LOAD PREVIOUS PAGE

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-lev study, the three types of exercise program (treatment) would be randomly assigned 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 a b levels for factor 2, the experiment will involve collecting data on ab treatment

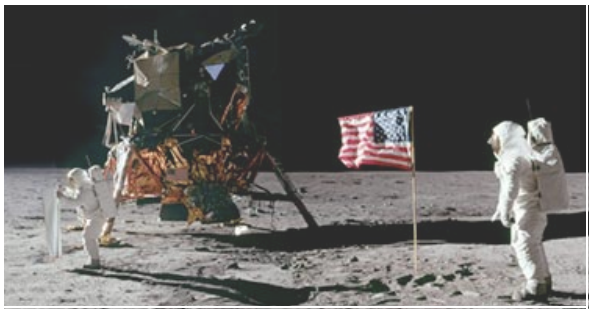
Analysis of variance and significance testing

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.

To Examine Space.



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 estimat regression equation for simple linear regression is a straight line approximation to th relationship between y and x.

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.

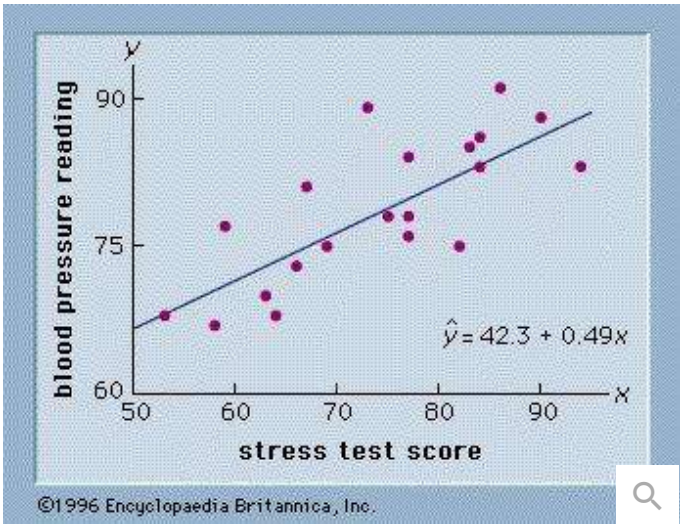


Figure 4: A scatter diagram showing the relationship between stress and blood pressure.

Encyclopædia Britannica, Inc.

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.

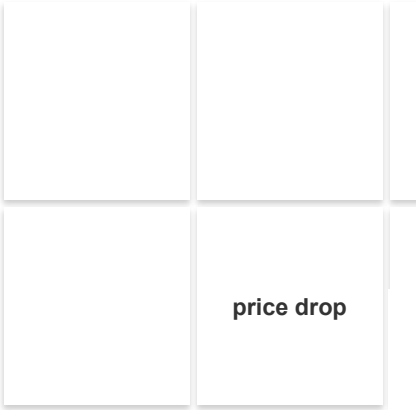
When expressed as a percentage, the coefficient of determination can be interprete as the percentage of the total sum of squares that can be explained using the estimated regression equation. For the stress-level research study, the value of r^2 is

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.

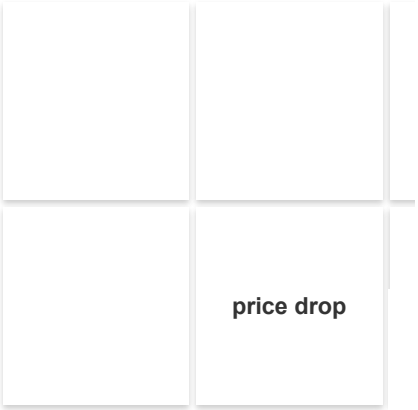


Residual analysis

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.

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.

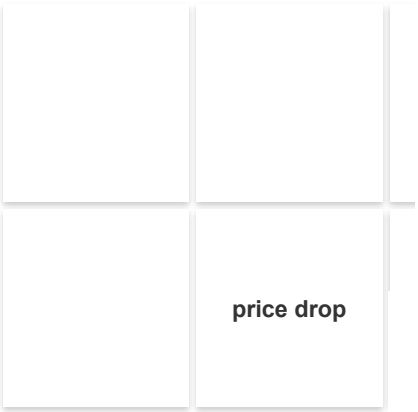


Model building

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 pred sales of an iced tea that is available in either bottles or cans. Clearly, the independen

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 tha two variables are perfectly related in a negative linear sense, and a correlation

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 be applicable. For nonstationary series, sometimes differences between successive

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.

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

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

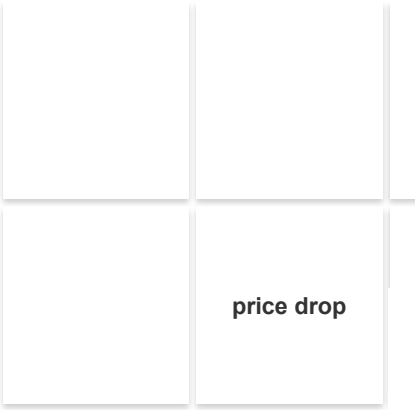
All control charts are constructed in a similar fashion. For example, the centre line of an $\bar{x}$-chart corresponds to the mean of the process when the process is in control and producing output of acceptable quality. The vertical axis of the control chart identifies the scale of measurement for 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.

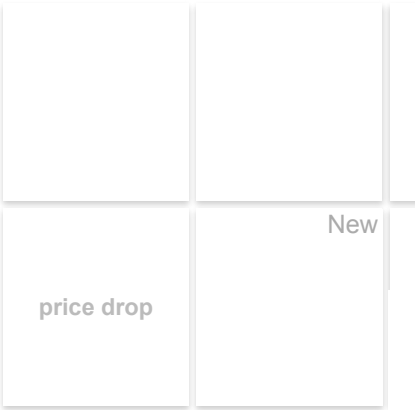
A goal in the design of sample surveys is to obtain a sample that is representative of the population so that precise inferences can be made. Sampling error is the difference between a population parameter and a sample statistic used to estimate it. For example, the difference between a population mean 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



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.



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 o

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

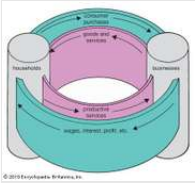
Dennis J. Sweeney

Thomas A. Williams

LEARN MORE in these related Britannica articles:

probability and statistics: The rise of statistics

During the 19th century, statistics grew up as the empirical science of the state and gained preeminence as a form of social knowledge. Population and economic numbers had been collected, though often not in a systematic way, since ancient times and in…



economics: Postwar developments

…theory, mathematical model building, and statistical testing of economic predictions. The development of econometrics had an impact on economics in general, since those who formulated new theories began to cast them in terms that allowed empirical testing….