Politecnico di Bari

Dipartimento di Ingegneria Elettrica e dell'Informazione

Corso di Laurea Triennale in Ingegneria dei Sistemi Medicali

Bioinformatics and Big Data Analytics

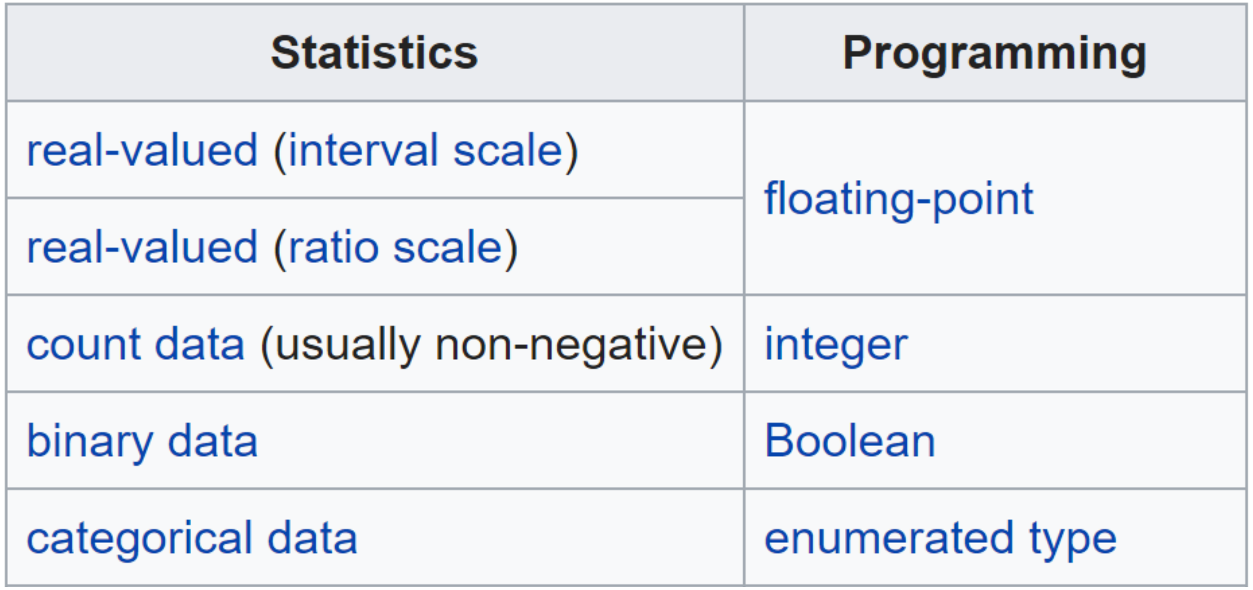
An Introduction to Multivariate Statistical Analysis

Eng. Nicola **Altini**, *Ph.D. Student*
Eng. Giacomo Donato **Cascarano**, *Ph.D. Student*
Prof. *Eng*. Vitoantonio **Bevilacqua**, *Ph.D.*

Anno Accademico 2019/2020

Data

- By data we mean observations made upon our environment – observations which are the result of measurements.
- These observations answer questions such as *how much*, *how many*, *how long*, *how often*, *how fast, etc.* and have the characteristics that they can be represented by numbers.

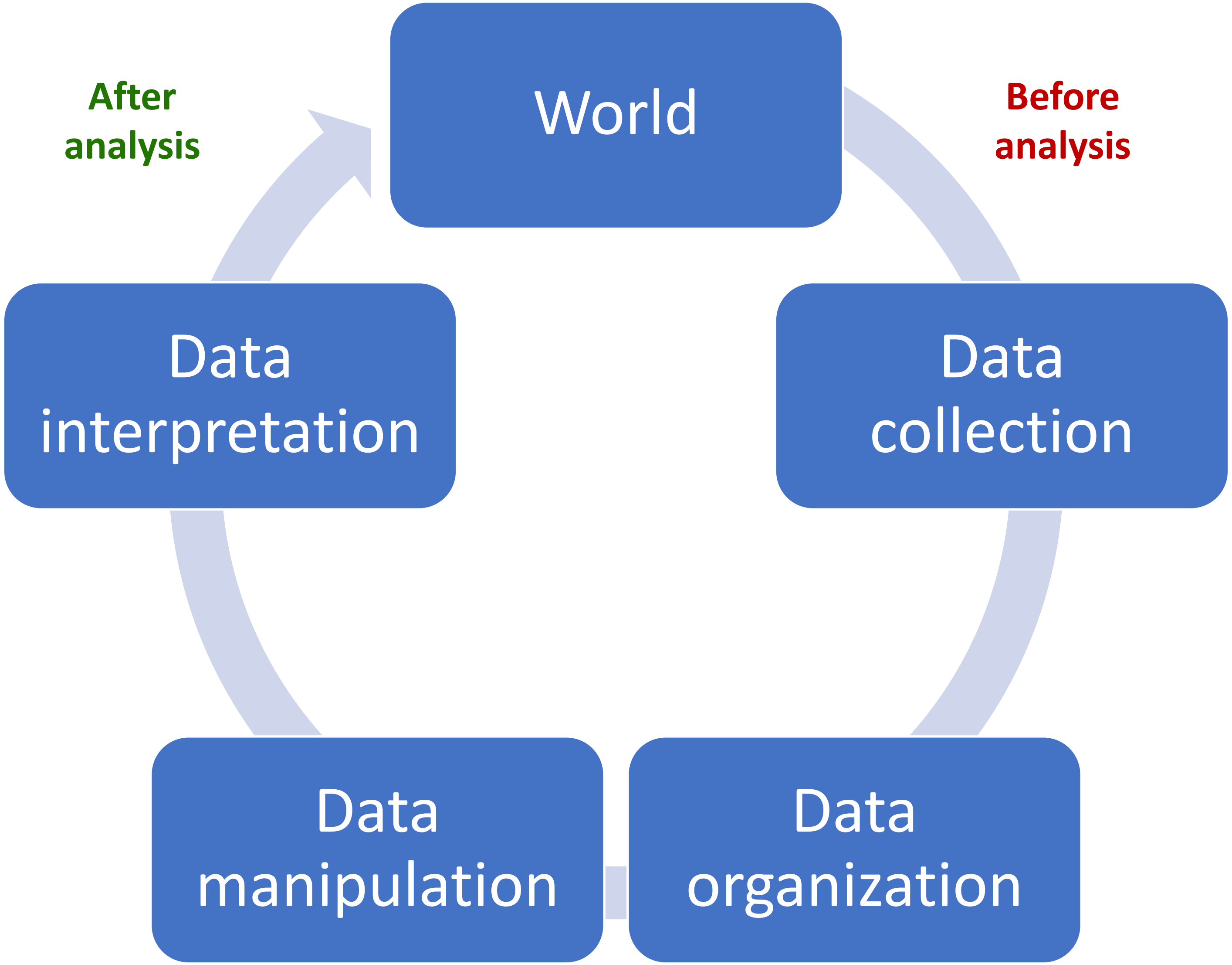


Statistics	Programming
real-valued (interval scale)	floating-point
real-valued (ratio scale)	
count data (usually non-negative)	integer
binary data	Boolean
categorical data	enumerated type

Objectives of statistical analysis

- ***Data Collection***
- ***Data organization*** and ***Data manipulation***
  - ***Data reduction***. It involves the summarization of data, in order to describe the original observations without sacrificing critical information.
  - ***Inference***. We want to use statistical analysis as an inferential measuring tool. All measurements are subject to error. Stastical analysis provides the methods for stating the degree of precision of our measurements.
  - ***Identification of relationships between and among sets of data***. Does knowledge about one set of data allow us to infer or predict characteristics about another set of data?
- ***Data interpretation***

Objectives of statistical analysis



World

Data collection

Data organization

Data manipulation

Data interpretation

Before analysis

After analysis

Descriptive vs inferential analysis

- Any particular statistical analysis can be characterized as **descriptive** or **inferential**, depending upon the nature of the data collection procedure and the objective of the data analysis. Before, we need to introduce the concepts of ***population*** and ***sample***.
- A ***population***, also referred as *universe*, is any well-defined collection of things. Well-defined means that the members of the population are spelled out, or are defined by an unequivocal statement.
- A ***sample***, is a subset of the *population*. Due to economical and practical constraints, we will make observations on a *sample* from that *population* and then make inferences about the population as a whole.
  - There are several techniques to take a sample from a population. A very common, simple and useful technique is *random sampling*.

Descriptive vs inferential analysis

- In the case of **descriptive analysis**, we are concerned with the direct exhaustive measurements of *population* characteristics. These defining characteristics fo population are called *parameters*.
- Observations must be made on every single member of the population in order to precisely state the value of the parameter.
- In the case of **inferential statistical analysis**, we are concerned with measuring the characteristics of only a *sample* from the population, and then making inferences, or estimates, about the corresponding value of the characteristics in the population from which the sample was drawn.
- A characteristic measured only on a sample is referred to as a *statistic*. A *sample statistic* is an estimate of *population parameter*.

Objects, Variables and Scales

- **Objects**. An object can be any data source we use for our statistical analysis. It can refer to individuals, physical or biological things, geographic locations, time periods, or events; that is, anything upon which observations can be made.
- **Variables**. The properties or characteristics of an object that can assume more different values are referred to as variables. The variables have meaning only with respect to objects. Examples of variables linked to objects are scholastic achievement of students, the effectiveness of medical treatments, the lifetime of a cell.
- **Scales**. A scale is a scheme for numerical representation of the values of a variable. We can think of a scale as a set of numbers.

Scales

- We can distinguish between four type of scales: *nominal*, *ordinal*, *interval* and *ratio*.
- **Nominal scales**. Many objects have characteristics that differ in *kind* only (for instance, think at the variable of gender). If we attach numbers to the alternative labels of this type of variable, the numbers have no meaning other than as a distinguishing label. Variables with nominal scale are referred as *categorical variables*.
  - *Qualitative variables*.
  - *Non-metric scale*.
- **Ordinal scales**. If the values of a variable can be rranged in a meaningful order, then such a variable can be represented by the numbers on an *ordinal scale*.
  - *Quantitative variables*.
  - *Non-metric scale*.
  - *Transitive property*.

Scales

- **Interval scales**. A step beyond the ordinal scale is the interval scale, in which equal differences between scale values have equal meaning.
  - *Quantitative variables*.
  - *Metric scale*.
  - *Transitive property*.
- A limitation of the interval scale, though, is that *ratios* of the scale values have no meaning.
- This is because this scale has an arbitrary zero point, one that does not really represent a zero quantity. So, the values occurring on the scale cannot be interpreted in any absolute sense. As an example, think about Fahrenheit and Celsius scales for temperature measuring.

Scales

- **Ratio scales**. This scale carries out more information of the previous ones. In fact, in addition to the transitive property of ordinal and interval scales, the ratio scales has the added property tht ratios of its values do have meaning, and equal ratios have equal meaning due to the presence of a genuine zero point on the scale.
  - *Quantitative variables*.
  - *Metric scale*.
  - *Transitive property*.
  - *Meaningful ratios*.
- Examples: a scale of seconds can be used to measure the variable of time duration.

Class of variables

- **Qualitative vs. quantitative**.
  - Qualitatively scaled variables
  - Quantitatively scaled variables
- **Discrete vs. continuous**.
  - A *discrete variable* only takes on a finite number of values (or countably infinite number of values).
  - A *continuous variable* is one which can take on an uncountable set of values.
- **Dichotomous variables**.
  - A variable which can take only two possible values. Also called *boolean variables* or *binary variables*.
  - An important kind of dichotomous variable is the *dummy variable*. It is created by converting a given level of a qualitative variable into a binary variable.

Notation

- **Univariate analysis**

$$
\begin{array}{cc}
Objects & Values\ on\ variable\ x \\
O^{(1)} & x^{(1)} \\
O^{(2)} & x^{(2)} \\
\vdots & \vdots \\
O^{(n)} & x^{(n)}
\end{array}
$$

- **Multivariate analysis**

$$
\begin{array}{cccccc}
Objects & x_1 & x_2 & \cdots & x_k \\
O^{(1)} & x_1^{(1)} & x_2^{(1)} & \cdots & x_1^{(1)} \\
O^{(2)} & x_1^{(2)} & x_2^{(2)} & & x_k^{(2)} \\
\vdots & \vdots & \vdots & \cdots & \vdots \\
O^{(n)} & x_1^{(n)} & x_2^{(n)} & & x_k^{(n)}
\end{array}
$$

Statistical Inference

- The problem of statistical inference – the drawing of conclusions about populations based on sample observations – can be approached from two different directions.
- **Parameter Estimation**. On the one hand, the population parameters can be estimated directly, with the sample observations serving as approximation to the true population characteristic.
- **Hypothesis Testing**. On the other hand, the sample observations can serve to support or discredit (i.e., **test**) *a priori hypotheses* about the population.

Point Estimation

- For a given population there may be any number of different parameters in which we might be interested – e.g., its mean, median, mode, variance, standard deviation, range, etc.
- One approach to this estimation problem is to obtain a single value based upon a sample of observations, a value which we feel is the best possible approximation of the true value of the population parameter.
- This type of point estimate does not provide us with information as to how close we might expect it to be the population parameter, but it does at least suggest its value.
- Examples:
  - Mean of a sample $\bar{x}$ is a point estimate of the population mean μ.
  - Sample standard deviation s and variance s^2 are respective estimates of the population standard deviation σ and variance σ^2.

Sample Mean, Variance and Median

- Sample Mean

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x^{(i)}$$

- Sample Variance

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(x^{(i)} - \bar{x}\right)^2$$

- Median: is the value separating the higher half from the lower half of a data sample.

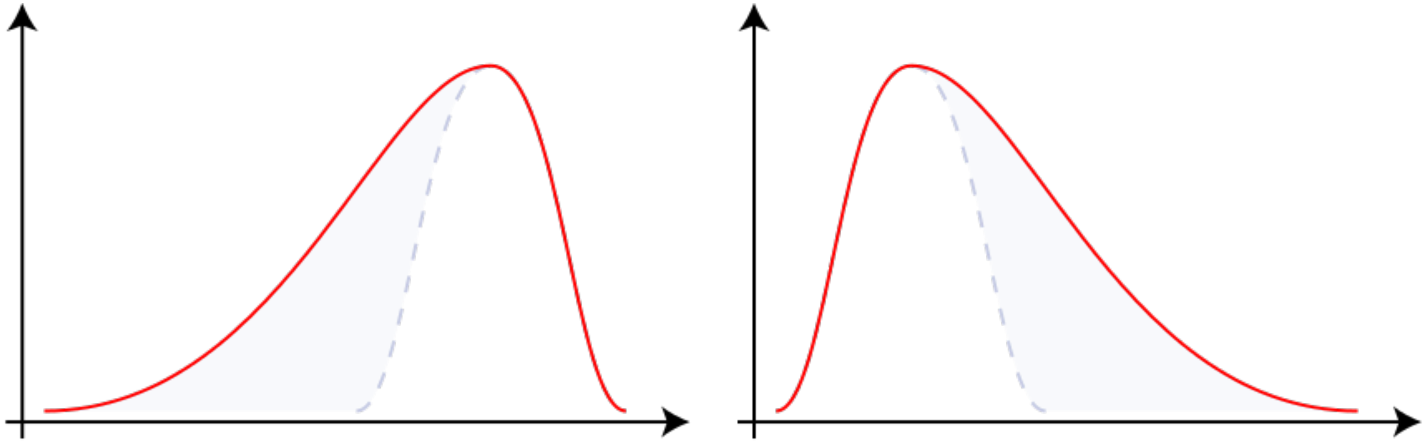
Point Estimation

- We want a point estimate to be an *unbiased* estimate of the population parameter in question. By unbiased we mean that the long run average or *expected value* of the estimate will equal the population parameter.
  - Note that this property is certainly true for sample mean, sample variance and sample standard deviation, but it is not valid for the range. In fact, the range of a small sample would tend to underestimate the range of the parent population from which the sample was drawn.
- We also want a point estimate to be relatively efficient compared to alternative point estimates. By efficient we mean the speed with which the estimate becomes a more accurate estimate of the population parameter as the sample size increases.

Sample skewness

- Sample skewness

$$b_1 = \frac{m_3}{s^3} = \frac{\frac{1}{n}\sum_{i=1}^{n}\left(x^{(i)} - \bar{x}\right)^3}{\left[\frac{1}{n-1}\sum_{i=1}^{n}(x^{(i)} - \bar{x})^2\right]^{\frac{3}{2}}}$$



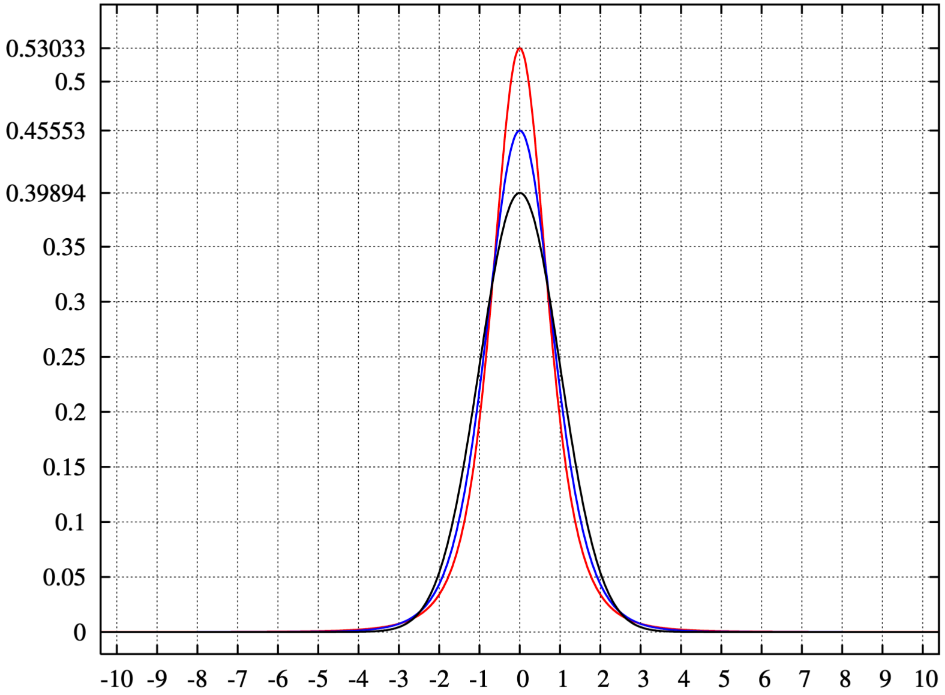
Negative Skew

Positive Skew

Sample Excess Kurtosis

- Sample Excess Kurtosis

$$g_2 = \frac{m_4}{m_2^2} - 3 = \frac{\frac{1}{n}\sum_{i=1}^{n}\left(x^{(i)} - \bar{x}\right)^4}{\left[\frac{1}{n}\sum_{i=1}^{n}\left(x^{(i)} - \bar{x}\right)^2\right]^2} - 3$$



Probability Density Function for the Pearson type VII distribution with excess kurtosis of **infinity (red)**; **2 (blue)**; and **0 (black)**

Interval Estimation

- A point estimation of a population parameter is better than no estimate at all, but we prefer to have some idea of how close the true parameter might be to the estimate
- We could state with a certain probability that it is within a particular distance of the parameter. Such an estimate is referred to as an **interval estimate**, an interval of values within which we can state with a certain degree of confidence (i.e., probability) that the parameter falls.
- For example, rather than using a sample mean $\bar{x}$ as a point estimate of the population mean μ, we could state an interval of values within which we strongly believe the true mean to fall, such as:

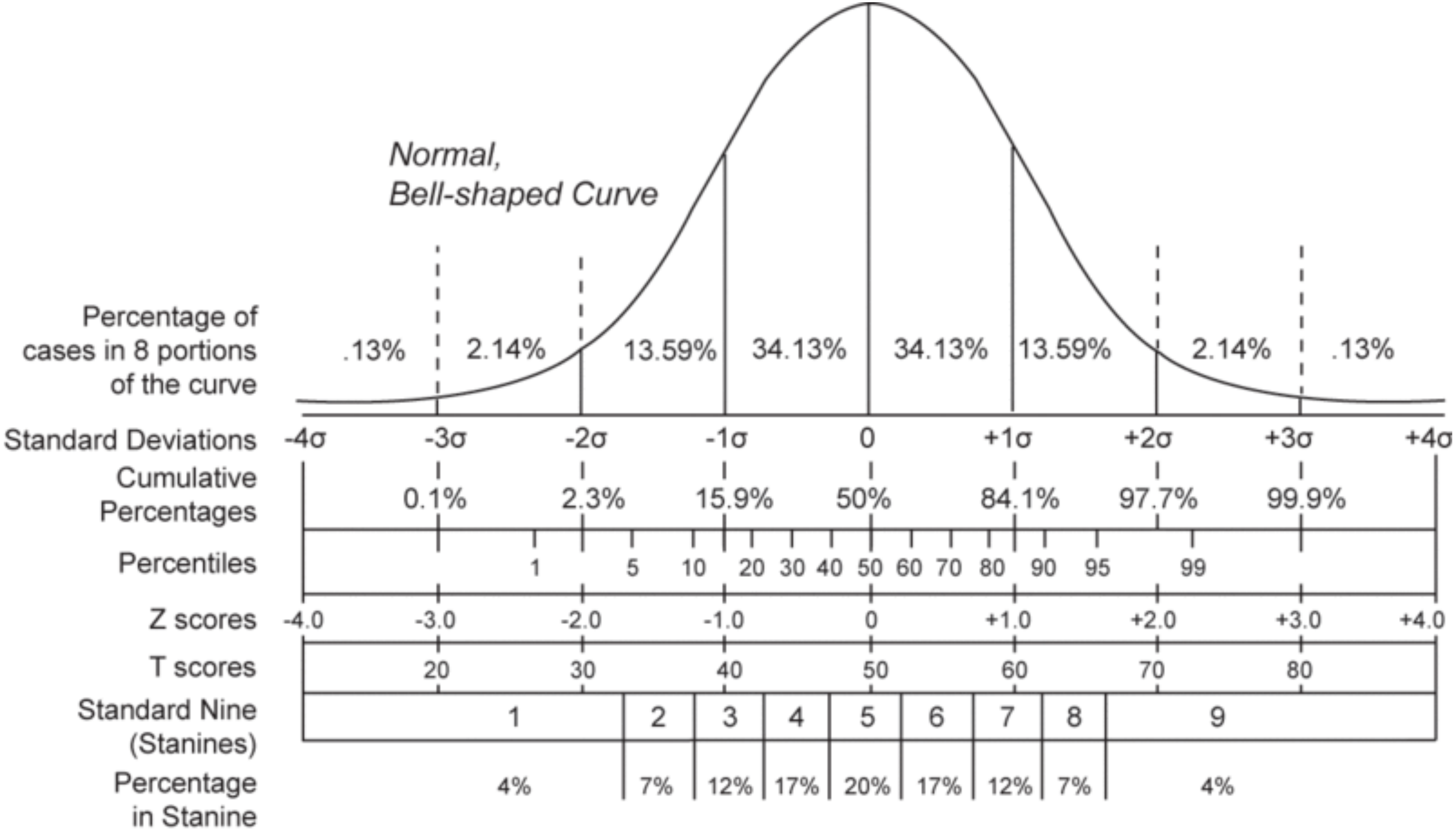
$$a < \mu < b$$

- The key phrase is *«probably lies between»*.

Confidence intervals and limits

- The two key elements of an interval estimate, then, include the specification of the limits or boundary values of the interval, and the associated probability that the population parameter is contained in that interval of values.
- This interval of values is referred to as ***confidence interval***, and the boundary values which define it are referred to as ***condifence limits*** of the interval.
- For each confidence interval there is an associated probability indicating how certain we are that the population parameter falls within the interval,

Confidence intervals and limits



Hypothesis Testing

- Hypothesis testing is an indirect approach to the problem of statistical inference. Rather than using our sample observations to derive statistics which are approximations of the population parameter in question, we will use our samle statistics to support or discredit *a priori hypotheses*, or speculations, about the true value of the population parameter.
- The hypotheses about the population parameters that we wish to test can be based either on prior observations or on theoretical grounds. Whatever is the basis for the hypothesis, our sample observations will be used to test the likelihood or tenability of its being true.
- If it is forced to be untenable, from a probability point of view, then we are forced to believe in alternative hypothesis.

References

- Sam Kash Kachigan. *Multivariate Statistical Analysis: A Conceptual Introduction*. Radius Press, 1991.
- Wikipedia
  - Data type.
    URL: https://en.wikipedia.org/wiki/Data_type
  - Continuos or discrete variable.
    URL: https://en.wikipedia.org/wiki/Continuous_or_discrete_variable
  - Multivariate statistics.
    URL: https://en.wikipedia.org/wiki/Multivariate_statistics