Friedrich-Alexander-Universität
Department Artificial Intelligence
in Biomedical Engineering | AIBE

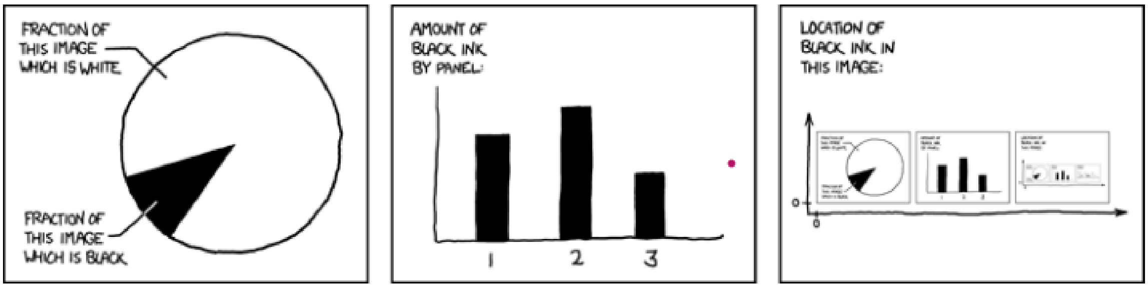
Machine Learning
Data Analytics

FAU

Chapter 6:

Descriptive statistics

Overview

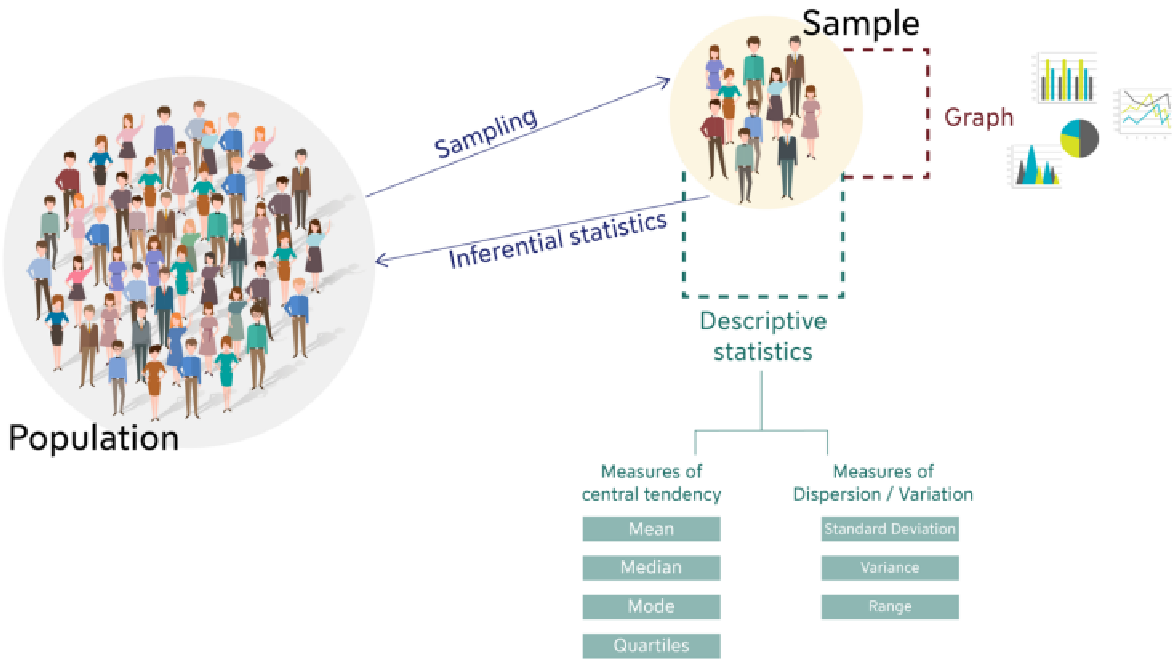

FRACTION OF THIS IMAGE WHICH IS WHITE
FRACTION OF THIS IMAGE WHICH IS BLACK
AMOUNT OF BLACK INK BY PANEL:
1
2
3
LOCATION OF BLACK INK IN THIS IMAGE:
0
0

Introduction to Descriptive statistics

Descriptive statistics **describe**, **show**, and **summarize** the basic features of a dataset found in a given study, presented in a summary that describes the data sample and its measurements. It helps analysts to understand the data better.

Descriptive statistics **represent the available data sample** and does not include theories, inferences, probabilities, or conclusions. That’s a job for inferential statistics.

Descriptive statistics summarize findings.

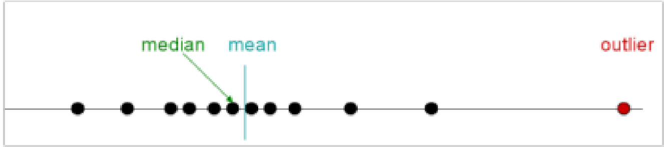


There are two o ways to summarize data:
- Exploratory data analysis (graphs, ...)
- Statistics

Summarizing the center

We identify the central position of any data set while describing a set of data. This is known as the measure of central tendency. A measure of central tendency describes a set of data by identifying the central position in the data set as a single value. There are three ways how to summarize the central tendency of a dataset, which are **mean** (average score), **median** (middle score) and **mode** (most frequent score). In different ways they each tell us what value in a data set is typical or representative of the data set.

Choosing the best measure of central tendency depends on the type of data we have. Therefore, it is important to understand the strengths and limits of each measure.



1. Mean

The arithmetic mean (short mean) is the same as the average value of a data set. The mean of a given data is the sum of all observations divided by the number of observations:

$$m = \frac{1}{n} \sum_{i=1}^{n} Xi$$

Example: [4, 12, 34, 15, 49, 50, 24, 11] m=(4+12+34+15+49+50+24+11)/8 = 24,875

Pros:	**Cons:**
• Considers every score ➔ most accurate summary of data • Resistant to sampling variation: removing one sample changes the mean far less than mode or median	• Affected by extreme scores and skewed distributions • Can only be used with interval and ratio data

2. Median

The value of the middlemost observation, obtained after arranging the data in ascending or descending order, is called the median of the data. Ordering a data set $x1 \leq x2 \leq x3 \leq \ldots \leq x_n$ from lowest to highest value, the median $x^{\sim}$ is the data point separating the upper half of the data values from the lower half.

If n is odd: $x_{(n+1)/2}$

If n is even: $(x_{n/2} + x_{n/2+1})$ / 2

Example 1: [4, 46, 55, 1, 13, 10, 20]
-> sorting order [1, 4, 10, 13, 20, 46, 55]

$x^{\sim}$ = 13

Example 2: [4, 12, 34, 15, 49, 50, 24, 11]
-> sorting order [4, 11, 12, 15, 24, 34, 49, 50]

$x^{\sim}$ = (15+24)/2 = 19,5

Pros:	**Cons:**
• Relatively unaffected by outliers (very low or high scores) and skewed distributions	• Does not consider all scores of the data set
• Can be used with ordinal, interval, and ratio data	• Not very stable

3. Mode

The value which appears most often in the given data i.e., the observation with the highest frequency is called a mode of data.

Case 1: Ungrouped Data

To identify the mode in ungrouped data, the mode is the observation which occurs maximum times. For example, in the data: 6, 8, 9, 3, 4, 6, 7, 6, 3, the value 6 appears the greatest number of times. Thus, mode = 6. An easy way to remember mode is: Most Often Data Entered. Note: A data may have no mode, 1 mode, or more than 1 mode. Depending upon the number of modes the data has, it can be called unimodal, bimodal, trimodal, or multimodal.

Case 2: Grouped Data

When the data is continuous, the mode can be found using the following steps:

Step 1: Find modal class i.e., the class with maximum frequency.
Step 2: Find mode using the following formula:

$$Mode = I + \left(\frac{f_m - f_1}{2f_m - f_1 - f_2}\right) h$$

With: l = lower limit of modal class, f_m = frequ. of modal class, f_1 = frequ. of class preceding modal class, f_2 = frequ. of class succeeding modal class, h = class width

Pros:	**Cons:**
• Easy to calculate and understand • Can be used with nominal data	• There can be more than one mode • Mode can change dramatically by adding only one dataset • Independent of all other data in the set

Frequency distributions

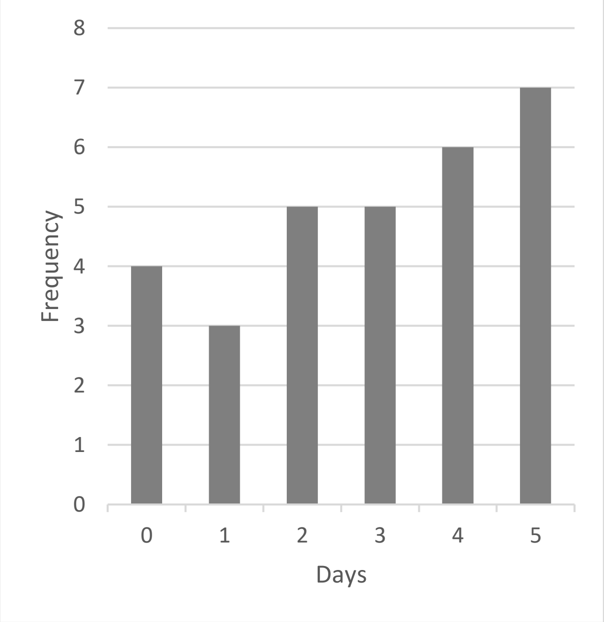
A frequency distribution is an overview of all distinct values in some variable and the number of times they occur.

The first step in drawing a frequency distribution is to construct a **frequency table**. A frequency table is a way of organizing the data by listing every possible score (including those not actually obtained in the sample) as a column of numbers and the frequency of occurrence of each score as another. Computing the frequency of a score is simply a matter of counting the number of times that score appears in the set of data. It is necessary to include scores with zero frequency to draw the frequency polygons correctly.

Example: days needed to answer my email
Data: 5 2 2 3 4 4 3 2 0 3 0 3 2 1 5 1 3 1 5 5 2 4 0 0 4 5 4 4 5 5

The frequency table gives the occurrence of each value:

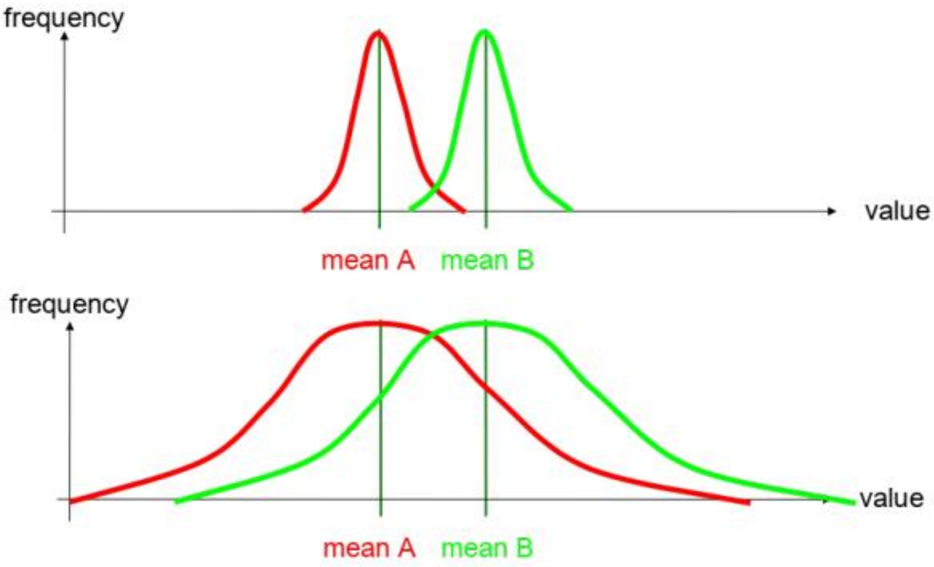
Days	Frequency	Frequency (%)
0	4	13 %
1	3	10 %
2	5	17 %
3	5	17 %
4	6	20 %
5	7	23 %



Measures of variability

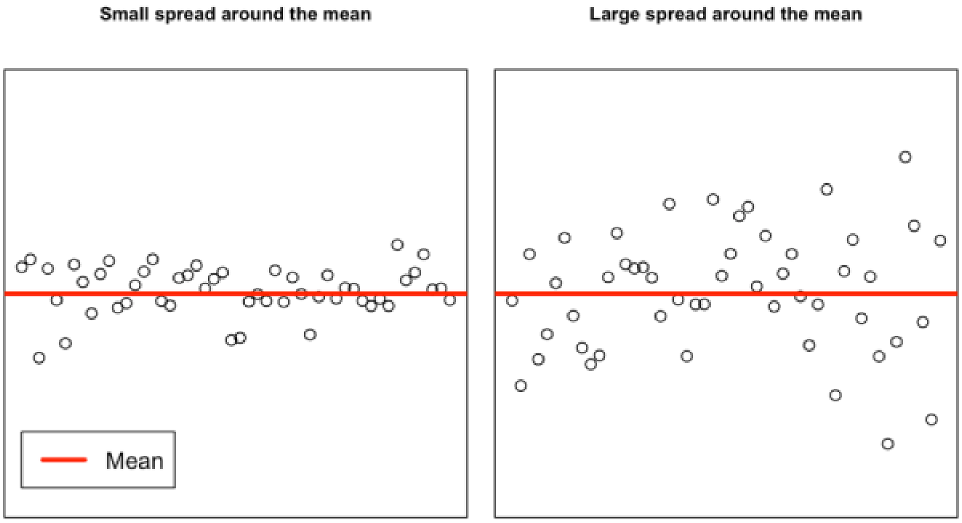
When looking at the following dataset: [5, 10, 10, 15, 9, 10, 10, 11], the mode, median and mean all equal 10. Thus, even if they are all different measures, they can be the same in a given dataset.

Thus, further classifications of a dataset are necessary to measure the accuracy of the mean of a dataset.



1. Standard deviation and variance

Variance is a measure of how data points vary from the mean, whereas standard deviation is the measure of the distribution of statistical data. Both variance and standard deviation measure the **accuracy** and the **variability** of the dataset. The basic difference between variance and the standard deviation is in their units. The standard deviation is represented in the same units as the mean of data, while the variance is represented in squared units.



If $x1, x2, \ldots \mathrm{x_n}$ are the data in a sample with mean m:

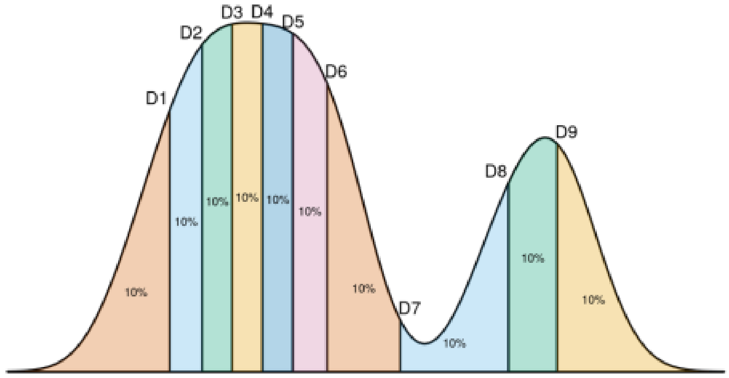
- Mean Deviation = $1/\mathrm{n} * \sum(\mathrm{x_i}\text{-}\mathrm{m})$ (= Difference between mean and scores)
- Variance $\mathrm{s}^2 = 1/\mathrm{n} \sum(\mathrm{x_i}\text{-}\mathrm{m})^2$ $(= \sigma^2)$
- Standard deviation s = $\sqrt{\mathrm{Variance}}$ $(= \sigma)$

The variance and standard deviation are important in statistics because they serve as the basis for other types of statistical calculations. For example, the standard deviation is necessary for converting test scores into Z-scores. The variance and standard deviation also play an important role when conducting statistical tests such as t-tests.

> The variance and standard deviation are important because they tell us things about the data set that we can't learn just by looking at the mean, or average.
>
> As an example, imagine that you have three younger siblings: one sibling who is 13, and twins who are 10. In this case, the average age of your siblings would be 11. Now imagine that you have three siblings, ages 17, 12, and 4.
>
> In this case, the average age of your siblings would still be 11, but the variance and standard deviation would be larger.

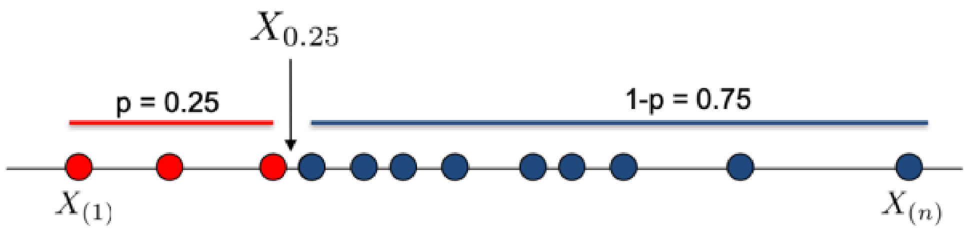
2. Quantiles

Quantiles belong to the measures of position in statistics. They divide a certain amount of data so that one part p is less than or equal to and the other part 1-p is greater than or equal to the quantile.



A p-quantile is defined as the x value of the distribution which includes p*N observations, with 0<p < 1 and N being the number of observations.

There are a few special quantiles which have their own definition: **Quartiles** refer to quarters of the distribution, deciles to the tenth part:



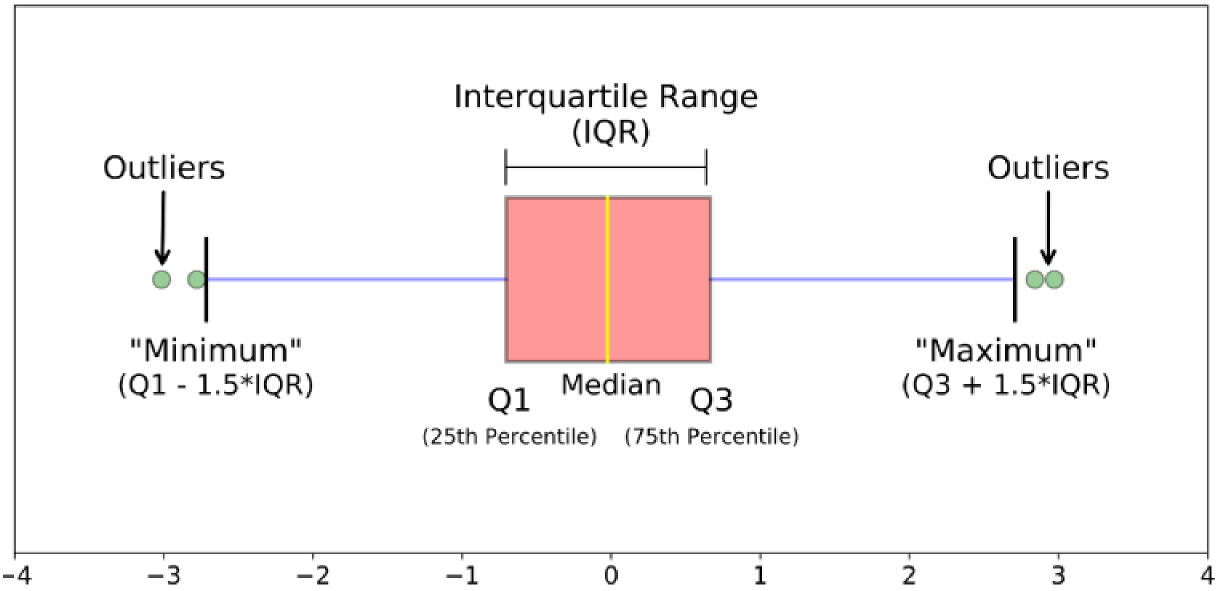
A **percentile** is one of the equal divisions of an amount, expressed on a scale from 0 to 100. The 90th percentile of an amount is all amounts between zero percent and ninety percent.

The **median** is a special quantile, that divides a sample of data into 2 groups containing equal numbers of observations.

3. Boxplots

To visualize the dataset and its components, boxplots can be used for a representation of the distribution in the dataset.

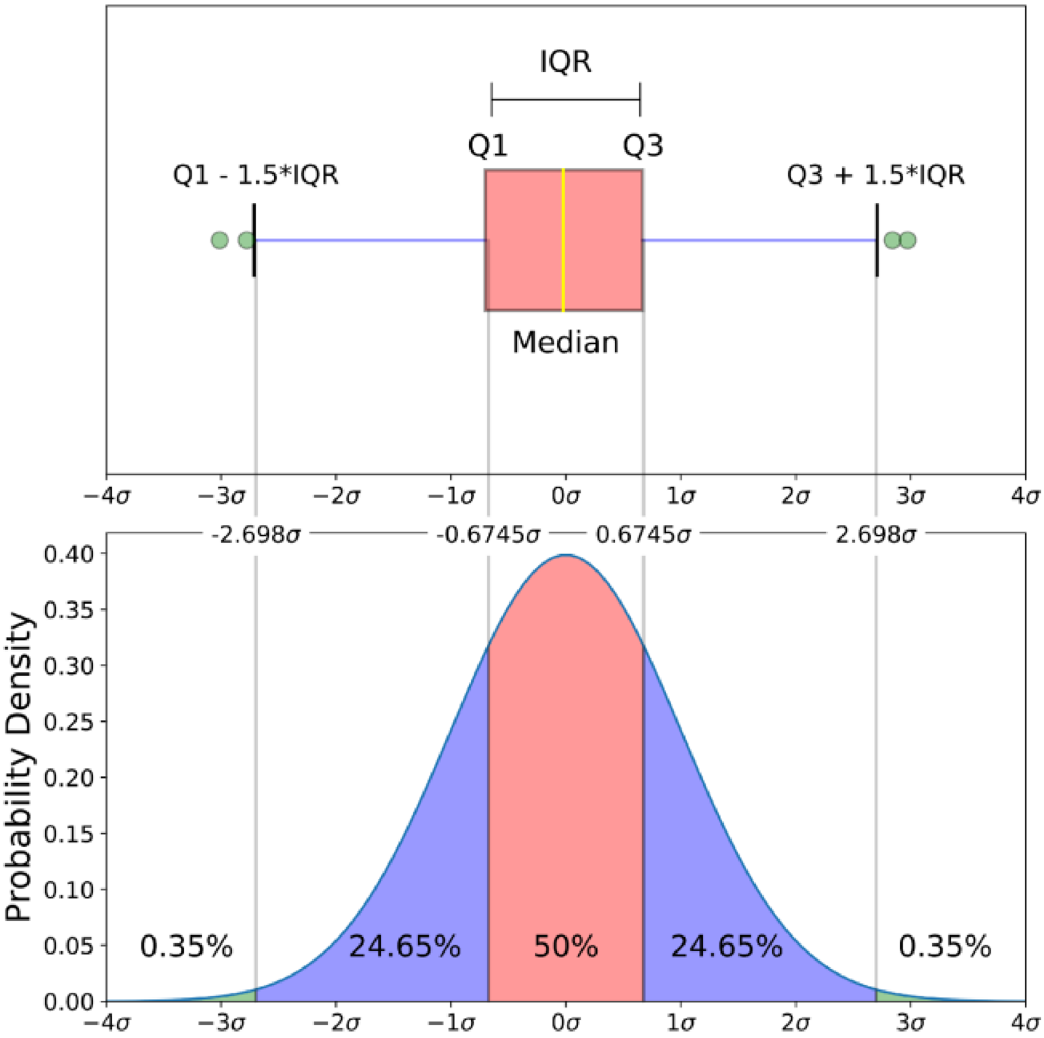
The box plot, also called box-whisker plot or box graph is a diagram that allows the clear presentation of the most important robust measures of location and dispersion. The minimum, the lower quartile, the median, the upper quartile, and the maximum are shown.[1]



Boxplots consist of:

- Minimum (Q0 or 0th percentile): the lowest data point in the data set excluding any outliers
- Maximum (Q4 or 100th percentile): the highest data point in the data set excluding any outliers
- Median (Q2 or 50th percentile): the middle value in the data set
- First quartile (Q1 or 25th percentile) • Third quartile (Q3 or 75th percentile)
- Interquartile range (IQR) the distance between the upper and lower quartiles

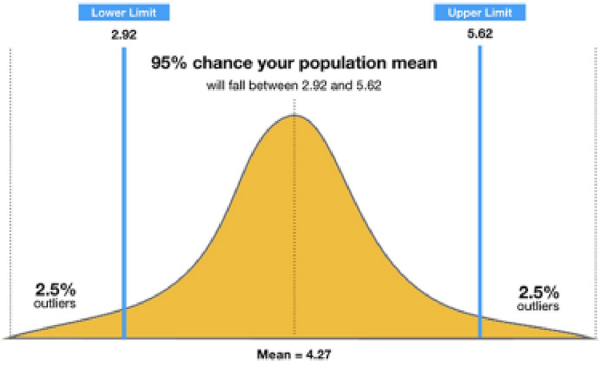
A box-plot usually includes two parts, a box and a set of whiskers as shown. The box is drawn from Q1 to Q3 with a horizontal line drawn in the middle to denote the median. The whiskers can be defined in various ways.



4. Confidence interval

A confidence interval refers to the probability that a population parameter will fall between a set of values for a certain proportion of times.

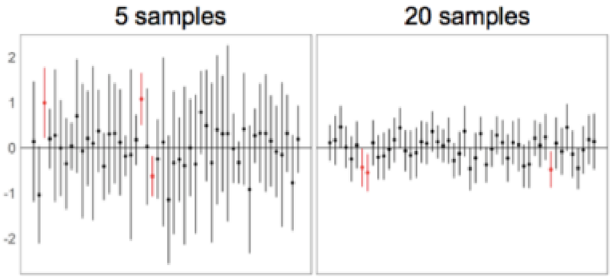
Confidence intervals measure the degree of uncertainty or certainty in a sampling method. They can take any number of probability limits, with the most common being a 95% or 99% confidence level. Confidence intervals are conducted using statistical methods, such as a t-test.



A confidence interval is a range of values, bounded above and below the statistic's mean, that likely would contain an unknown population parameter. Confidence level refers to the percentage of probability, or certainty, that the confidence interval would contain the true population parameter when you draw a random sample many times. Or, in the vernacular, "we are 99% certain (confidence level) that most of these samples (confidence intervals) contain the true population parameter."

$$CI = [\overline{X} - t_{df}\frac{s}{\sqrt{n}}; \overline{X} + t_{df}\frac{s}{\sqrt{n}}]$$

Example: performing k = 50 experiments



- ➔ True mean (x = 0, s = 1) will be covered by 95% CI
- ➔ Smaller CI if you take more samples in an experiment
- ➔ CI is smaller for high number of samples n

References

[1.] Galarnyk M. Understanding Boxplots. Medium. Published July 6, 2020. Accessed June 8, 2022. https://towardsdatascience.com/understanding-boxplots-5e2df7bcbd51

Carmines, E. and Zeller, R. (1979). Reliability and Validity Assessment. Newbury Park: Sage Publications

Colosi, L (1997) The Layman's Guide to Social Research Methods http://www.wiley.com/college/westen/0471387541/instructor/ch02/ar_02.html

Field, A. and Hole, G. (2003). How to Design and Report Experiments. Sage Publications

Carmines, E. and Zeller, R. (1979). Reliability and Validity Assessment. Newbury Park: Sage Publications

Field, A. and Hole, G. (2003). How to Design and Report Experiments. Sage Publications

Alan Dix, Janet Finlay, Gregory Abowd and Russell Beale. (1998) Human Computer, Interaction (second edition), Prentice Hall, ISBN 0132398648 (new Edition announced for October 2003)

Ben Shneiderman. (1998) Designing the User Interface, 3rd Ed., Addison Wesley; ISBN: 0201694972

Discount Usability Engineering http://www.useit.com/papers/guerrilla_hci.html

Heuristic Evaluation http://www.useit.com/papers/heuristic/

ISO 13407 Chapter 9, www.id-book.co