Introduction and History of Statistics

Introduction

Statistics is a mathematical science pertaining to the collection, organization, analysis, interpretation and presentation of data. It also provides tools for explanation, prediction and forecasting based on data. Statistics are applicable to a wide variety of disciplines, from academics to medicine, government and business.

Statistical methods can be used to summarize or describe a collection of data; which is called descriptive statistics. Additionally, statistics are often used to draw inferences about the process or population being studied; this is called inferential statistics. Descriptive, and inferential (for explanation or prediction) statistics comprise applied statistics.

Mathematical statistics is concerned with the theoretical basis of the discipline and studies the conditions and assumptions of statistical methods. In this class, we focus on applied statistics and will briefly talk about mathematical statistics.

In the application of statistics to a scientific, industrial, or societal problem, it is necessary to begin with a process or population to be studied. This might be a population of people in a country, of undergraduate students at major universities, or of manufacturing output by a particular factory during a given period. It could also be a process observed at various times. Data collected about this last kind of "population" constitute what is called a time series.

For practical reasons, rather than compiling data about an entire population (such as in a census), a chosen subset of the population is studied. The procedure for selecting a subset is called sampling, and the chosen subset is called a sample. Data are collected about the sample in an observational or experimental setting. The data are then subjected to statistical analysis, which serves two related purposes: description and inference.

Descriptive statistics are used to summarize the data, either numerically or graphically, to describe the sample. Basic examples of numerical descriptors include graphical summarizations, which in turn include a variety of charts, graphs, percentiles, means and standard deviations.

Inferential statistics are used to model patterns in the data in order to draw inferences about the larger population. After all, the sample is not the population and we hope to know about the population based on data collected from the sample. That is, we “infer” something about the population from the sample. These inferences may take the form of answers to yes/no questions (hypothesis testing), estimates of numerical characteristics (estimation), descriptions of association (correlation) or modeling of relationships (regression analysis). Other modeling techniques include analysis of variance (ANOVA), time series, and data mining.

The concept of correlation is of particular interest. Statistical analysis

may reveal that two variables (that is, two properties of the population under consideration) tend to vary together, as if they are connected. For example, a study of education and age of death among people might find that those with less education tend to have shorter lives than people with more education. The two variables are said to be correlated (which is a positive correlation in this case). However, one cannot immediately infer the existence of a causal relationship between the two variables. The correlated phenomena could be caused by a third, previously unconsidered phenomenon, called a concomitant or confounding variable. In this case, access to health care could be a concomitant variable.

Another concept is the distinction between a sample versus a population. If the sample is representative of the population, then we can make inferences and conclusions about the population based on information about the sample. A major problem lies in determining the extent to which the chosen sample is representative. There are systematical ways (called sampling methods) to make sure it is the case. In experimental designs, it is also critical to make sure that participants are randomly assigned to experimental conditions (e.g., control and treatments). Statistics offers methods to estimate and correct for randomness in the sample.

A third fundamental mathematical concept employed in understanding such randomness is probability. Mathematical statistics is the branch of applied mathematics that uses probability theory and analysis to examine the theoretical basis of statistics.

The use of any statistical method is valid only when the research design and population under consideration satisfies the basic statistical assumptions of the method. Inappropriate design or misuse of statistics can produce subtle but serious errors in description and interpretation. Subtle in the sense that sometimes even experienced researchers make such errors; serious in the sense that misinterpretation can impact on subsequent outcomes (e.g., forecasting).

As a student of statistics, you must learn to use your intuition. The statistical significance of a trend in the data, which measures the extent to which the trend could be caused by randomness of the sample, may not agree with one's intuitive sense of its significance. When one notes such a discrepancy, one should look carefully at the research design and statistical analysis. The set of basic statistical skills and skepticism needed by people to deal with research information in their fields is referred to as statistical literacy.

The basic steps of an experiment are:

1. Planning the research, including determining information sources, research subject selection, and ethical considerations for the proposed research and method.

2. Implementing the experimental design which includes concentrating on the system model and the interaction of independent and dependent variables.
3. Summarizing a collection of observations (descriptive statistics).
4. Making inference about the variables studied (inferential statistics).
5. Documenting / presenting the results of the study.

Some of the subset or specialized fields of applied statistics include:

Actuarial science
Biostatistics
Business statistics
Chemometrics (chemistry statistics)
Data analysis
Data mining (applying pattern recognition to discover knowledge from data)
Demographics
Econometrics (economic statistics)
Energy Statistics
Engineering Statistics
Epidemiology
Geographic Statistics
Psychological statistics
Reliability engineering
Social statistics
Sports statistics
Statistical literacy
Statistical modeling
Survey analysis
Structural data analysis
Survival analysis

The rapid and sustained increases in computing power starting from the 1950s have had a substantial impact on the practice of statistical science. Early statistical models were almost always from the class of linear models, but powerful computers, coupled with suitable numerical algorithms, have allowed researchers to explore nonlinear models and even create new model types such as generalized linear models and multilevel models.

Increased computing power has also led to the growing popularity of computationally-intensive methods. These methods are developed to solve for problems with complex unknowns whose solutions may depend on other unknowns in the model. Researchers in mathematical statistics study related issues such as estimation methods and algorithms to solve for those unknowns. There are also researchers who study Bayesian statistics, a school of thoughts that is based on the idea that existing knowledge about the population (called priors) should be used together with data from the sample to form new knowledge (called posteriors). The computer evolution coupled with powerful new statistical software has implications for the future of mathematical science with new emphasis on "experimental" and "empirical" statistics.