

STATISTICS and PROBABILITY

Definition

Statistics is the **science** and practice of developing **human knowledge** through the use of empirical **data** expressed in quantitative form.

It is based on **statistical theory** which is supposed to be a branch of applied **mathematics**.

Within statistical theory, **randomness** and **uncertainty** are modeled by **probability theory**.

Because one aim of statistics is to produce the "best" information from available data, some authors consider statistics a branch of **decision theory**.

Statistical practice includes the **planning, summarizing, and interpreting** of *observations*, allowing for variability and uncertainty.

Origin

The word *statistics* comes from

- the modern **Latin** phrase *statisticum collegium* (lecture about state affairs), from which came
- the **Italian** word *statista*, which means "**statesman**" or "**politician**" (compare to **status**) and
- the **German** *Statistik*, originally designating the analysis of data about the state. It acquired the meaning

of the collection and classification of data generally in the early nineteenth century.

The collection of data about states and localities continues, largely through [national and international statistical services](#); in particular,

[censuses](#) provide regular information about the [population](#).

Statistical methods

We describe our knowledge (and ignorance) mathematically and attempt to learn more from whatever we can observe. This requires us to

- [plan our observations](#) to control their variability ([experiment design](#)),
- [summarize a collection of observations](#) to feature their commonality by suppressing details ([descriptive statistics](#)), and
- reach consensus about what [the observations tell us](#) about the world we observe ([statistical inference](#)).

In some forms of descriptive statistics, notably [data mining](#), the second and third of these steps become so prominent that the first step (**planning**) appears to become less important. In these disciplines, data often are collected outside the control of the person doing the analysis, and the result of the analysis may be more an

operational model than a consensus report about the world.

Probability

The probability of an event is often defined as a number between one and zero.

In reality however there is virtually nothing that has a probability of 1 or 0.

Could you say that the sun will certainly rise in the morning?

Your answer may be “Yes”.

But what if an extremely unlikely event destroys the sun?

What if there is a nuclear war and the sky is covered in ash and smoke?

We often round the probability of such things up or down because they are so likely or unlikely to occur, that it's easier to recognize them as a probability of one or zero.

However, this can often lead to misunderstandings and dangerous behavior, because people are unable to distinguish between, e.g., a probability of 10^{-4} and a probability of 10^{-9} , despite the very practical difference between them. If you expect to cross the road about 10^5 or 10^6 times in your life, then reducing your risk of being run over per road crossing to 10^{-9} will make you

safe for your whole life, while a risk per road crossing of 10^{-4} will make it very likely that you will have an accident, despite the intuitive feeling that 0.01% is a very small risk.

Specialized disciplines

Some sciences use applied statistics so extensively that they have specialized terminology. These disciplines include:

Biostatistics

Business statistics

Economic statistics

Engineering statistics

Statistical physics

Demography

Psychological statistics

Social statistics (for all the *social* sciences)

Process analysis and chemometrics (for analysis of data from analytical chemistry and chemical engineering)

Reliability Engineering

Statistics form a key basis tool in business and manufacturing as well. It is used to understand measurement systems variability, control processes (as

in [statistical process control](#) or SPC), for summarizing data, and to make data-driven decisions. In these roles it is a key tool, and perhaps the only reliable tool.

Software

Modern statistics is supported by computers to perform some of the very large and complex calculations required.

Whole branches of statistics have been made possible by computing, for example [neural networks](#).

The computer revolution has implications for the future of statistics, with a new emphasis on 'experimental' statistics.

A list of statistical packages in common use

[R programming language](#)

[S programming language](#)

[MATLAB](#)

[GNU Octave](#)

[Macanova](#)

[MS Excel](#)

[OpenOffice Calc](#)

[SAS](#)

SPSS

STATA

MiniTab

StatPro - Free Excel add-in (*<http://bl-bus-dotnet3.ads.iu.edu/albrightbooks/Default.htm#Add-Ins>*)