Mathematical statistics

Mathematical statistics is the application of probability theory, a branch of mathematics, to statistics, as opposed to techniques for collecting statistical data. Specific mathematical techniques which are used for this include mathematical analysis, linear algebra, stochastic analysis, differential equations, and measure theory.[1][2]
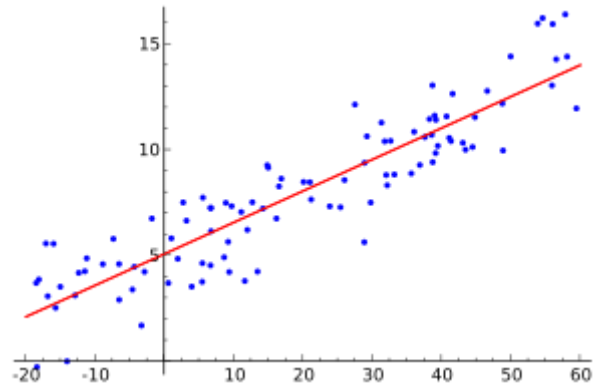

Illustration of linear regression on a data set. Regression analysis is an important part of mathematical statistics.

Contents

Introduction

Statistical data collection is concerned with the planning of studies, especially with the design of randomized experiments and with the planning of surveys using random sampling. The initial analysis of the data often follows the study protocol specified prior to the study being conducted. The data from a study can also be analyzed to consider secondary hypotheses inspired by the initial results, or to suggest new studies. A secondary analysis of the data from a planned study uses tools from data analysis, and the process of doing this is mathematical statistics.

Data analysis is divided into:

- descriptive statistics - the part of statistics that describes data, i.e. summarises the data and their typical properties.
- inferential statistics - the part of statistics that draws conclusions from data (using some model for the data): For example, inferential statistics involves selecting a model for the data, checking whether the data fulfill the conditions of a particular model, and with quantifying the involved uncertainty (e.g. using confidence intervals).

While the tools of data analysis work best on data from randomized studies, they are also applied to other kinds of data. For example, from natural experiments and observational studies, in which case the inference is dependent on the model chosen by the statistician, and so subjective.[3]

Topics

The following are some of the important topics in mathematical statistics:[4][5]

Probability distributions

A probability distribution is a function that assigns a probability to each measurable subset of the possible outcomes of a random experiment, survey, or procedure of statistical inference. Examples are found in experiments whose sample space is non-numerical, where the distribution would be a categorical distribution; experiments whose sample space is encoded by discrete random variables, where the distribution can be specified by a probability mass function; and experiments with sample spaces encoded by continuous random variables, where the distribution can be specified by a probability density function. More complex experiments, such as those involving stochastic processes defined in continuous time, may demand the use of more general probability measures.

A probability distribution can either be univariate or multivariate. A univariate distribution gives the probabilities of a single random variable taking on various alternative values; a multivariate distribution (a joint probability distribution) gives the probabilities of a random vector—a set of two or more random variables—taking on various combinations of values. Important and commonly encountered univariate probability distributions include the binomial distribution, the hypergeometric distribution, and the normal distribution. The multivariate normal distribution is a commonly encountered multivariate distribution.

Special distributions

- Normal distribution, the most common continuous distribution
- Bernoulli distribution, for the outcome of a single Bernoulli trial (e.g. success/failure, yes/no)
- Binomial distribution, for the number of "positive occurrences" (e.g. successes, yes votes, etc.) given a fixed total number of independent occurrences
- Negative binomial distribution, for binomial-type observations but where the quantity of interest is the number of failures before a given number of successes occurs
- Geometric distribution, for binomial-type observations but where the quantity of interest is the number of failures before the first success; a special case of the negative binomial distribution, where the number of successes is one.
- Discrete uniform distribution, for a finite set of values (e.g. the outcome of a fair die)
- Continuous uniform distribution, for continuously distributed values
- Poisson distribution, for the number of occurrences of a Poisson-type event in a given period of time
- Exponential distribution, for the time before the next Poisson-type event occurs
- Gamma distribution, for the time before the next k Poisson-type events occur
- Chi-squared distribution, the distribution of a sum of squared standard normal variables; useful e.g. for inference regarding the sample variance of normally distributed samples (see chi-squared test)
- Student's t distribution, the distribution of the ratio of a standard normal variable and the square root of a scaled chi squared variable; useful for inference regarding the mean of normally distributed samples with unknown variance (see Student's t-test)
- Beta distribution, for a single probability (real number between 0 and 1); conjugate to the Bernoulli distribution and binomial distribution

Statistical inference

Statistical inference is the process of drawing conclusions from data that are subject to random variation, for example, observational errors or sampling variation.[6] Initial requirements of such a system of procedures for inference and induction are that the system should produce reasonable answers when applied to well-defined situations and that it should be general enough to be applied across a range of situations. Inferential statistics are used to test hypotheses and make estimations using sample data. Whereas descriptive statistics describe a sample, inferential statistics infer predictions about a larger population that the sample represents.

The outcome of statistical inference may be an answer to the question "what should be done next?", where this might be a decision about making further experiments or surveys, or about drawing a conclusion before implementing some organizational or governmental policy. For the most part, statistical inference makes propositions about populations, using data drawn from the population of interest via some form of random sampling. More generally, data about a random process is obtained from its observed behavior during a finite period of time. Given a parameter or hypothesis about which one wishes to make inference, statistical inference most often uses:

- a statistical model of the random process that is supposed to generate the data, which is known when randomization has been used, and
- a particular realization of the random process; i.e., a set of data.

Regression

In statistics, **regression analysis** is a statistical process for estimating the relationships among variables. It includes many techniques for modeling and analyzing several variables, when the focus is on the relationship between a dependent variable and one or more independent variables More specifically, regression analysis helps one understand how the typical value of the dependent variable (or 'criterion variable') changes when any one of the independent variables is varied, while the other independent variables are held fixed. Most commonly, regression analysis estimates the conditional expectation of the dependent variable given the independent variables – that is, the average value of the dependent variable when the independent variables are fixed. Less commonly, the focus is on a quantile, or other location parameter of the conditional distribution of the dependent variable given the independent variables. In all cases, the estimation target is a function of the independent variables called the **regression function** In regression analysis, it is also of interest to characterize the variation of the dependent variable around the regression function which can be described by a probability distribution

Many techniques for carrying out regression analysis have been developed. Familiar methods, such as linear regression, are parametric, in that the regression function is defined in terms of a finite number of unknown parameters that are estimated from the data (e.g. using ordinary least squares). Nonparametric regression refers to techniques that allow the regression function to lie in a specified set of functions, which may be infinite-dimensional

Nonparametric statistics

Nonparametric statistics are values calculated from data in a way that is not based on parameterized families of probability distributions. They include both descriptive and inferential statistics. The typical parameters are the mean, variance, etc. Unlike parametric statistics, nonparametric statistics make no assumptions about the probability distributions of the variables being assessed.

Non-parametric methods are widely used for studying populations that take on a ranked order (such as movie reviews receiving one to four stars). The use of non-parametric methods may be necessary when data have a ranking but no clear numerical interpretation, such as when assessing preferences. In terms of levels of measurement, non-parametric methods result in "ordinal" data.

As non-parametric methods make fewer assumptions, their applicability is much wider than the corresponding parametric methods. In particular, they may be applied in situations where less is known about the application in question. Also, due to the reliance on fewer assumptions, non-parametric methods are more robust.

Another justification for the use of non-parametric methods is simplicity. In certain cases, even when the use of parametric methods is justified, non-parametric methods may be easier to use. Due both to this simplicity and to their greater robustness, non-parametric methods are seen by some statisticians as leaving less room for improper use and misunderstanding.

Statistics, mathematics, and mathematical statistics

Mathematical statistics is a key subset of the discipline of statistics. Statistical theorists study and improve statistical procedures with mathematics, and statistical research often raises mathematical questions. Statistical theory relies on probability and decision theory.

Mathematicians and statisticians like Gauss, Laplace, and C. S. Peirce used decision theory with probability distributions and loss functions (or utility functions). The decision-theoretic approach to statistical inference was reinvigorated by Abraham Wald and his successors,[7][8][9][10][11][12][13] and makes extensive use of scientific computing, analysis, and optimization; for the design of experiments, statisticians use algebra and combinatorics.

See also

- Asymptotic theory (statistics)

References

1. Lakshmikantham,, ed. by D. Kannan,... V (2002). *Handbook of stochastic analysis and applications*. New York: M. Dekker. ISBN 0824706609.
2. Schervish, Mark J. (1995). *Theory of statistics* (Corr. 2nd print. ed.). New York: Springer. ISBN 0387945466.
3. Freedman, D.A. (2005) *Statistical Models: Theory and Practice*, Cambridge University Press. ISBN 978-0-521-67105-7
4. Hogg, R. V., A. Craig, and J. W. McKean. "Intro to Mathematical Statistics." (2005).
5. Larsen, Richard J. and Marx, Morris L. "An Introduction to Mathematical Statistics and Its Applications" (2012). Prentice Hall.
6. Upton, G., Cook, I. (2008) *Oxford Dictionary of Statistics*, OUP. ISBN 978-0-19-954145-4
7. Wald, Abraham (1947). *Sequential analysis*. New York: John Wiley and Sons. ISBN 0-471-91806-7. "See Dover reprint, 2004: ISBN 0-486-43912-7"
8. Wald, Abraham (1950). *Statistical Decision Functions*. John Wiley and Sons, New York.
9. Lehmann, Erich (1997). *Testing Statistical Hypotheses* (2nd ed.). ISBN 0-387-94919-4.
10. Lehmann, Erich; Cassella, George (1998). *Theory of Point Estimation* (2nd ed.). ISBN 0-387-98502-6.
11. Bickel, Peter J.; Doksum, Kjell A. (2001). *Mathematical Statistics: Basic and Selected Topics*. **1** (Second (updated printing 2007) ed.). Pearson Prentice-Hall.
12. Le Cam, Lucien (1986). *Asymptotic Methods in Statistical Decision Theory*. Springer-Verlag. ISBN 0-387-96307-3.
13. Liese, Friedrich & Miescke, Klaus-J. (2008) *Statistical Decision Theory: Estimation, Testing, and Selection*. Springer.

Further reading

- Borovkov, A. A. (1999). *Mathematical Statistics*. CRC Press. ISBN 90-5699-018-7
- Virtual Laboratories in Probability and Statistics (Univ. of Ala.-Huntsville)
- StatiBot, interactive online expert system on statistical tests.

Retrieved from "https://en.wikipedia.org/w/index.php?title=Mathematical_statistics&oldid=876831477"