Statistic overview

Statistics is a branch of mathematics that focuses on the scientific methodologies for collecting, organizing, summarizing, presenting, and analyzing data. It also involves drawing valid conclusions and making informed decisions based on such analyses. In today's world, statistics are omnipresent, encompassing areas such as economic statistics, geophysical statistics, employment statistics, accident statistics, financial statistics, population statistics, and many others.

Base Concept

- **Population**: refers to a set of individuals, objects, events, or measurements that share a common characteristic or variable of interest. It is the complete group relevant to a researcher's question or experiment. A statistical population can consist of a group of existing objects (such as a group of workers) or a hypothetical and potentially infinite group of objects, conceived as a generalization from experience (such as the set of all possible hands in a game of poker). A population is not static; it can change over time, as it represents a dynamic flow.
- **Statistical Unit**: A statistical unit is a single member of a population, the group of entities being studied. It is a key concept in statistics, serving as the source of data or information from which conclusions are drawn. Depending on the nature of the study, statistical units can be individuals, households, businesses, events, or other entities. In each statistical unit we observe and study variables that can be multiple.

Example

We can study the books in a library collection. For each book (statistical unit) we could examine multiple variables such as:

- Number of pages
- Publication year
- Genre
- and more...

- **Distribution**: refers to the way in which the values of a random variable are spread or arranged. It describes the frequency or probability of different outcomes in a dataset and provides insight into the underlying patterns or characteristics of the data. A distribution can be classified in two ways based on the nature of the random variable it describes:
  - In a *discrete distribution*, the random variable can take on a finite or countably infinite number of distinct values.
  - In a *continuous distribution*, the random variable can take on an infinite number of values within a given range.
- **Average**: The average is a measure of central tendency that indicates a typical value within a dataset. It serves to summarize the data using a single representative value.

- *Arithmetic* mean, is a measure of central tendency that calculates the total of a set of values divided by the number of values in that set. This yields a single value that encapsulates the data, representing a typical value within the dataset.

$$\text{Arithmetical mean} \quad = \quad \sum_{i=1}^{n} \frac{x_i}{n}$$

- *Median*: Is a more general concept. The median is the middle value when a data set is ordered from least to greatest.
- *Mode*: The **mode** is a measure of central tendency that represents the value that appears most frequently in a dataset. The mode focuses on frequency.

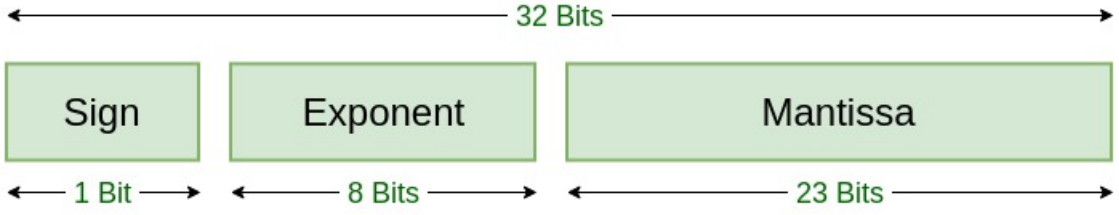
Computational problems with floating point representation

Floating Point Representation is a method used in computer systems to represent real numbers, allowing for a wide range of values by using a format that separates a number into its significant digits (the **mantissa**) and its exponent.

This type of representation is commonly used in statistics when we compute averages or perform other statistical calculations. Unfortunately, this type of representation has multiple problems that can affect the results:

- *Precision Issue*: In operations involving very large and very small numbers, precision may be lost.
- Rounding Errors: Floating point arithmetic can lead to rounding errors because not all decimal numbers can be represented exactly in binary. This is especially problematic in statistical calculations where precise values are essential.
- *Overflow and Underflow*: During calculations, particularly with large datasets or extreme values, you might encounter overflow (resulting in infinity) or underflow (resulting in zero). This can distort the results
- Accumulation of Errors: When performing multiple calculations, such as summing a large dataset, small rounding errors can accumulate, potentially leading to significant inaccuracies in the final result. This is known as numerical instability.

Structure



Single Precision
IEEE 754 Floating-Point Standard

- the *sign* indicates if the number is positive or negative
- *mantissa* is composed by the significant digits of the number
- *exponent* determinates the scale of the number

Numerical Solution and Knuth’s Contribution

Numerical solutions refer to mathematical techniques used to obtain approximate solutions to problems that may be difficult or impossible to solve analytically. **Donald Knut**, a renowned computer scientist, is famous for his significant contributions to algorithms and numerical analysis. His work laid the groundwork for many aspects of computer science and programming.
In his famous book called: "*The Art of Computer Programming*" provided treatments and techniques for minimazing errors. Below are a few of them:

- *Careful algorithm design*: Choose algorithms that maintain numerical stability, meaning they produce small changes in output for small changes in input.
- *Error Analysis*: Knuth underline the importance of errors analysis to understand how errors propagate through computations. By estimating the upper bounds of rounding errors, programmers can better gauge the reliability of their results.
- *Numerical representation*: Knuth provides guidelines for effectively using floating point representation, including how to avoid pitfalls such as overflow and underflow.
- *Adaptive methods*: Use adaptive algorithms that adjust their parameters based on the behavior of the data or the error observed in intermediate computations. This help to maintain accuracy.
- *Verification and Validation*: Implementing checks against known results or using multiple methods to solve the same problem can help verify the accuracy of computations. This technique can catch errors due to incorrect algorithm implementation or unexpected numerical behavior.
- *Data Structure choices*: Using data structures that minimize computational overhead and reduce the likelihood of errors during numerical operations is crucial.