomial; other commonly used continuous distributions include the uniform, exponential, gamma, chi-square, beta, *t*, and F.

Estimation

It is often of interest to learn about the characteristics of a large group of elements such as individuals, households, buildings, products, parts, customers, and so on. All the elements of interest in a particular study

form the population. Because of time, cost, and other considerations, data often cannot be collected from every element of the population. In such cases, a subset of the population, called a sample, is used to provide the data. Data from the sample are then used to develop estimates of the characteristics of the larger population. The process of using a sample to make inferences about a population is called statistical inference.

Characteristics such as the population mean, the population variance, and the population proportion are called parameters of the population. Characteristics of the sample such as the sample mean, the sample variance, and the sample proportion are called sample statistics. There are two types of estimates: point and interval. A point estimate is a value of a sample statistic that is used as a single estimate of a population parameter. No statements are made about the quality or precision of a point estimate. Statisticians prefer interval estimates because interval estimates are accompanied by a statement concerning the degree of confidence that the interval contains the population parameter being estimated. Interval estimates of population parameters are called confidence intervals.

Sampling and sampling distributions

Although sample survey methods will be discussed in more detail below in the section Sample survey methods, it should be noted here that the methods of statistical inference, and estimation in particular, are based on the notion that a probability sample has been taken. The key characteristic of a probability sample is that each element in the population has a known probability of being included in the sample. The most fundamental type is a simple random sample.

For a population of size N, a simple random sample is a sample selected such that each possible sample of size n has the same probability of being selected. Choosing the elements from the population one at a time so that each element has the same probability of being selected will provide a simple random sample. Tables of random numbers, or computer-generated random numbers, can be used to guarantee that each element has the same probability of being selected.

A sampling distribution is a probability distribution for a sample statistic. Knowledge of the sampling distribution is necessary for the construction of an interval estimate for a population parameter. This is why a probability sample is needed; without a probability sample, the sampling distribution cannot be determined and an interval estimate of a parameter cannot be constructed.

Estimation of a population mean

The most fundamental point and interval estimation process involves the estimation of a population mean. Suppose it is of interest to estimate the population mean, μ, for a quantitative variable. Data collected from a simple random sample can be used to compute the sample mean, $\bar{x}$, where the value of $\bar{x}$ provides a point estimate of μ.

When the sample mean is used as a point estimate of the population mean, some error can be expected owing to the fact that a sample, or subset of the population, is used to compute the point estimate. The absolute value of the difference between the sample mean, $\bar{x}$, and the population mean, μ, written $|\bar{x} - \mu|$, is called the sampling error. Interval estimation incorporates a probability statement about the magnitude of

the sampling error. The sampling distribution of $\bar{x}$ provides the basis for such a statement.

Statisticians have shown that the mean of the sampling distribution of $\bar{x}$ is equal to the population mean, μ, and that the standard deviation is given by σ/Square root of$\sqrt{n}$, where σ is the population standard deviation. The standard deviation of a sampling distribution is called the standard error. For large sample sizes, the central limit theorem indicates that the sampling distribution of $\bar{x}$ can be approximated by a normal probability distribution. As a matter of practice, statisticians usually consider samples of size 30 or more to be large.

In the large-sample case, a 95% confidence interval estimate for the population mean is given by $\bar{x}$ ± 1.96σ/Square root of$\sqrt{n}$. When the population standard deviation, σ, is unknown, the sample standard deviation is used to estimate σ in the confidence interval formula. The quantity 1.96σ/Square root of$\sqrt{n}$ is often called the margin of error for the estimate. The quantity σ/Square root of$\sqrt{n}$ is the standard error, and 1.96 is the number of standard errors from the mean necessary to include 95% of the values in a normal distribution. The interpretation of a 95% confidence interval is that 95% of the intervals constructed in this manner will contain the population mean. Thus, any interval computed in this manner has a 95% confidence of containing the population mean. By changing the constant from 1.96 to 1.645, a 90% confidence interval can be obtained. It should be noted from the formula for an interval estimate that a 90% confidence interval is narrower than a 95% confidence interval and as such has a slightly smaller confidence of including the population mean. Lower levels of confidence lead to even more narrow intervals. In practice, a 95% confidence interval is the most widely used.

Owing to the presence of the $n^{1/2}$ term in the formula for an interval estimate, the sample size affects the margin of error. Larger sample sizes lead to smaller margins of error. This observation forms the basis for procedures used to select the sample size. Sample sizes can be chosen such that the confidence interval satisfies any desired requirements about the size of the margin of error.

The procedure just described for developing interval estimates of a population mean is based on the use of a large sample. In the small-sample case—i.e., where the sample size *n* is less than 30—the *t* distribution is used when specifying the margin of error and constructing a confidence interval estimate. For example, at a 95% level of confidence, a value from the *t* distribution, determined by the value of *n*, would replace the 1.96 value obtained from the normal distribution. The *t* values will always be larger, leading to wider confidence intervals, but, as the sample size becomes larger, the *t* values get closer to the corresponding values from a normal distribution. With a sample size of 25, the *t* value used would be 2.064, as compared with the normal probability distribution value of 1.96 in the large-sample case.

Estimation of other parameters

For qualitative variables, the population proportion is a parameter of interest. A point estimate of the population proportion is given by the sample proportion. With knowledge of the sampling distribution of the sample proportion, an interval estimate of a population proportion is obtained in much the same fashion as for a population mean. Point and interval estimation procedures such as these can be applied to other population parameters as well. For instance, interval estimation of a

population variance, standard deviation, and total can be required in other applications.

Estimation procedures for two populations

The estimation procedures can be extended to two populations for comparative studies. For example, suppose a study is being conducted to determine differences between the salaries paid to a population of men and a population of women. Two independent simple random samples, one from the population of men and one from the population of women, would provide two sample means, $\bar{x}_1$ and $\bar{x}_2$. The difference between the two sample means, $\bar{x}_1 - \bar{x}_2$, would be used as a point estimate of the difference between the two population means. The sampling distribution of $\bar{x}_1 - \bar{x}_2$ would provide the basis for a confidence interval estimate of the difference between the two population means. For qualitative variables, point and interval estimates of the difference between population proportions can be constructed by considering the difference between sample proportions.

Hypothesis testing

Hypothesis testing is a form of statistical inference that uses data from a sample to draw conclusions about a population parameter or a population probability distribution. First, a tentative assumption is made about the parameter or distribution. This assumption is called the null hypothesis and is denoted by H_0. An alternative hypothesis (denoted H_a), which is the opposite of what is stated in the null hypothesis, is then defined. The hypothesis-testing procedure involves using sample data to determine whether or not H_0 can be rejected. If H_0 is rejected, the statistical conclusion is that the alternative hypothesis H_a is true.

For example, assume that a radio station selects the music it plays based on the assumption that the average age of its listening audience is 30 years. To determine whether this assumption is valid, a hypothesis test could be conducted with the null hypothesis given as H_0: $\mu = 30$ and the alternative hypothesis given as H_a: $\mu \neq 30$. Based on a sample of individuals from the listening audience, the sample mean age, $\bar{x}$, can be computed and used to determine whether there is sufficient statistical evidence to reject H_0. Conceptually, a value of the sample mean that is "close" to 30 is consistent with the null hypothesis, while a value of the sample mean that is "not close" to 30 provides support for the alternative hypothesis. What is considered "close" and "not close" is determined by using the sampling distribution of $\bar{x}$.

Ideally, the hypothesis-testing procedure leads to the acceptance of H_0 when H_0 is true and the rejection of H_0 when H_0 is false. Unfortunately, since hypothesis tests are based on sample information, the possibility of errors must be considered. A type I error corresponds to rejecting H_0 when H_0 is actually true, and a type II error corresponds to accepting H_0 when H_0 is false. The probability of making a type I error is denoted by α, and the probability of making a type II error is denoted by β.

In using the hypothesis-testing procedure to determine if the null hypothesis should be rejected, the person conducting the hypothesis test specifies the maximum allowable probability of making a type I error, called the level of significance for the test. Common choices for the level of significance are $\alpha = 0.05$ and $\alpha = 0.01$. Although most applications of hypothesis testing control the probability of making a type I error, they do not always control the probability of making a type II error. A graph

known as an operating-characteristic curve can be constructed to show how changes in the sample size affect the probability of making a type II error.

A concept known as the *p*-value provides a convenient basis for drawing conclusions in hypothesis-testing applications. The *p*-value is a measure of how likely the sample results are, assuming the null hypothesis is true; the smaller the *p*-value, the less likely the sample results. If the *p*-value is less than α, the null hypothesis can be rejected; otherwise, the null hypothesis cannot be rejected. The *p*-value is often called the observed level of significance for the test.

A hypothesis test can be performed on parameters of one or more populations as well as in a variety of other situations. In each instance, the process begins with the formulation of null and alternative hypotheses about the population. In addition to the population mean, hypothesis-testing procedures are available for population parameters such as proportions, variances, standard deviations, and medians.

Hypothesis tests are also conducted in regression and correlation analysis to determine if the regression relationship and the correlation coefficient are statistically significant (see below Regression and correlation analysis). A goodness-of-fit test refers to a hypothesis test in which the null hypothesis is that the population has a specific probability distribution, such as a normal probability distribution. Nonparametric statistical methods also involve a variety of hypothesis-testing procedures.

Bayesian methods

The methods of statistical inference previously described are often referred to as classical methods. Bayesian methods (so called after the English mathematician Thomas Bayes) provide alternatives that allow one to combine prior information about a population parameter with information contained in a sample to guide the statistical inference process. A prior probability distribution for a parameter of interest is specified first. Sample information is then obtained and combined through an application of Bayes's theorem to provide a posterior probability distribution for the parameter. The posterior distribution provides the basis for statistical inferences concerning the parameter.

A key, and somewhat controversial, feature of Bayesian methods is the notion of a probability distribution for a population parameter. According to classical statistics, parameters are constants and cannot be represented as random variables. Bayesian proponents argue that, if a parameter value is unknown, then it makes sense to specify a probability distribution that describes the possible values for the parameter as well as their likelihood. The Bayesian approach permits the use of objective data or subjective opinion in specifying a prior distribution. With the Bayesian approach, different individuals might specify different prior distributions. Classical statisticians argue that for this reason Bayesian methods suffer from a lack of objectivity. Bayesian proponents argue that the classical methods of statistical inference have built-in subjectivity (through the choice of a sampling plan) and that the advantage of the Bayesian approach is that the subjectivity is made explicit.

Bayesian methods have been used extensively in statistical decision theory (see below Decision analysis). In this context, Bayes's theorem provides a mechanism for combining a prior probability distribution for

the states of nature with sample information to provide a revised (posterior) probability distribution about the states of nature. These posterior probabilities are then used to make better decisions.

Experimental design

Data for statistical studies are obtained by conducting either experiments or surveys. Experimental design is the branch of statistics that deals with the design and analysis of experiments. The methods of experimental design are widely used in the fields of agriculture, medicine, biology, marketing research, and industrial production.

In an experimental study, variables of interest are identified. One or more of these variables, referred to as the factors of the study, are controlled so that data may be obtained about how the factors influence another variable referred to as the response variable, or simply the response. As a case in point, consider an experiment designed to determine the effect of three different exercise programs on the cholesterol level of patients with elevated cholesterol. Each patient is referred to as an experimental unit, the response variable is the cholesterol level of the patient at the completion of the program, and the exercise program is the factor whose effect on cholesterol level is being investigated. Each of the three exercise programs is referred to as a treatment.

Three of the more widely used experimental designs are the completely randomized design, the randomized block design, and the factorial design. In a completely randomized experimental design, the treatments are randomly assigned to the experimental units. For instance, applying this design method to the cholesterol-level study, the three types of exercise program (treatment) would be randomly assigned to the experimental units (patients).

The use of a completely randomized design will yield less precise results when factors not accounted for by the experimenter affect the response variable. Consider, for example, an experiment designed to study the effect of two different gasoline additives on the fuel efficiency, measured in miles per gallon (mpg), of full-size automobiles produced by three manufacturers. Suppose that 30 automobiles, 10 from each manufacturer, were available for the experiment. In a completely randomized design the two gasoline additives (treatments) would be randomly assigned to the 30 automobiles, with each additive being assigned to 15 different cars. Suppose that manufacturer 1 has developed an engine that gives its full-size cars a higher fuel efficiency than those produced by manufacturers 2 and 3. A completely randomized design could, by chance, assign gasoline additive 1 to a larger proportion of cars from manufacturer 1. In such a case, gasoline additive 1 might be judged to be more fuel efficient when in fact the difference observed is actually due to the better engine design of automobiles produced by manufacturer 1. To prevent this from occurring, a statistician could design an experiment in which both gasoline additives are tested using five cars produced by each manufacturer; in this way, any effects due to the manufacturer would not affect the test for significant differences due to gasoline additive. In this revised experiment, each of the manufacturers is referred to as a block, and the experiment is called a randomized block design. In general, blocking is used in order to enable comparisons among the treatments to be made within blocks of homogeneous experimental units.

Factorial experiments are designed to draw conclusions about more than one factor, or variable. The term factorial is used to indicate that all possible combinations of the factors are considered. For instance, if there are two factors with a levels for factor 1 and b levels for factor 2, the

experiment will involve collecting data on *ab* treatment combinations. The factorial design can be extended to experiments involving more than two factors and experiments involving partial factorial designs.

Analysis of variance and significance testing

A computational procedure frequently used to analyze the data from an experimental study employs a statistical procedure known as the analysis of variance. For a single-factor experiment, this procedure uses a hypothesis test concerning equality of treatment means to determine if the factor has a statistically significant effect on the response variable. For experimental designs involving multiple factors, a test for the significance of each individual factor as well as interaction effects caused by one or more factors acting jointly can be made. Further discussion of the analysis of variance procedure is contained in the subsequent section.

Regression and correlation analysis

Regression analysis involves identifying the relationship between a dependent variable and one or more independent variables. A model of the relationship is hypothesized, and estimates of the parameter values are used to develop an estimated regression equation. Various tests are then employed to determine if the model is satisfactory. If the model is deemed satisfactory, the estimated regression equation can be used to predict the value of the dependent variable given values for the independent variables.

Regression model

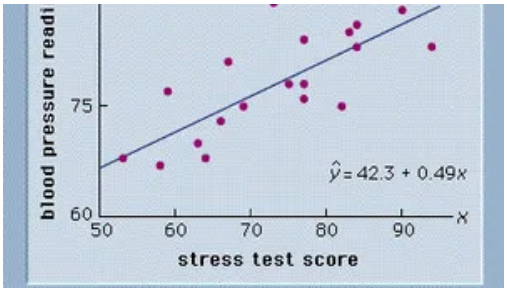
In simple linear regression, the model used to describe the relationship between a single dependent variable *y* and a single independent variable

x is $y = \beta_0 + \beta_1 x + \varepsilon$. β_0 and β_1 are referred to as the model parameters, and ε is a probabilistic error term that accounts for the variability in y that cannot be explained by the linear relationship with x. If the error term were not present, the model would be deterministic; in that case, knowledge of the value of x would be sufficient to determine the value of y.

In multiple regression analysis, the model for simple linear regression is extended to account for the relationship between the dependent variable y and p independent variables $x_1, x_2, \ldots, x_p$. The general form of the multiple regression model is $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p + \varepsilon$. The parameters of the model are the $\beta_0, \beta_1, \ldots, \beta_p$, and ε is the error term.

Least squares method

Either a simple or multiple regression model is initially posed as a hypothesis concerning the relationship among the dependent and independent variables. The least squares method is the most widely used procedure for developing estimates of the model parameters. For simple linear regression, the least squares estimates of the model parameters β_0 and β_1 are denoted b_0 and b_1. Using these estimates, an estimated regression equation is constructed: $\hat{y} = b_0 + b_1 x$. The graph of the estimated regression equation for simple linear regression is a straight line approximation to the relationship between y and x.



scatter diagram with estimated regression equation

As an illustration of regression analysis and the least squares method, suppose a university medical centre is investigating the relationship between stress and blood pressure. Assume that both a stress test

score and a blood pressure reading have been recorded for a sample of 20 patients. The data are shown graphically in Figure 4, called a scatter diagram. Values of the independent variable, stress test score, are given on the horizontal axis, and values of the dependent variable, blood pressure, are shown on the vertical axis. The line passing through the data points is the graph of the estimated regression equation: $\hat{y} = 42.3 + 0.49x$. The parameter estimates, $b_0 = 42.3$ and $b_1 = 0.49$, were obtained using the least squares method.

A primary use of the estimated regression equation is to predict the value of the dependent variable when values for the independent variables are given. For instance, given a patient with a stress test score of 60, the predicted blood pressure is 42.3 + 0.49(60) = 71.7. The values predicted by the estimated regression equation are the points on the line in Figure 4, and the actual blood pressure readings are represented by the points scattered about the line. The difference between the observed value of y and the value of y predicted by the estimated regression equation is called a residual. The least squares method chooses the parameter estimates such that the sum of the squared residuals is minimized.

Analysis of variance and goodness of fit

A commonly used measure of the goodness of fit provided by the estimated regression equation is the coefficient of determination. Computation of this coefficient is based on the analysis of variance procedure that partitions the total variation in the dependent variable, denoted SST, into two parts: the part explained by the estimated regression equation, denoted SSR, and the part that remains unexplained, denoted SSE.

The measure of total variation, SST, is the sum of the squared deviations of the dependent variable about its mean: $\Sigma(y - \bar{y})^2$. This quantity is known as the total sum of squares. The measure of unexplained variation, SSE, is referred to as the residual sum of squares. For the data in Figure 4, SSE is the sum of the squared distances from each point in the scatter diagram (see Figure 4) to the estimated regression line: $\Sigma(y - \hat{y})^2$. SSE is also commonly referred to as the error sum of squares. A key result in the analysis of variance is that SSR + SSE = SST.

The ratio r^2 = SSR/SST is called the coefficient of determination. If the data points are clustered closely about the estimated regression line, the value of SSE will be small and SSR/SST will be close to 1. Using r^2, whose values lie between 0 and 1, provides a measure of goodness of fit; values closer to 1 imply a better fit. A value of $r^2 = 0$ implies that there is no linear relationship between the dependent and independent variables.

When expressed as a percentage, the coefficient of determination can be interpreted as the percentage of the total sum of squares that can be explained using the estimated regression equation. For the stress-level research study, the value of r^2 is 0.583; thus, 58.3% of the total sum of squares can be explained by the estimated regression equation $\hat{y} = 42.3 + 0.49x$. For typical data found in the social sciences, values of r^2 as low as 0.25 are often considered useful. For data in the physical sciences, r^2 values of 0.60 or greater are frequently found.

Significance testing

In a regression study, hypothesis tests are usually conducted to assess the statistical significance of the overall relationship represented by the

regression model and to test for the statistical significance of the individual parameters. The statistical tests used are based on the following assumptions concerning the error term: (1) ε is a random variable with an expected value of 0, (2) the variance of ε is the same for all values of x, (3) the values of ε are independent, and (4) ε is a normally distributed random variable.

The mean square due to regression, denoted MSR, is computed by dividing SSR by a number referred to as its degrees of freedom; in a similar manner, the mean square due to error, MSE, is computed by dividing SSE by its degrees of freedom. An F-test based on the ratio MSR/MSE can be used to test the statistical significance of the overall relationship between the dependent variable and the set of independent variables. In general, large values of F = MSR/MSE support the conclusion that the overall relationship is statistically significant. If the overall model is deemed statistically significant, statisticians will usually conduct hypothesis tests on the individual parameters to determine if each independent variable makes a significant contribution to the model.

Residual analysis

The analysis of residuals plays an important role in validating the regression model. If the error term in the regression model satisfies the four assumptions noted earlier, then the model is considered valid. Since the statistical tests for significance are also based on these assumptions, the conclusions resulting from these significance tests are called into question if the assumptions regarding ε are not satisfied.

The ith residual is the difference between the observed value of the dependent variable, y_i, and the value predicted by the estimated

regression equation, $\hat{y}_i$. These residuals, computed from the available data, are treated as estimates of the model error, ε. As such, they are used by statisticians to validate the assumptions concerning ε. Good judgment and experience play key roles in residual analysis.

Graphical plots and statistical tests concerning the residuals are examined carefully by statisticians, and judgments are made based on these examinations. The most common residual plot shows $\hat{y}$ on the horizontal axis and the residuals on the vertical axis. If the assumptions regarding the error term, ε, are satisfied, the residual plot will consist of a horizontal band of points. If the residual analysis does not indicate that the model assumptions are satisfied, it often suggests ways in which the model can be modified to obtain better results.

Model building

In regression analysis, model building is the process of developing a probabilistic model that best describes the relationship between the dependent and independent variables. The major issues are finding the proper form (linear or curvilinear) of the relationship and selecting which independent variables to include. In building models it is often desirable to use qualitative as well as quantitative variables.

As noted above, quantitative variables measure how much or how many; qualitative variables represent types or categories. For instance, suppose it is of interest to predict sales of an iced tea that is available in either bottles or cans. Clearly, the independent variable “container type” could influence the dependent variable “sales.” Container type is a qualitative variable, however, and must be assigned numerical values if it is to be used in a regression study. So-called dummy variables are used to

represent qualitative variables in regression analysis. For example, the dummy variable x could be used to represent container type by setting $x = 0$ if the iced tea is packaged in a bottle and $x = 1$ if the iced tea is in a can. If the beverage could be placed in glass bottles, plastic bottles, or cans, it would require two dummy variables to properly represent the qualitative variable container type. In general, k - 1 dummy variables are needed to model the effect of a qualitative variable that may assume k values.

The general linear model $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p + \varepsilon$ can be used to model a wide variety of curvilinear relationships between dependent and independent variables. For instance, each of the independent variables could be a nonlinear function of other variables. Also, statisticians sometimes find it necessary to transform the dependent variable in order to build a satisfactory model. A logarithmic transformation is one of the more common types.

Correlation

Correlation and regression analysis are related in the sense that both deal with relationships among variables. The correlation coefficient is a measure of linear association between two variables. Values of the correlation coefficient are always between −1 and +1. A correlation coefficient of +1 indicates that two variables are perfectly related in a positive linear sense, a correlation coefficient of −1 indicates that two variables are perfectly related in a negative linear sense, and a correlation coefficient of 0 indicates that there is no linear relationship between the two variables. For simple linear regression, the sample correlation coefficient is the square root of the coefficient of determination, with the

sign of the correlation coefficient being the same as the sign of b_1, the coefficient of x_1 in the estimated regression equation.

Neither regression nor correlation analyses can be interpreted as establishing cause-and-effect relationships. They can indicate only how or to what extent variables are associated with each other. The correlation coefficient measures only the degree of linear association between two variables. Any conclusions about a cause-and-effect relationship must be based on the judgment of the analyst.

Time series and forecasting

A time series is a set of data collected at successive points in time or over successive periods of time. A sequence of monthly data on new housing starts and a sequence of weekly data on product sales are examples of time series. Usually the data in a time series are collected at equally spaced periods of time, such as hour, day, week, month, or year.

A primary concern of time series analysis is the development of forecasts for future values of the series. For instance, the federal government develops forecasts of many economic time series such as the gross domestic product, exports, and so on. Most companies develop forecasts of product sales.

While in practice both qualitative and quantitative forecasting methods are utilized, statistical approaches to forecasting employ quantitative methods. The two most widely used methods of forecasting are the Box-Jenkins autoregressive integrated moving average (ARIMA) and econometric models.

ARIMA methods are based on the assumption that a probability model generates the time series data. Future values of the time series are

assumed to be related to past values as well as to past errors. A time series must be stationary, *i.e.*, one which has a constant mean, variance, and autocorrelation function, in order for an ARIMA model to be applicable. For nonstationary series, sometimes differences between successive values can be taken and used as a stationary series to which the ARIMA model can be applied.

Econometric models develop forecasts of a time series using one or more related time series and possibly past values of the time series. This approach involves developing a regression model in which the time series is forecast as the dependent variable; the related time series as well as the past values of the time series are the independent or predictor variables.

Nonparametric methods

The statistical methods discussed above generally focus on the parameters of populations or probability distributions and are referred to as parametric methods. Nonparametric methods are statistical methods that require fewer assumptions about a population or probability distribution and are applicable in a wider range of situations. For a statistical method to be classified as a nonparametric method, it must satisfy one of the following conditions: (1) the method is used with qualitative data, or (2) the method is used with quantitative data when no assumption can be made about the population probability distribution. In cases where both parametric and nonparametric methods are applicable, statisticians usually recommend using parametric methods because they tend to provide better precision. Nonparametric methods are useful, however, in situations where the assumptions required by parametric methods appear questionable. A few of the more commonly used nonparametric methods are described below.

Assume that individuals in a sample are asked to state a preference for one of two similar and competing products. A plus (+) sign can be recorded if an individual prefers one product and a minus (−) sign if the individual prefers the other product. With qualitative data in this form, the nonparametric sign test can be used to statistically determine whether a difference in preference for the two products exists for the population. The sign test also can be used to test hypotheses about the value of a population median.

The Wilcoxon signed-rank test can be used to test hypotheses about two populations. In collecting data for this test, each element or experimental unit in the sample must generate two paired or matched data values, one from population 1 and one from population 2. Differences between the paired or matched data values are used to test for a difference between the two populations. The Wilcoxon signed-rank test is applicable when no assumption can be made about the form of the probability distributions for the populations. Another nonparametric test for detecting differences between two populations is the Mann-Whitney-Wilcoxon test. This method is based on data from two independent random samples, one from population 1 and another from population 2. There is no matching or pairing as required for the Wilcoxon signed-rank test.

Nonparametric methods for correlation analysis are also available. The Spearman rank correlation coefficient is a measure of the relationship between two variables when data in the form of rank orders are available. For instance, the Spearman rank correlation coefficient could be used to determine the degree of agreement between men and women concerning their preference ranking of 10 different television shows. A

Spearman rank correlation coefficient of 1 would indicate complete agreement, a coefficient of −1 would indicate complete disagreement, and a coefficient of 0 would indicate that the rankings were unrelated.

Statistical quality control

Statistical quality control refers to the use of statistical methods in the monitoring and maintaining of the quality of products and services. One method, referred to as acceptance sampling, can be used when a decision must be made to accept or reject a group of parts or items based on the quality found in a sample. A second method, referred to as statistical process control, uses graphical displays known as control charts to determine whether a process should be continued or should be adjusted to achieve the desired quality.

Acceptance sampling

Assume that a consumer receives a shipment of parts called a lot from a producer. A sample of parts will be taken and the number of defective items counted. If the number of defective items is low, the entire lot will be accepted. If the number of defective items is high, the entire lot will be rejected. Correct decisions correspond to accepting a good-quality lot and rejecting a poor-quality lot. Because sampling is being used, the probabilities of erroneous decisions need to be considered. The error of rejecting a good-quality lot creates a problem for the producer; the probability of this error is called the producer’s risk. On the other hand, the error of accepting a poor-quality lot creates a problem for the purchaser or consumer; the probability of this error is called the consumer’s risk.

The design of an acceptance sampling plan consists of determining a sample size n and an acceptance criterion c, where c is the maximum

number of defective items that can be found in the sample and the lot still be accepted. The key to understanding both the producer's risk and the consumer's risk is to assume that a lot has some known percentage of defective items and compute the probability of accepting the lot for a given sampling plan. By varying the assumed percentage of defective items in a lot, several different sampling plans can be evaluated and a sampling plan selected such that both the producer's and consumer's risks are reasonably low.

Statistical process control

Statistical process control uses sampling and statistical methods to monitor the quality of an ongoing process such as a production operation. A graphical display referred to as a control chart provides a basis for deciding whether the variation in the output of a process is due to common causes (randomly occurring variations) or to out-of-the-ordinary assignable causes. Whenever assignable causes are identified, a decision can be made to adjust the process in order to bring the output back to acceptable quality levels.

Control charts can be classified by the type of data they contain. For instance, an $\bar{x}$-chart is employed in situations where a sample mean is used to measure the quality of the output. Quantitative data such as length, weight, and temperature can be monitored with an $\bar{x}$-chart. Process variability can be monitored using a range or R-chart. In cases in which the quality of output is measured in terms of the number of defectives or the proportion of defectives in the sample, an np-chart or a p-chart can be used.

All control charts are constructed in a similar fashion. For example, the centre line of an $\bar{x}$-chart corresponds to the mean of the process when the process is in control and producing output of acceptable quality. The vertical axis of the control chart identifies the scale of measurement for the variable of interest. The upper horizontal line of the control chart, referred to as the upper control limit, and the lower horizontal line, referred to as the lower control limit, are chosen so that when the process is in control there will be a high probability that the value of a sample mean will fall between the two control limits. Standard practice is to set the control limits at three standard deviations above and below the process mean. The process can be sampled periodically. As each sample is selected, the value of the sample mean is plotted on the control chart. If the value of a sample mean is within the control limits, the process can be continued under the assumption that the quality standards are being maintained. If the value of the sample mean is outside the control limits, an out-of-control conclusion points to the need for corrective action in order to return the process to acceptable quality levels.

Sample survey methods

As noted above in the section Estimation, statistical inference is the process of using data from a sample to make estimates or test hypotheses about a population. The field of sample survey methods is concerned with effective ways of obtaining sample data. The three most common types of sample surveys are mail surveys, telephone surveys, and personal interview surveys. All of these involve the use of a questionnaire, for which a large body of knowledge exists concerning the phrasing, sequencing, and grouping of questions. There are other types of sample surveys that do not involve a questionnaire. For example, the sampling of accounting records for audits and the use of a computer to

sample a large database are sample surveys that use direct observation of the sampled units to collect the data.

A goal in the design of sample surveys is to obtain a sample that is representative of the population so that precise inferences can be made. Sampling error is the difference between a population parameter and a sample statistic used to estimate it. For example, the difference between a population mean and a sample mean is sampling error. Sampling error occurs because a portion, and not the entire population, is surveyed. Probability sampling methods, where the probability of each unit appearing in the sample is known, enable statisticians to make probability statements about the size of the sampling error. Nonprobability sampling methods, which are based on convenience or judgment rather than on probability, are frequently used for cost and time advantages. However, one should be extremely careful in making inferences from a nonprobability sample; whether or not the sample is representative is dependent on the judgment of the individuals designing and conducting the survey and not on sound statistical principles. In addition, there is no objective basis for establishing bounds on the sampling error when a nonprobability sample has been used.

Most governmental and professional polling surveys employ probability sampling. It can generally be assumed that any survey that reports a plus or minus margin of error has been conducted using probability sampling. Statisticians prefer probability sampling methods and recommend that they be used whenever possible. A variety of probability sampling methods are available. A few of the more common ones are reviewed here.

Simple random sampling provides the basis for many probability sampling methods. With simple random sampling, every possible sample of size *n* has the same probability of being selected. This method was discussed above in the section Estimation.

Stratified simple random sampling is a variation of simple random sampling in which the population is partitioned into relatively homogeneous groups called strata and a simple random sample is selected from each stratum. The results from the strata are then aggregated to make inferences about the population. A side benefit of this method is that inferences about the subpopulation represented by each stratum can also be made.

Cluster sampling involves partitioning the population into separate groups called clusters. Unlike in the case of stratified simple random sampling, it is desirable for the clusters to be composed of heterogeneous units. In single-stage cluster sampling, a simple random sample of clusters is selected, and data are collected from every unit in the sampled clusters. In two-stage cluster sampling, a simple random sample of clusters is selected and then a simple random sample is selected from the units in each sampled cluster. One of the primary applications of cluster sampling is called area sampling, where the clusters are counties, townships, city blocks, or other well-defined geographic sections of the population.

Decision analysis

Decision analysis, also called statistical decision theory, involves procedures for choosing optimal decisions in the face of uncertainty. In the simplest situation, a decision maker must choose the best decision from a finite set of alternatives when there are two or more possible

future events, called states of nature, that might occur. The list of possible states of nature includes everything that can happen, and the states of nature are defined so that only one of the states will occur. The outcome resulting from the combination of a decision alternative and a particular state of nature is referred to as the payoff.

When probabilities for the states of nature are available, probabilistic criteria may be used to choose the best decision alternative. The most common approach is to use the probabilities to compute the expected value of each decision alternative. The expected value of a decision alternative is the sum of weighted payoffs for the decision. The weight for a payoff is the probability of the associated state of nature and therefore the probability that the payoff occurs. For a maximization problem, the decision alternative with the largest expected value will be chosen; for a minimization problem, the decision alternative with the smallest expected value will be chosen.

Decision analysis can be extremely helpful in sequential decision-making situations—that is, situations in which a decision is made, an event occurs, another decision is made, another event occurs, and so on. For instance, a company trying to decide whether or not to market a new product might first decide to test the acceptance of the product using a consumer panel. Based on the results of the consumer panel, the company will then decide whether or not to proceed with further test marketing; after analyzing the results of the test marketing, company executives will decide whether or not to produce the new product. A decision tree is a graphical device that is helpful in structuring and analyzing such problems. With the aid of decision trees, an optimal decision strategy can be developed. A decision strategy is a contingency

plan that recommends the best decision alternative depending on what has happened earlier in the sequential process.

David R. Anderson Dennis J. Sweeney Thomas A. Williams

Citation Information
Article Title: statistics
Website Name: Encyclopaedia Britannica
Publisher: Encyclopaedia Britannica, Inc.
Date Published: 20 June 2023
URL: https://www.britannica.comhttps://www.britannica.com/science/statistics
Access Date: July 23, 2023