Business Intelligence

Exploratory Data Analysis by Methods of Descriptive Statistics

2

tdi@cphbusiness

Intended Learning Objectives

To find out

- What is Descriptive Statistics?
- How does it matter for BI?
- Which instruments of Descriptive Statistics are used for exploratory analysis of data in BI?

Agenda

Introduction to Descriptive Statistics

- Data distribution
- Data measures
  - Measures of central tendency
  - Measures of variability
- Correlation

Definition

Descriptive statistics refers to the process of summarizing numerical and categorical data in a concise and informative manner. It involves using various measures, such as measures of center, variability, shape, and location, to describe key features of the data. Descriptive statistics form the foundation for quantitative analysis and provide insights into the distribution and characteristics of a dataset.

- *AI generated definition based on: International Encyclopedia of Education(Fourth Edition), 2023*
- *Source: https://www.sciencedirect.com/topics/social-sciences/descriptive-statistics*

Data Distribution

Distribution of Frequency

Frequency Distribution

How many times certain value appears in the observations?

Discrete

Individually separated and distinct domain values

Example: a family could have 3 or 6 children, but not 4.5!

Color	Frequency
Brown	17
Red	18
Yellow	7
Green	7
Blue	2
Orange	4

Frequency Table

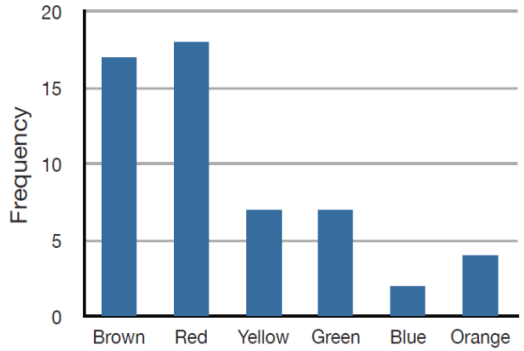


Figure 1. Distribution of 55 M&M's.

Histogram

Table 2. Response Times

568	720
577	728
581	729
640	777
641	808
645	824
657	825
673	865
696	875
703	1007

Table 3. Grouped frequency distribution

Range	Frequency
500-600	3
600-700	6
700-800	5
800-900	5
900-1000	0
1000-1100	1

bins

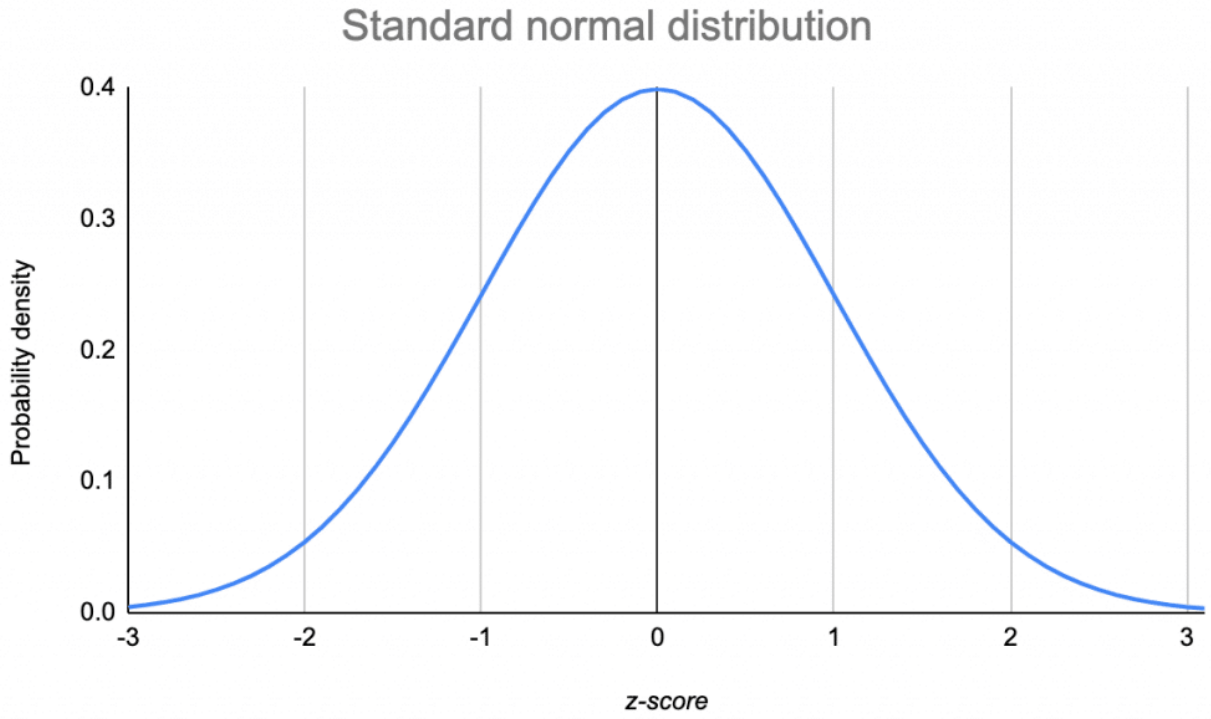
Continuous

Values forming an unbroken and uninterrupted domain

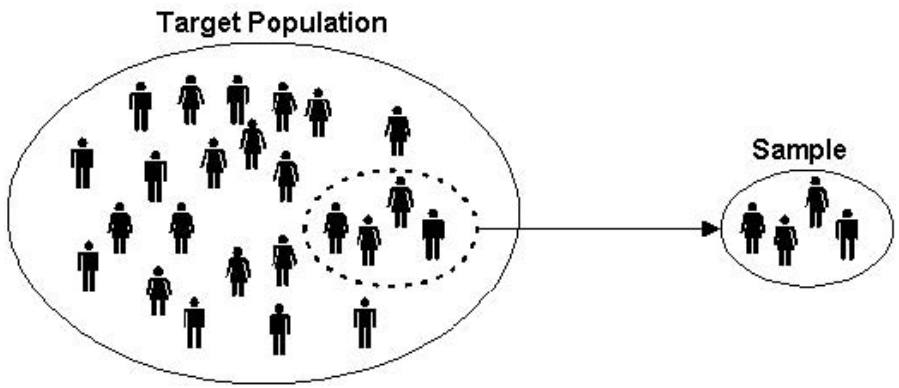
Example: response time – we can record 1.6 sec or 2378765 sec

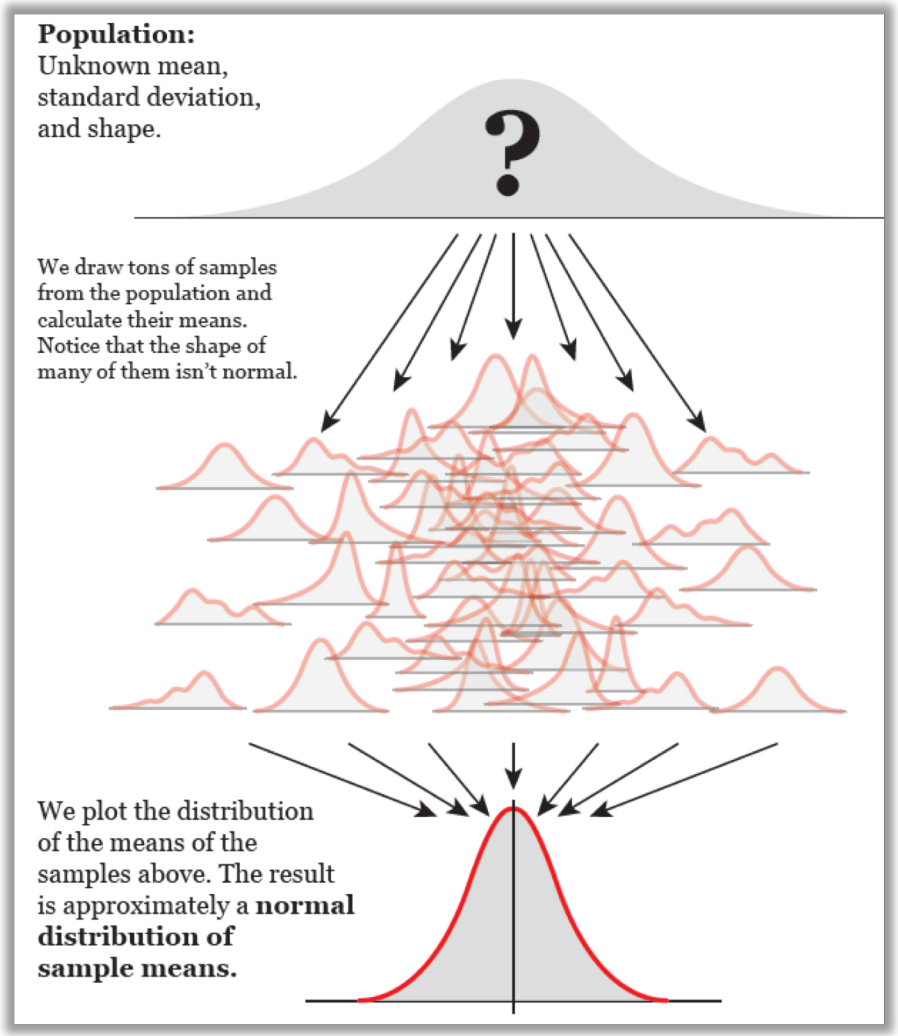
Normal Distribution
Gaussian Distribution

Normal distribution, also known as the Gaussian distribution, is a probability distribution that is symmetric about the mean, showing that data near the mean are more frequent in occurrence than data far from the mean.
The normal distribution appears as a "bell curve" when graphed.

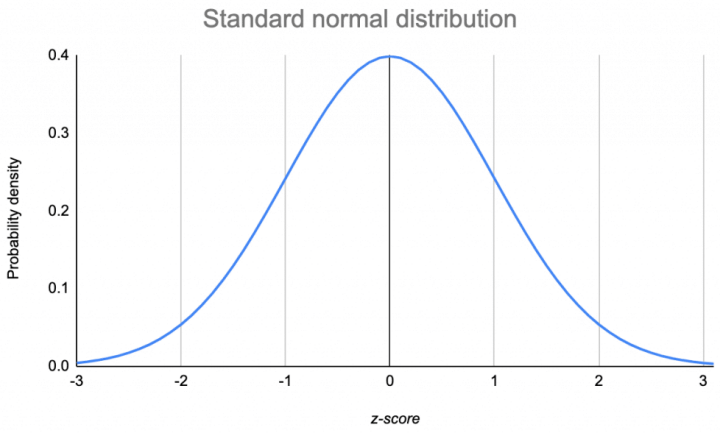


Normal Distribution





Descriptive Statistics

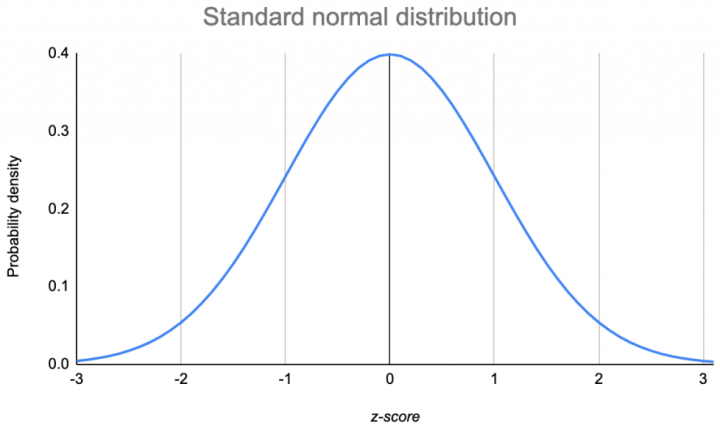


- A descriptive statistic summarizes and quantitatively describes the features of a data set
- Descriptive statistics is the process of using and analysing those statistics

Statistical Methods

- Study of central tendency in data sample
  - which is the centre of the sample, around which is the largest amount of data
  - calculate the measures of the central tendency – mean, media, mode
- Study of the distribution of the values outside the centre
  - measures of variability
  - interval, quartiles, dispersion, standard deviation
- Graphical data representation

Statistical Measures



Descriptive statistics consists of three basic categories of measures:

- frequency distribution - describe the occurrence of data within the data set (`count`)
- measures of central tendency - describe the centre of the data set (`mean, median, mode`)
- measures of variability or spread - describe the dispersion of the data set (`variance, standard deviation`)

Central Tendency Measures

Cphbusiness

Central Tendency Measures

Mean

$$\bar{x} = \frac{\sum_{i=1}^{N} x_i}{N} = \frac{x_1 + x_2 + \cdots + x_N}{N}.$$

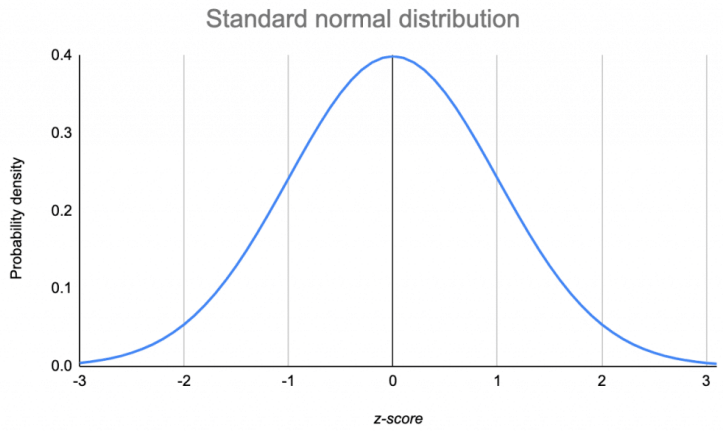
where x1, x2, ..., xN - values
N - size

Mode

the most frequent value in the sample

- Trimmed mean – removes outliers before calculating the mean
- Weighted mean – data points contribute differently to the sum

Central Tendency Measures



Median

$$median = L_1 + \left(\frac{N/2 - (\sum freq)_l}{freq_{median}} \right) width,$$

the value that divides the sorted sample into two equal halves

Cphbusiness

Data Aggregation and Granularity

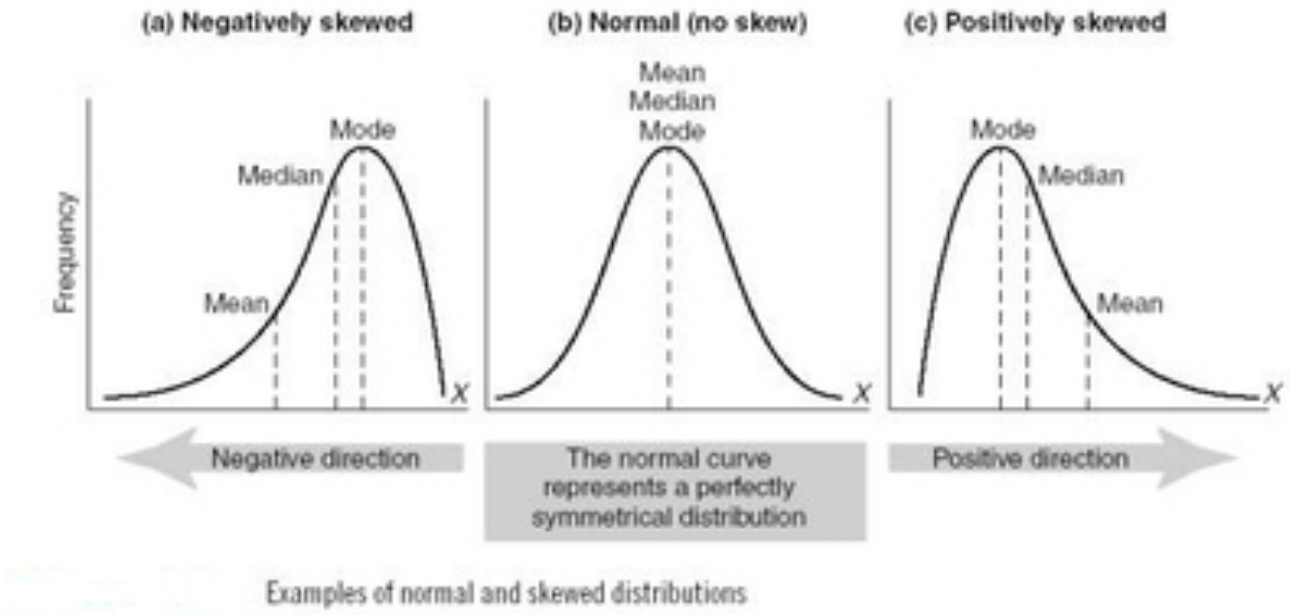
Descriptive statistics - mean, mode, median

- are form of aggregation
- while individual data points are shown by disaggregation

- Aggregation refers to how data is combined
  - can show the larger trends in the data
  - quantitative fields are aggregated

- Granularity refers to how detailed the data is
  - low and high granularity

Normal vs Skewed Distribution



Examples of normal and skewed distributions

For non symmetrical distributions empirically found:

`mean - mode = 3 * (mean - median)`

Exercise

- Here is a set of numbers representing the test results of one class of students:

24, 18, 12, 15, 19, 17, 17, 19, 21, 19, 18, 16, 22, 19, 20

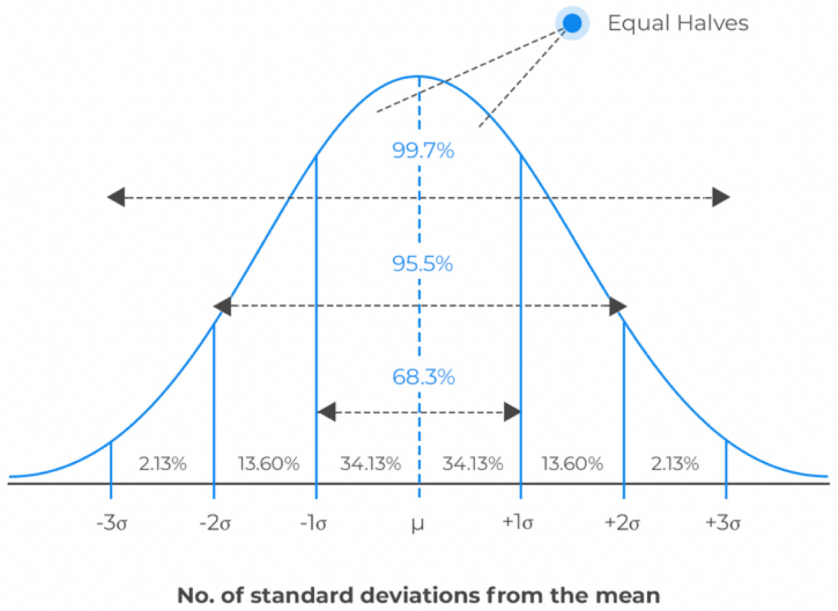
- Calculate the set's
  - size
  - sum
  - min
  - max
  - mean
  - median
  - mode
- Draw a bell curve plot representing the set

Measures of Variability

Variability Measures

Range R

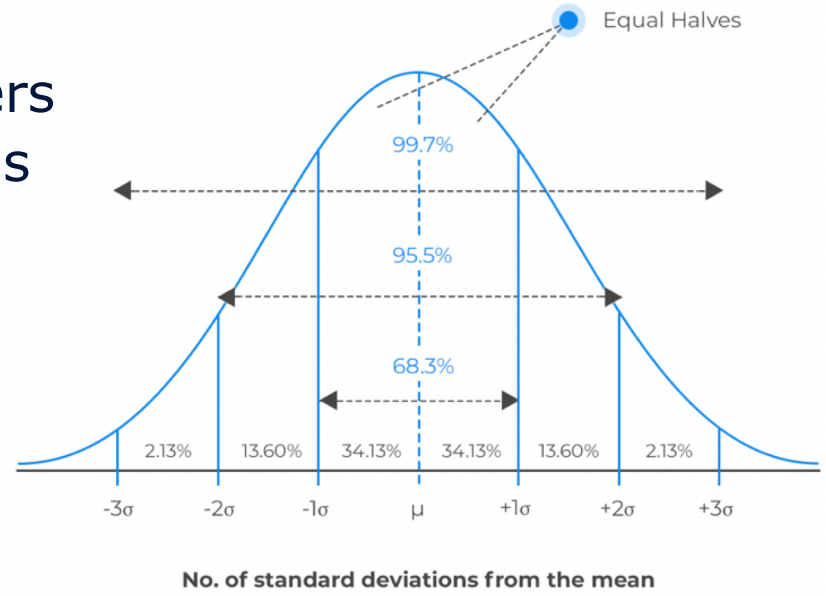
- difference between the max and min values



Variability Measures

Deviation S

- the amount by which a single measurement differs from a fixed value such as the mean
- standard error



Variance V

- based on the squared deviation
- $S^2 = V$
- $S = \sqrt{V}$

Degree of Freedom

n vs N

- make estimates for the whole population based on a sample
- bias

n - 1

- considers one constrain – the sample's mean
- reduces the bias

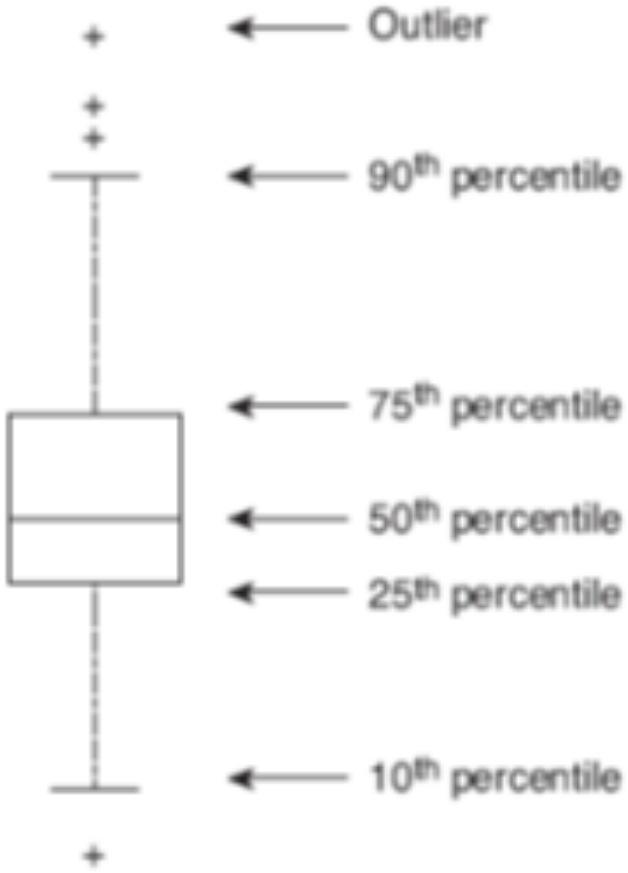
The "**Population** Standard Deviation":

$$\sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2}$$

The "**Sample** Standard Deviation":

$$s = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(x_i - \overline{x})^2}$$

Quantiles

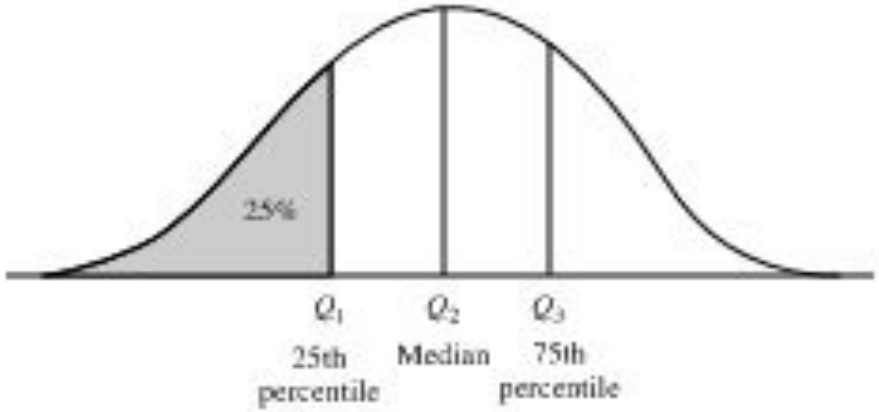


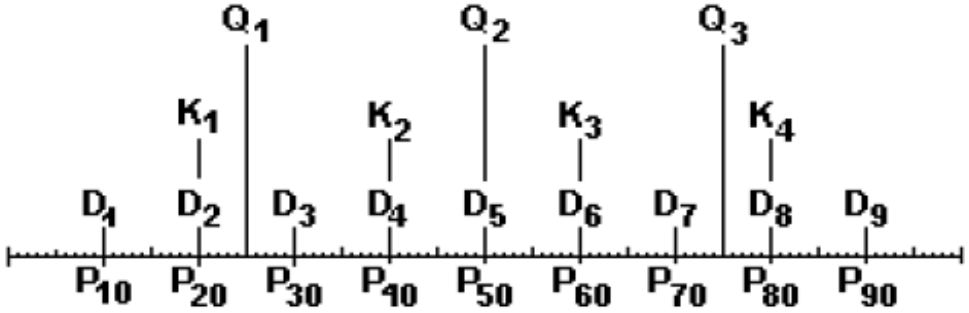
Q – quartiles

K – quintiles

D – deciles

P – percentiles

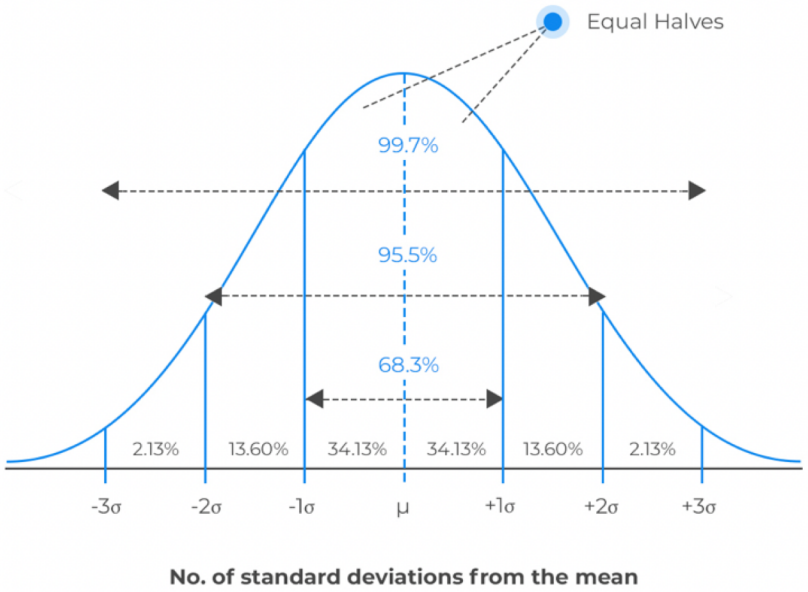


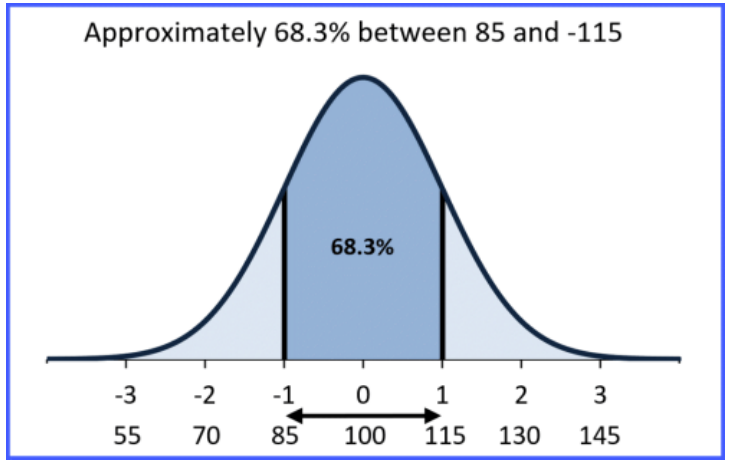


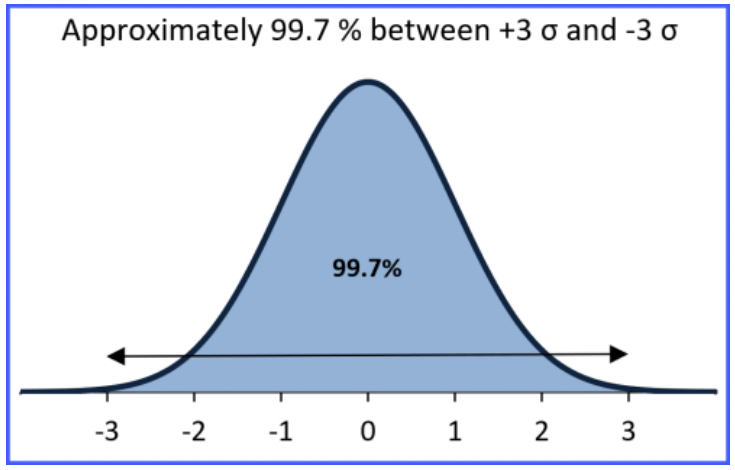
Cphbusiness

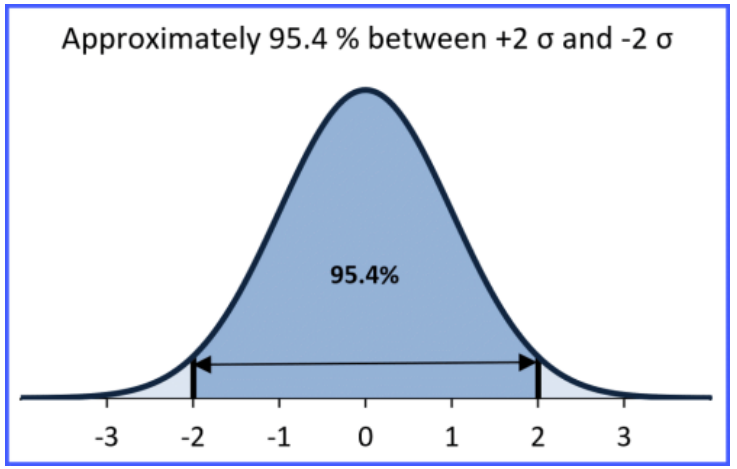
Outliers

Another Empirical Rule

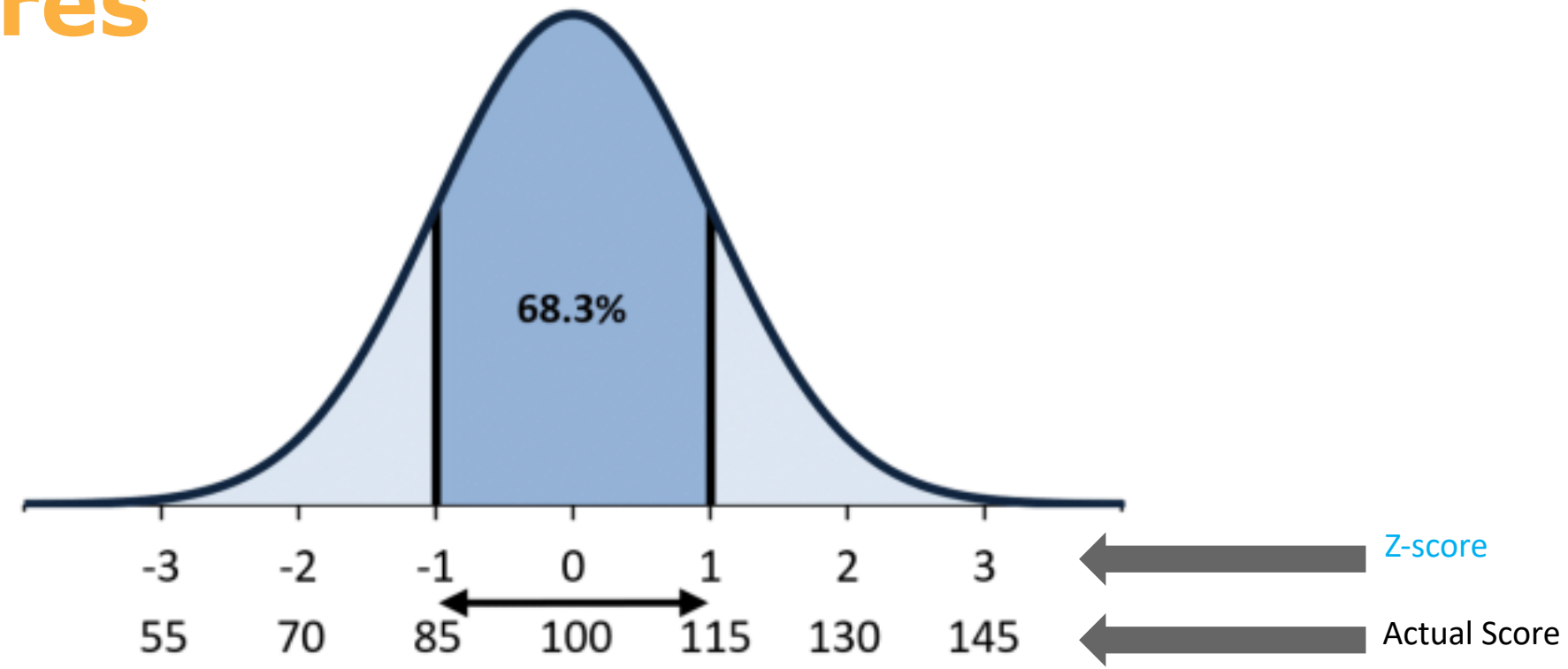






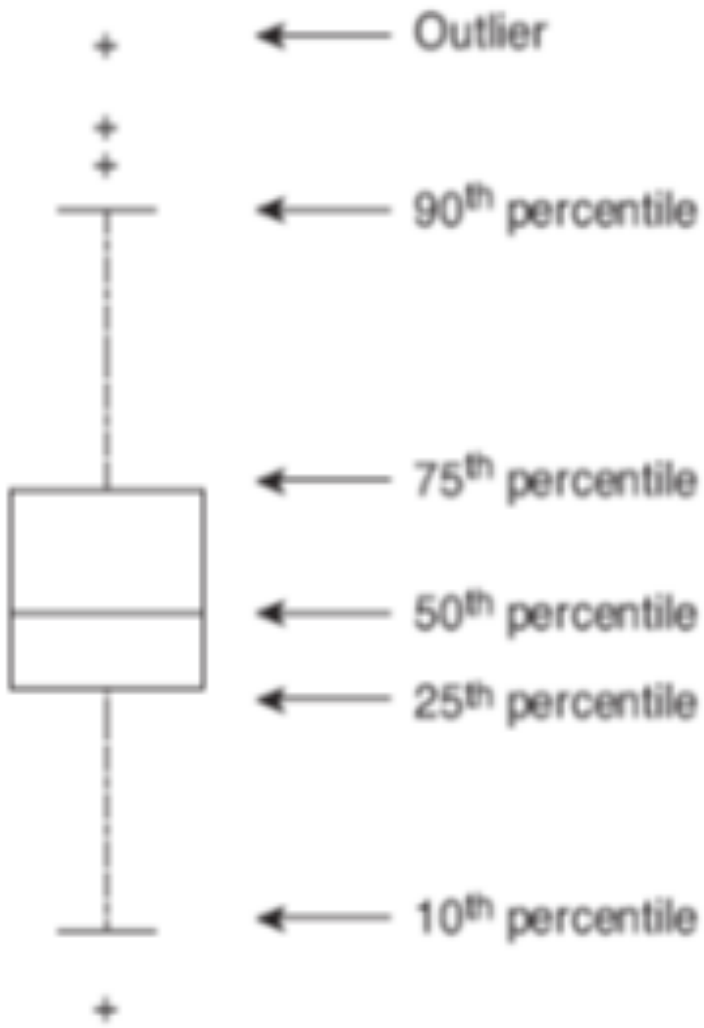


Z-scores



Cphbusiness

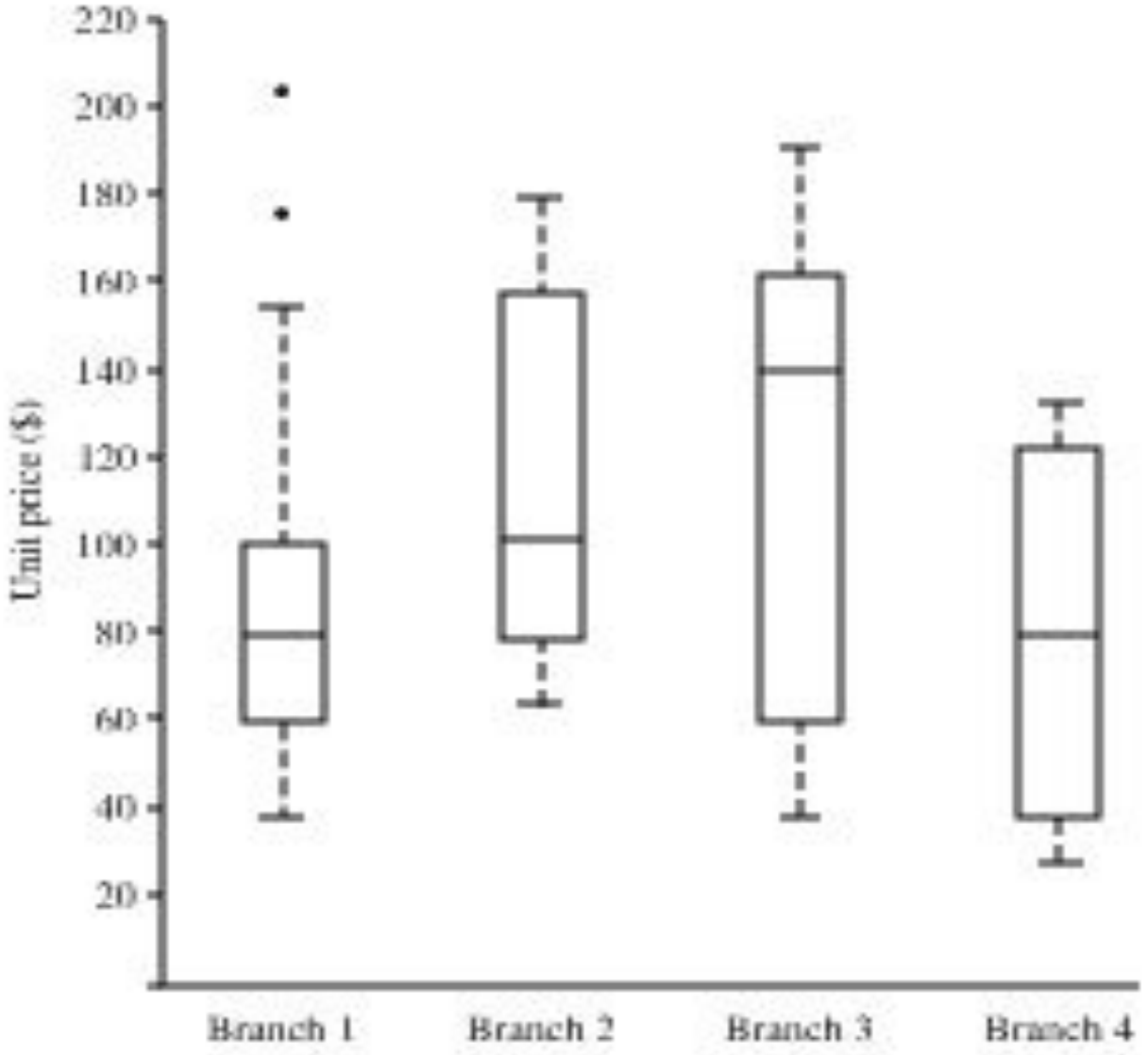
Box Plot



Box Plot

Good to use for

visual comparison of samples



Exercise

$$R = X_{max} - X_{min}$$

$$S = \sqrt{\frac{\sum (X_i - \overline{X})^2}{n-1}}$$

$$V = \frac{S}{\overline{X}}.100$$

- Here is the same set of numbers representing the test results of one class of students:

  24, 18, 12, 15, 19, 17, 17, 19, 21, 19, 18, 16, 22, 19, 20

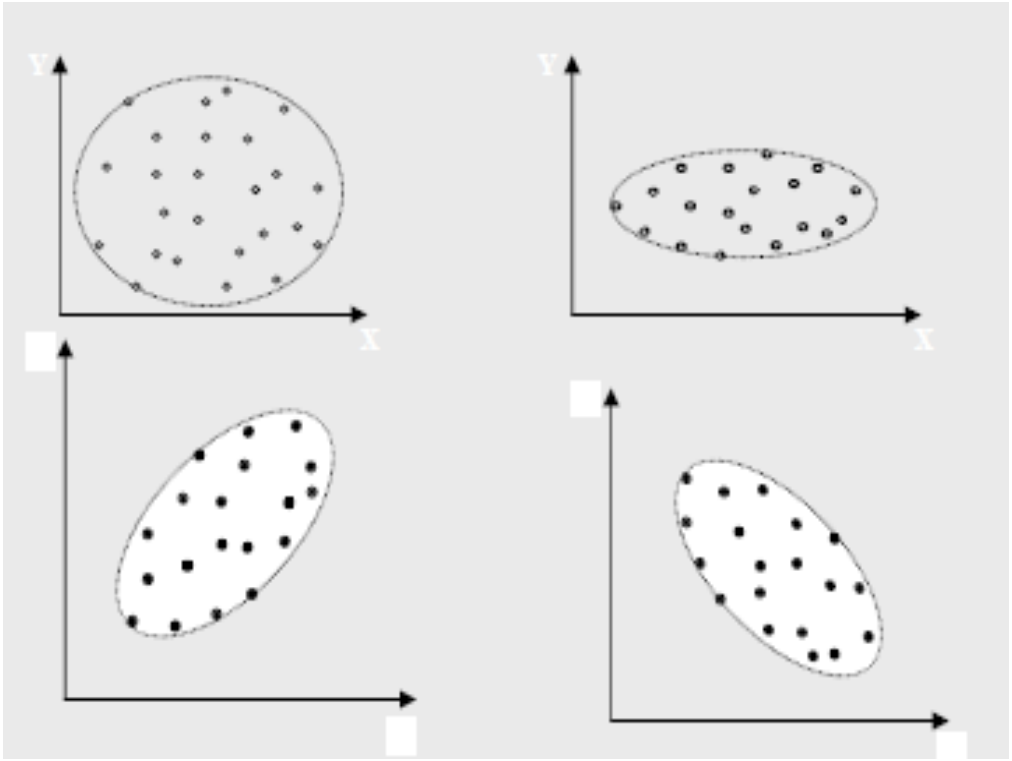
- Calculate the set's
  - range
  - standard deviation
  - variance

- Draw a box-whisker plot representing the set

Correlation

Cphbusiness

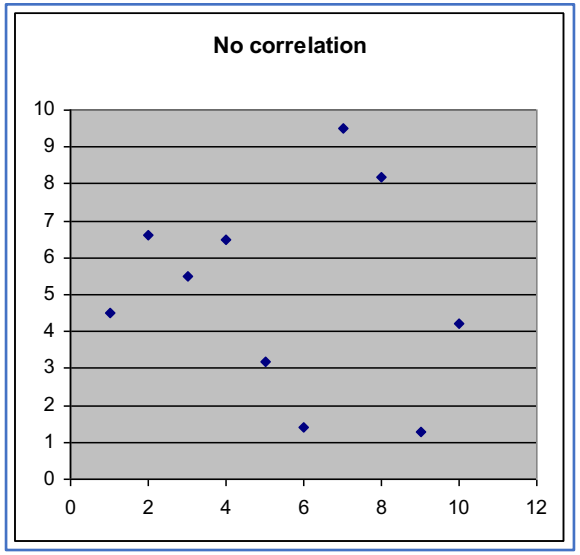
Dependency Between Two Samples

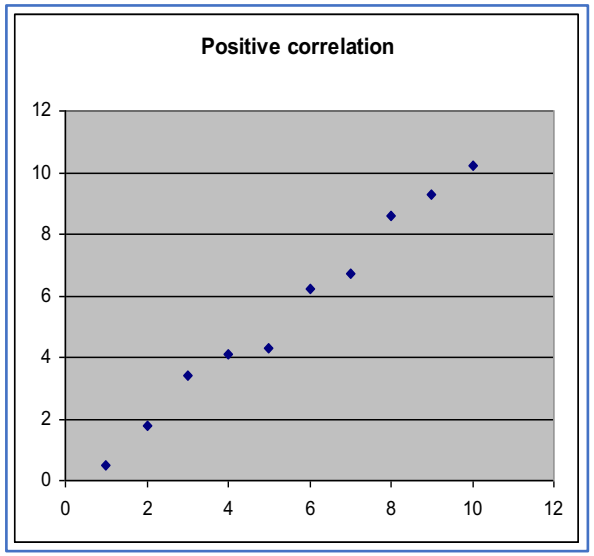


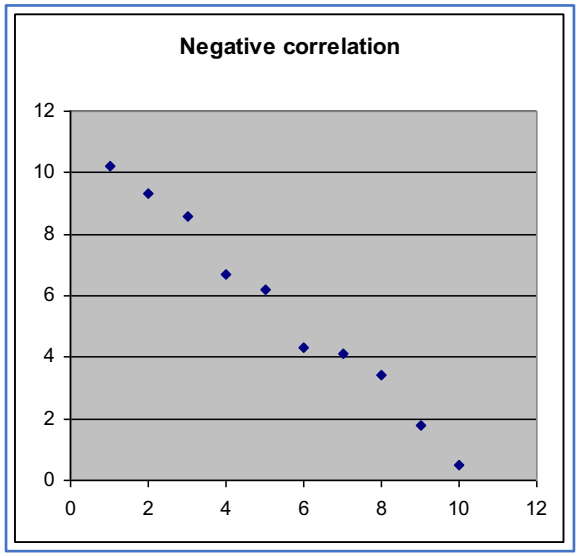
More Examples

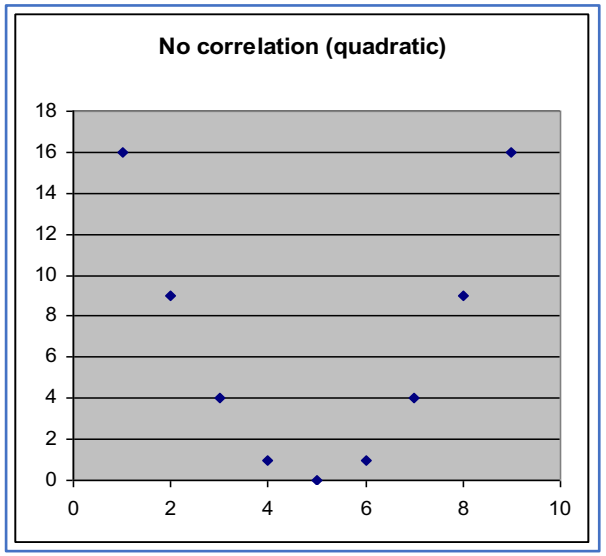
linear

non-linear dependency









Correlation

- Measure of the the strength and direction of the linear relationship between two numerical variables from the same observation data set
- Dimensionless
- Used only if both variables are numerical
- Estimated by a correlation coefficient denoted by $r = [-1, +1]$

Correlation Coefficient

- The value of r belongs to the interval [-1, +1]
  - close to 1 indicates that the variables are positively linearly related and the scatter plot of the two samples falls almost along a straight line with positive slope
  - close to -1 indicates that the variables are negatively linearly related and the scatter plot almost falls along a straight line with negative slope
  - correlation coefficient is close to 0 indicates a weak linear relationship

Correlation Coefficient

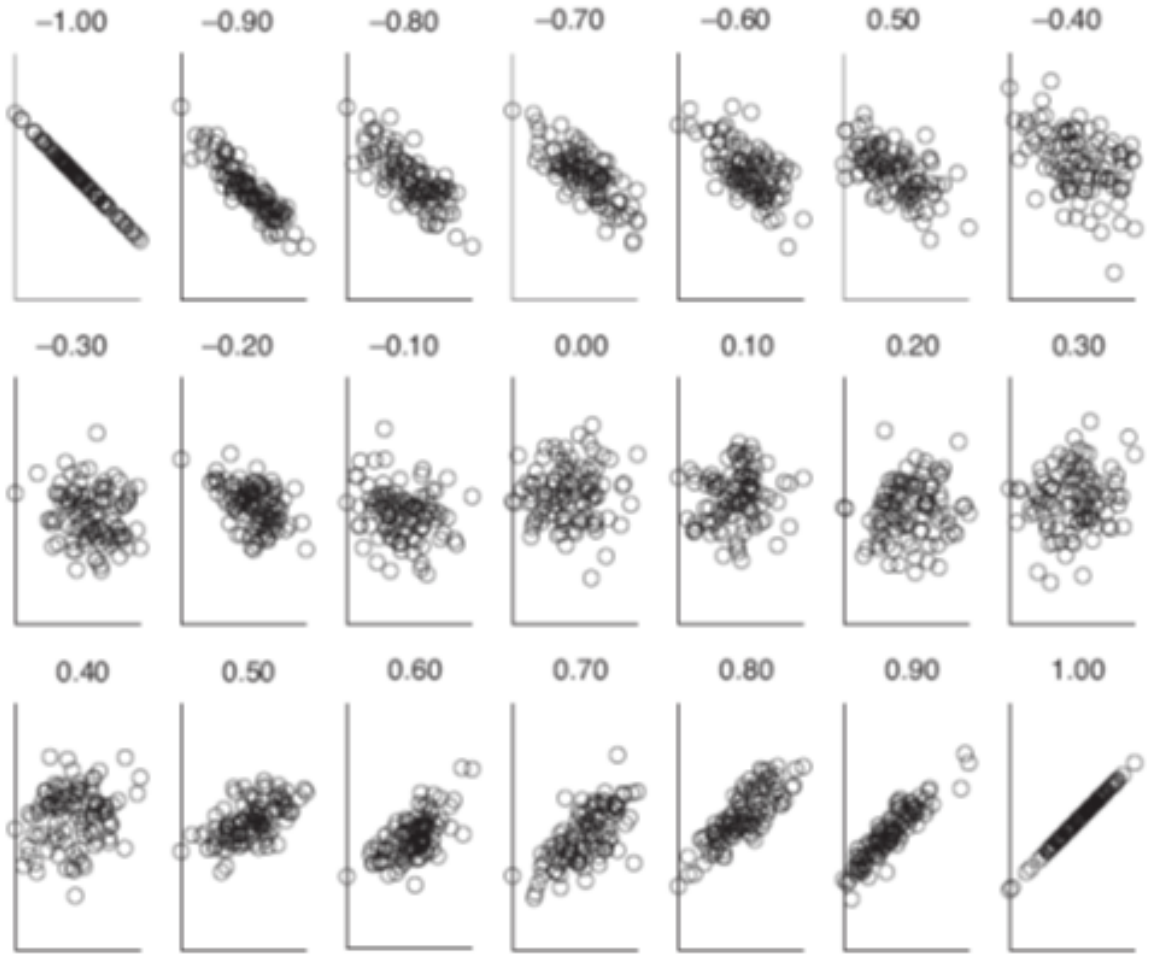
$$r = r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

where:

- n, x_i, y_i are defined as above
- $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ (the sample mean); and analogously for $\bar{y}$

Cphbusiness

Correlation Coefficient

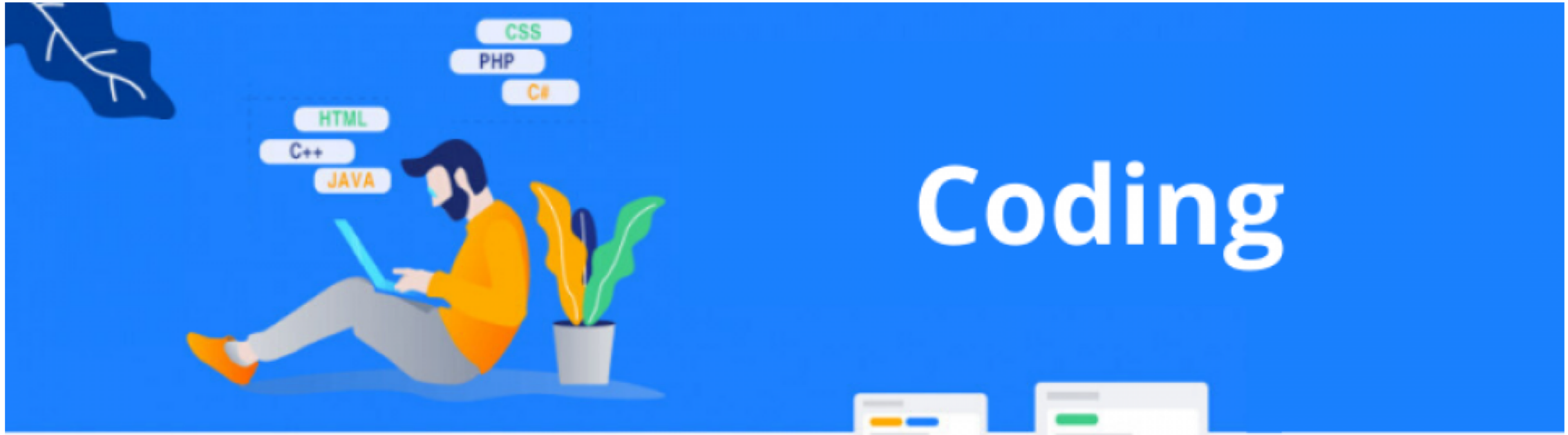


Cphbusiness

Reference

- http://www.r-tutor.com/elementary-statistics
- https://statistics.laerd.com/features-overview.php
- https://statistics.laerd.com/statistical-guides/measures-central-tendency-mean-mode-median.php
- https://statistics.laerd.com/statistical-guides/measures-of-spread-range-quartiles.php
- https://www.statisticshowto.com/
- http://calculator.net

Cphbusiness



Coding

Programming