MS-C1620 Statistical inference

1 Descriptive statistics

Jukka Kohonen

Department of Mathematics and Systems Analysis
School of Science
Aalto University

Academic year 2020–2021
Period III–IV

Course personnel

Lectures on Thursday 8–10 in ZOOM (link on course page).

- Lecturer: Jukka Kohonen
  Office hours: by email, jukka.kohonen@aalto.fi

Weekly exercise sessions.

- Head assistant: Paavo Raittinen
  paavo.raittinen@aalto.fi

See to the course page for materials and announcements.
https://mycourses.aalto.fi/course/view.php?id=29630

Grading

The course grade (0-5) is determined based on the total of exam points (0-24) and exercise points (0-6).

The total points correspond to the following course grades,

- 1 = 15 total points
- 2 = 16 total points
- 3 = 19 total points
- 4 = 22 total points
- 5 = 25 total points

Additionally, grade 1 is also awarded to those who get 12 points from the exam alone.

The exercise points are valid during the year 2021.

Exercises

The course exercises consist of two kinds of problems.

- **Homework problems**: the first problems of each week's exercise sheet are homework problem which must be completed before that week's exercise session. Exception: the first exercise.
- **Class problems**: the remaining problems of each week's exercise sheet are class problems that will be solved together in the exercise sessions.

Exercise points will be awarded from the sessions as follows:

- 1/2 exercise points: both active attendance and homework.
- 1/4 exercise points, only active attendance.

There are a total of 12 exercise sessions, meaning that one can obtain a total of 6 exercise points.

The exercise points are rounded up to the nearest integer.

Exercises

If none of the exercise group times is suitable for you but you would still like to get the exercise points, contact the head assistant. Note that you need an extremely good reason for getting points without attendance!

Lecture topics

1. Descriptive statistics
2. Confidence intervals and hypothesis testing
3.
4.
5.
6.
7.
8.
9.
10.
11.
12.

Contents

Contents

Statistics as a field

Statistics is the science of collecting, organizing, analyzing and interpreting data.

Statistical models are mathematical and based on probability theory.

The practice of statistics can be considered to have begun in ancient Babylon, Egypt and Rome, as population statistics. Data was collected for the government, for example, about birth rates. The word *statistics* comes from the Latin words *statisticum collegium* (council of state).

Descriptive vs. inferential statistics

Statistical methods and procedures can be divided roughly into two categories.

- **Descriptive statistics** aims at providing a concise summary of the data. The summary may be numerical or graphical or of some other form.
  - Examples of descriptive statistics: numerical tables, average values, deviations and visualizations.
- **Inferential statistics** draws conclusions about a population based on a sample. Statistical inference is based on mathematical modeling and probabilities.
  - Examples of inferential statistical methods: confidence intervals, hypothesis testing, linear regression.

Note that there still exists methods that do not really fit into either of the above categories. For example *principal component analysis* can be considered a method of **exploratory data analysis**.

Population and sample

Most of statistical projects revolve around trying to understand a population based on a sample.

- Population is the collection of all the people, items, or events about which one wants to make inferences (students at Aalto University).
- Sample, is a subset of the population (i.e. the people, items, or events) that one collects and analyzes to make inferences on the population (200 randomly chosen students).
- Observation is an element of the sample (Helena, a student at Aalto University).

In a typical project, **descriptive statistics** are first used to understand the sample and select suitable methods of **inferential statistics**, using which we then try to understand the population.

Variables and data

In statistical research, a sample consists of data which is made up of the observed values of selected variables. Sometimes the data points (the values of the selected variables) are also called observations.

Examples of variables:

- Temperature, height, blood pressure (numerical variables, perhaps also *continuous*)
- Clothing size (... S, M, L, ...), (ordinal variable)
- Gender, eye colour (categorical variables)

Variables of different types usually need to be analyzed with different methods.
(Look up also "levels of measurement")

Statistical research projects

Statistical research projects usually consist more or less of the following steps:

1. Define the research topic and the relevant research questions.
2. Define of the population and the variables of interest.
3. Plan of the sample collection such that it is representative of the population.
4. Collect the sample.
5. Organize the sample.
6. Study the sample using descriptive statistics.
7. Model and analyze the sample using inferential statistics.
8. Evaluate the results critically.
9. Communicate the (lack of) findings.

Different statistical studies

Statistical research projects can be conducted in several different ways, depending on the research questions, population, goals and resources.

- **Observational research** simply observes the sample without changing any existing conditions.
  - The lung cancer risk of smokers is compared to the lung cancer risk on non-smokers.
  - The effect of the reputation of an university to the salaries of its graduates is studied.
- **Controlled experiment** examines the effect of one variable to another by controlling existing conditions.
  - The effect of allergy medicine is compared to the effect of placebo by randomizing patients to two groups.
  - The effect of the type of soil to the growth rate of plants is studied by randomly planting plants of the same species to different types of soil.

Different statistical studies

- **Simulations** use mathematical modeling to mimic natural conditions or processes.
  - The spread of the Ebola virus is predicted by applying computer simulations.
  - The safety of a new car model is tested using crash test dummies.

The previous types of studies have various sub-cases. For example, a **survey** is an observational study where a representative sample of the population answers some particular questions.

Contents

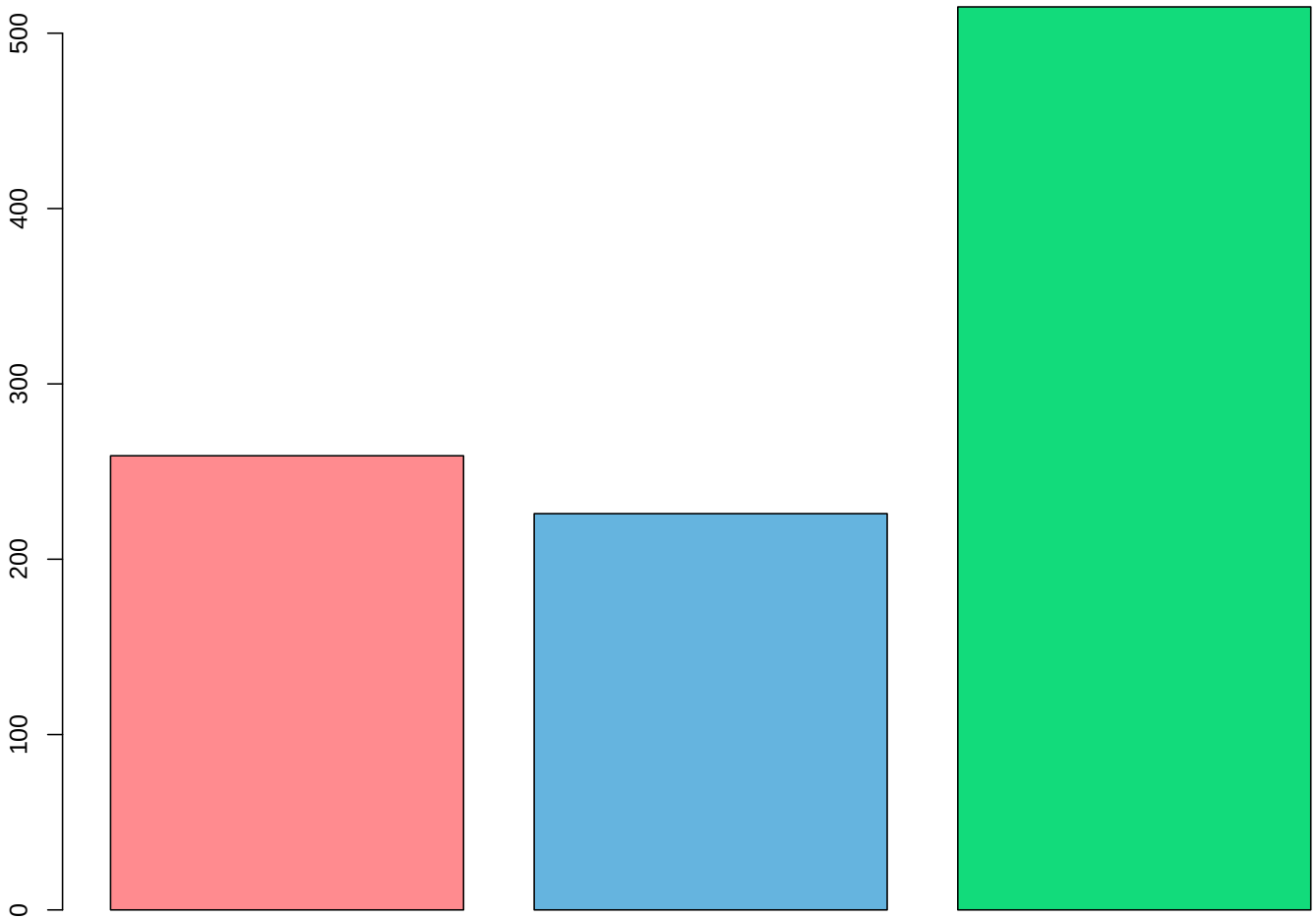
Visualization

The choice of a suitable visualization methods depends on the number of variables (univariate, bivariate, multivariate) and the types of the variables (continuous, discrete). Some examples:

- Discrete variable:
  - Bar plot
  - Pie chart
  - Dot chart
- Continuous variable:
  - Box plot
  - Histogram
- Bivariate (continuous $\times$ continuous):
  - Scatter plot
  - Two dimensional histogram
- Bivariate (continuous $\times$ categorical):
  - Multiple boxplots
  - Colored scatter plots

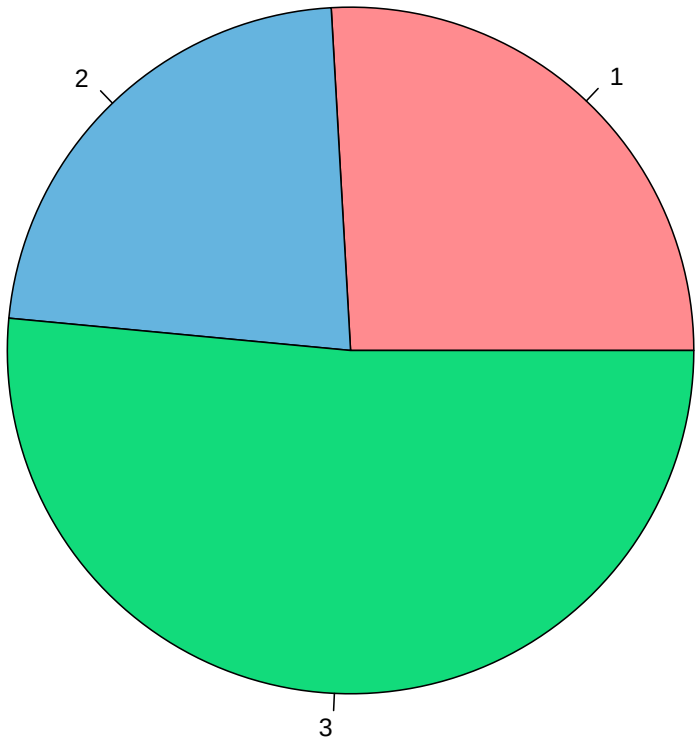
Bar plot

Single categorical variable.



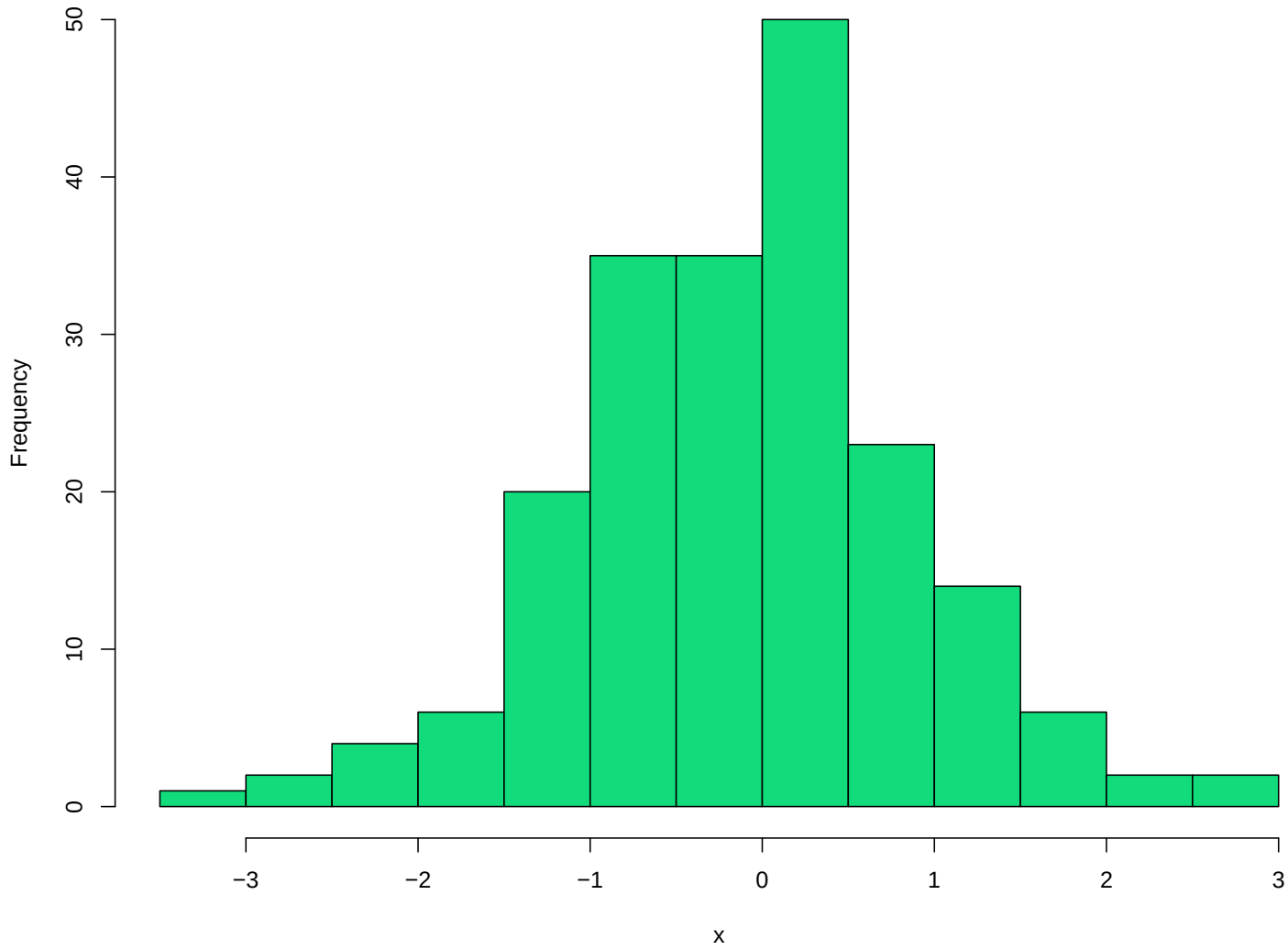
Pie chart

Single categorical variable.



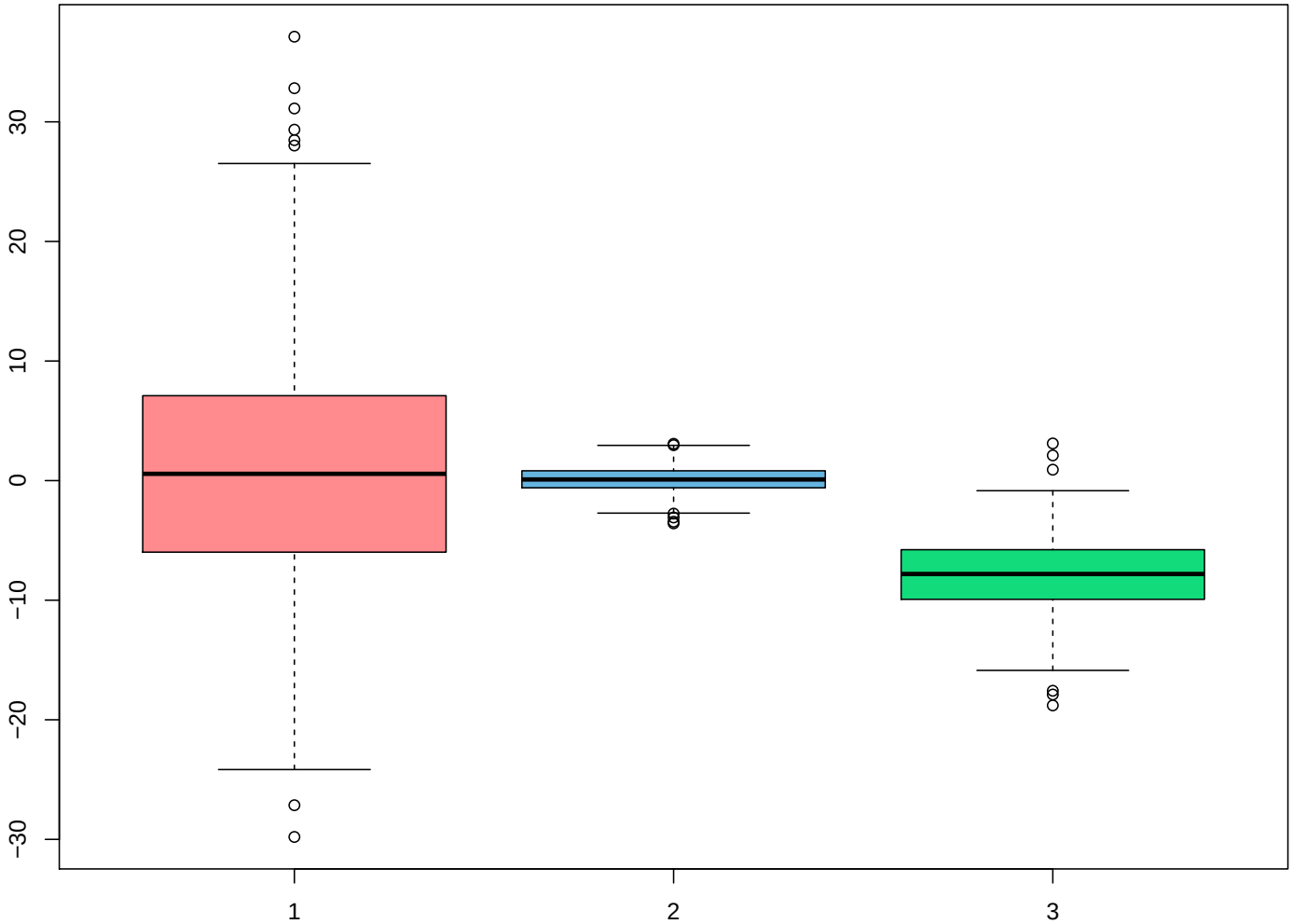
Histogram

Single continuous variable.



Box plot

Continuous $\times$ Categorical



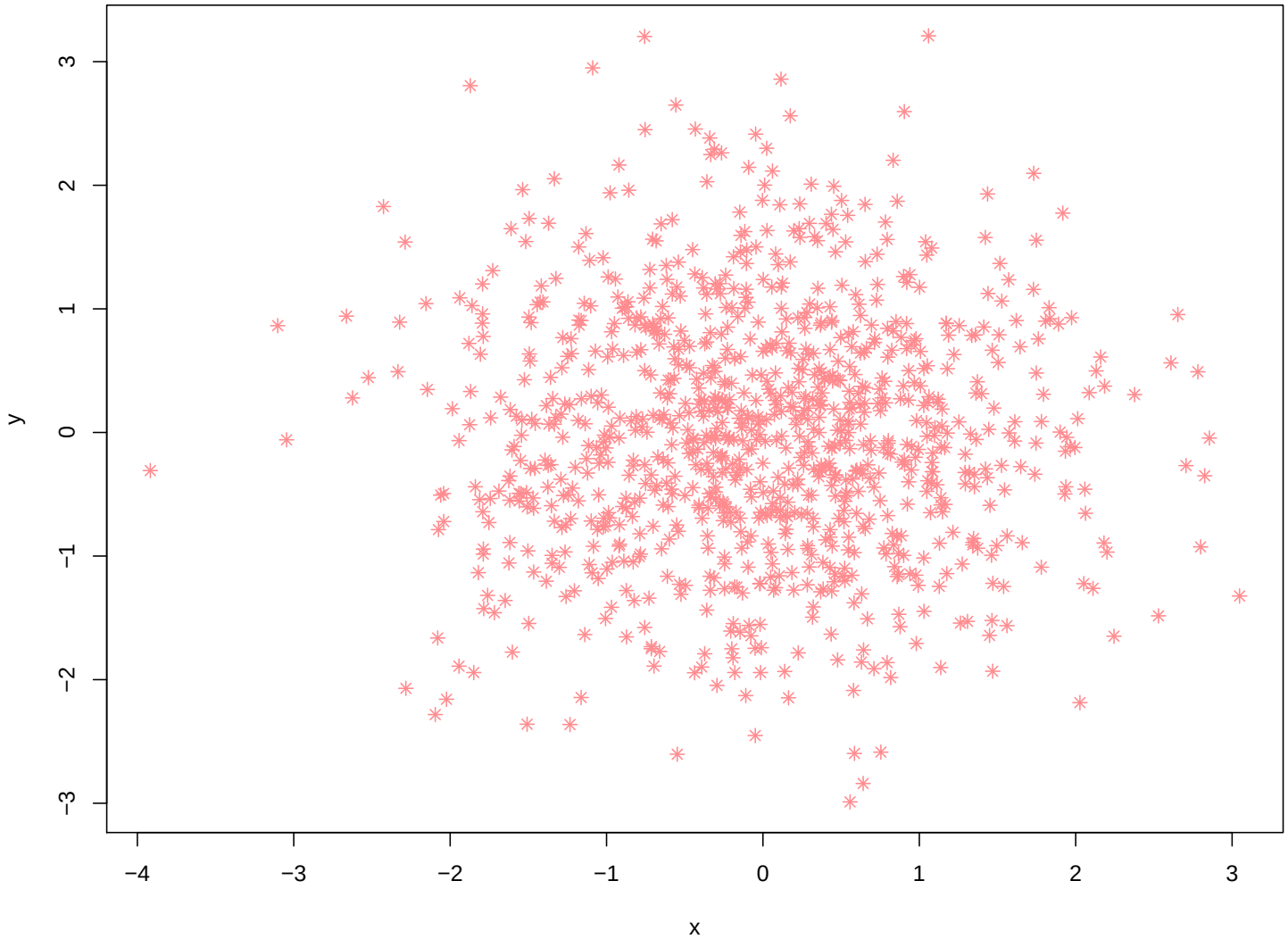
Interpretation of a box plot.

Let Q_1 and Q_3 denote the 25% and 75% sample quantiles and let x_{min} and x_{max} denote the minimum and maximum values of the sample

- The line in the middle of the box is the sample median.
- The lower and upper endpoints of the box are Q_1 and Q_3, i.e., the box contains 50% of the data.
- The upper "whisker" is located at $\min\{x_{max}, Q_3 + 1.5(Q_3 - Q_1)\}$.
- The lower "whisker" is located at $\max\{x_{min}, Q_1 - 1.5(Q_3 - Q_1)\}$.
- Outlying points (not inside the whiskers) are marked using circles.

The box plot allows the simultaneous inspection of location, scatter, symmetry and outliers.

Scatter plot



Good plotting practices

Remember to make your plots easy to understand.

Numerous examples of the kinds of plots you should not make can be found online. Check for example
https://en.wikipedia.org/wiki/Misleading_graph

Contents

Measures of location

Measures of location describe where the *center* of the data lies.

They are used to summarize the typical behavior in the data set.

- The average height of the sample.
- Salary level such that 50% of people have salaries above that.
- The most common number of children.

The following slides list various measures of location, with robust (tolerant to outliers) measures marked in **red**.

Different means

Let $x_1, x_2, \ldots, x_n$ be independent and identically distributed (i.i.d.) observations of a random variable x.

- The sample **mean**,

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i.$$

  The sample mean estimates the expected value $\mu = \mathrm{E}(x)$ of the random variable x.

- The α-**trimmed mean** is the mean of the sample after discarding the proportion α of both smallest and largest observations,
- **Weighted means** give variable weights for different observations,

$$\sum_{i=1}^{n} w_i x_i,$$

  such that $\sum_{i=1}^{n} w_i = 1$.

Median

Let $y_1 < y_2 < \ldots < y_n$ be ordered values of the data.

- The sample **median** is the middle value of the ordered values.
- If the number of observations is even, the sample median is the average of the two middle observations.
- The sample median estimates the population median m_x, the value with the following property

$$\mathrm{P}(x < m_x) \leq \frac{1}{2} \quad \text{and} \quad \mathrm{P}(x \leq m_x) \geq \frac{1}{2}.$$

Quantiles

- The sample **β-quantile**, $0 < \beta < 1$, is the data point y_k, where $k = \lceil \beta n \rceil$ and n is number of observations.
- The sample β-quantile estimates the population β-quantile β_x, defined as

$$\mathrm{P}(x < \beta_x) \leq \beta \quad \text{and} \, \mathrm{P}(x \leq \beta_x) \geq \beta.$$

- 0.25- and 0.75-quantiles are called first and third **quartiles**, and
- the **mid-hinge** is their average

$$\frac{y_{\lceil 0.25n \rceil} + y_{\lceil 0.75n \rceil}}{2}$$

Mode

The sample **mode** is the (possibly non-unique) value that has the highest frequency in the sample.

Mode estimates a value of a qualitative variable or discrete quantitative variable that has the highest probability.

Mode is rarely useful outside of categorical data.

Measures of scatter

Measures of scatter describe how far away from its center the data lies.

They are used to summarize the spread of the data set.

- The variability of height in a population
- The magnitude of measurement errors.

The following slides list various measures of scatter, with robust (tolerant to outliers) measures marked in **red**.

Variance

- The sample **variance**

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

  The sample variance estimates the population variance $\sigma^2 = \mathrm{E}[(x-\mu)^2]$.

- The sample **standard deviation**,

$$s = \sqrt{s^2},$$

  is often preferred over variance as it is measured in the same units as the data.

Median absolute deviation

The **median absolute deviation**, MAD, is the median of the sample $|x_1 - m_x|, |x_2 - m_x|, \ldots, |x_n - m_x|$.

MAD is often multiplied with by the constant 1.4826 to make it a consistent estimator of the standard deviation in a normal model.

Range

- The **sample range** is the interval $[x_{min}, x_{max}]$ and its length is

$$x_{max} - x_{mix}.$$

- The **interquartile range**, IQR, is the distance between the first and third quartile,

$$y_{\lceil 0.75n \rceil} - y_{\lceil 0.25n \rceil}$$

IQR is often multiplied with by the constant 0.7413 to make it a consistent estimator of the standard deviation in a normal model.

Measures of skewness and kurtosis

Thus far, roughly:

- First moment = measures of location
- Second moment = measures of spread

If we continue onward, the next two moments give us measures of *skewness* and *kurtosis*

Skewness describes the deviation of the data from symmetry.

Kurtosis describes the heaviness of the tails of the data.

The following slides list various measures of skewness and kurtosis.

Skewness

The sample **skewness** is

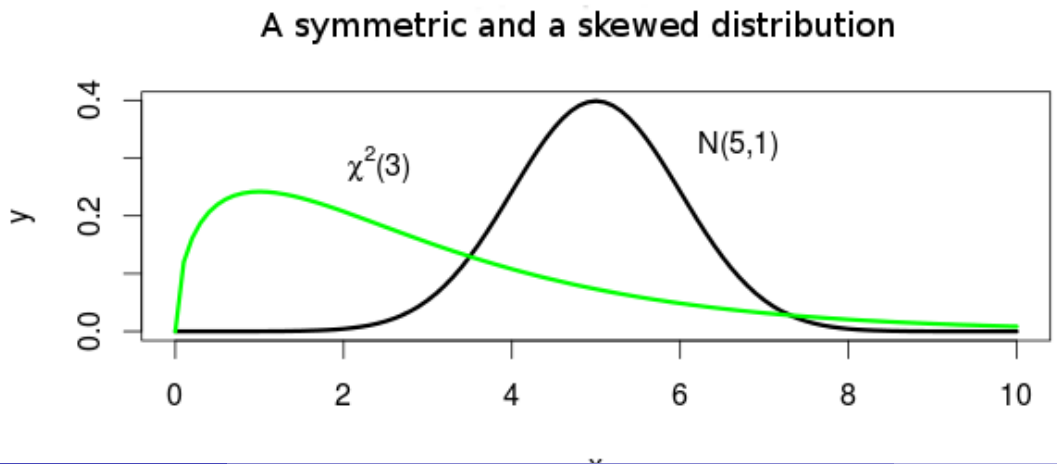
$$\hat{\gamma} = \frac{m_3}{s^3},$$

where

$$m_3 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^3.$$

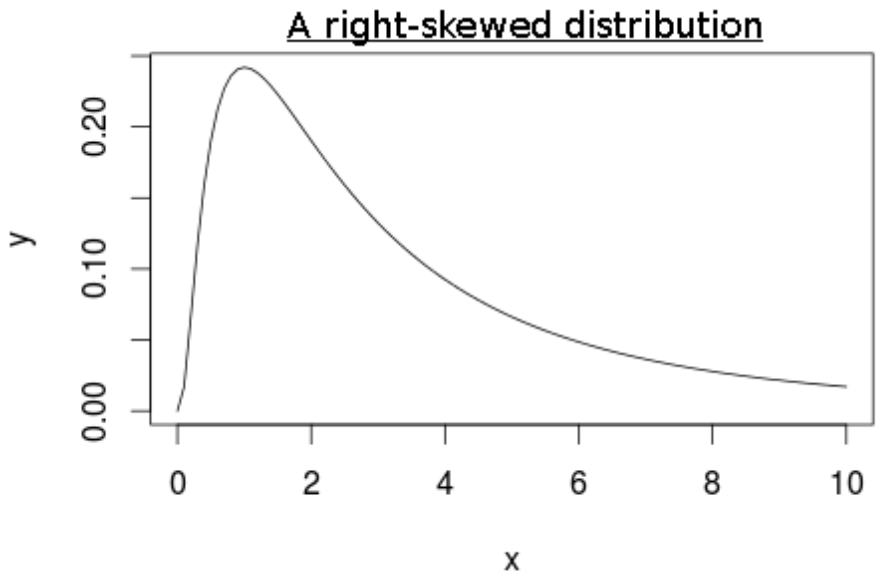
Sample skewness coefficient estimates the population skewness,

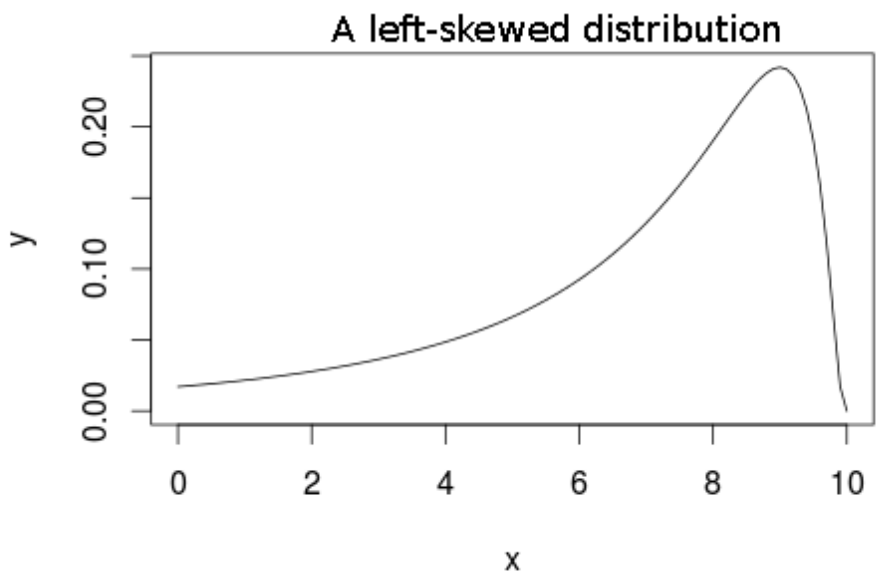
$$\gamma = \mathrm{E}\left[\left(\frac{x-\mu}{\sigma}\right)^3\right].$$



Interpretation of skewness

- If the skewness coefficient $\hat{\gamma} > 0$, then the distribution is *skewed to the right* (positively skewed). I.e. the distribution has a long right tail and the mass of the distribution is concentrated on the left.
- If $\hat{\gamma} < 0$, then the distribution is *skewed to the left* (negatively skewed). I.e. the distribution has a long left tail and the mass of the distribution is concentrated on the right.





Median skewness

The **median skewness**,

$$v_2 = \frac{3(\bar{x} - m_x)}{s}.$$

The underlying reasoning is that for symmetrical distributions the sample mean and the sample median estimate the same population value.

The mean and the median in the median skewness could be replaced with any two measures of location to obtain different measures of skewness.

Kurtosis

The sample **kurtosis coefficient** is

$$\hat{\kappa} = \frac{m_4}{s^4} - 3,$$

where

$$m_4 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^4.$$

The sample kurtosis coefficient estimates the population kurtosis

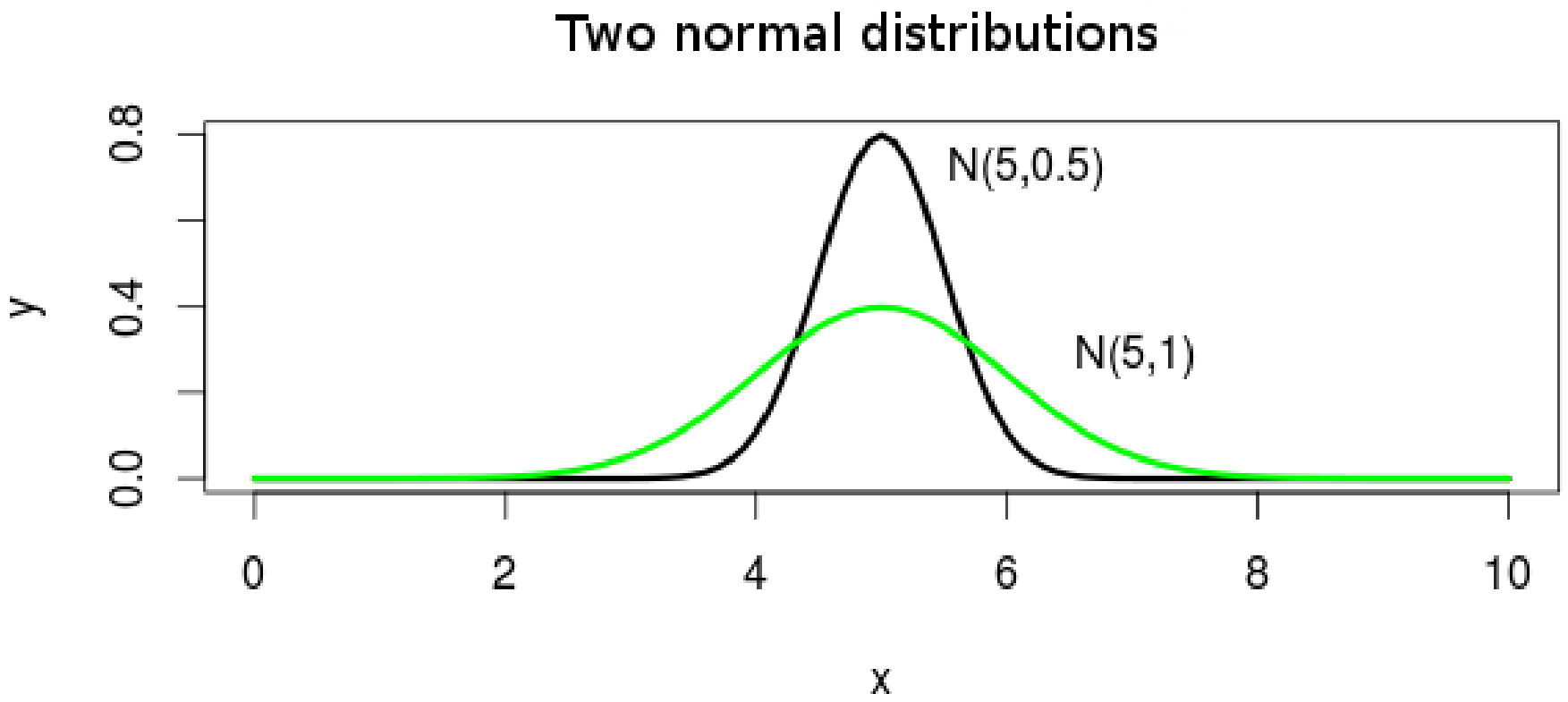
$$\kappa = \mathrm{E}\left[\left(\frac{x - \mu}{\sigma}\right)^4\right] - 3.$$

Interpretation of kurtosis

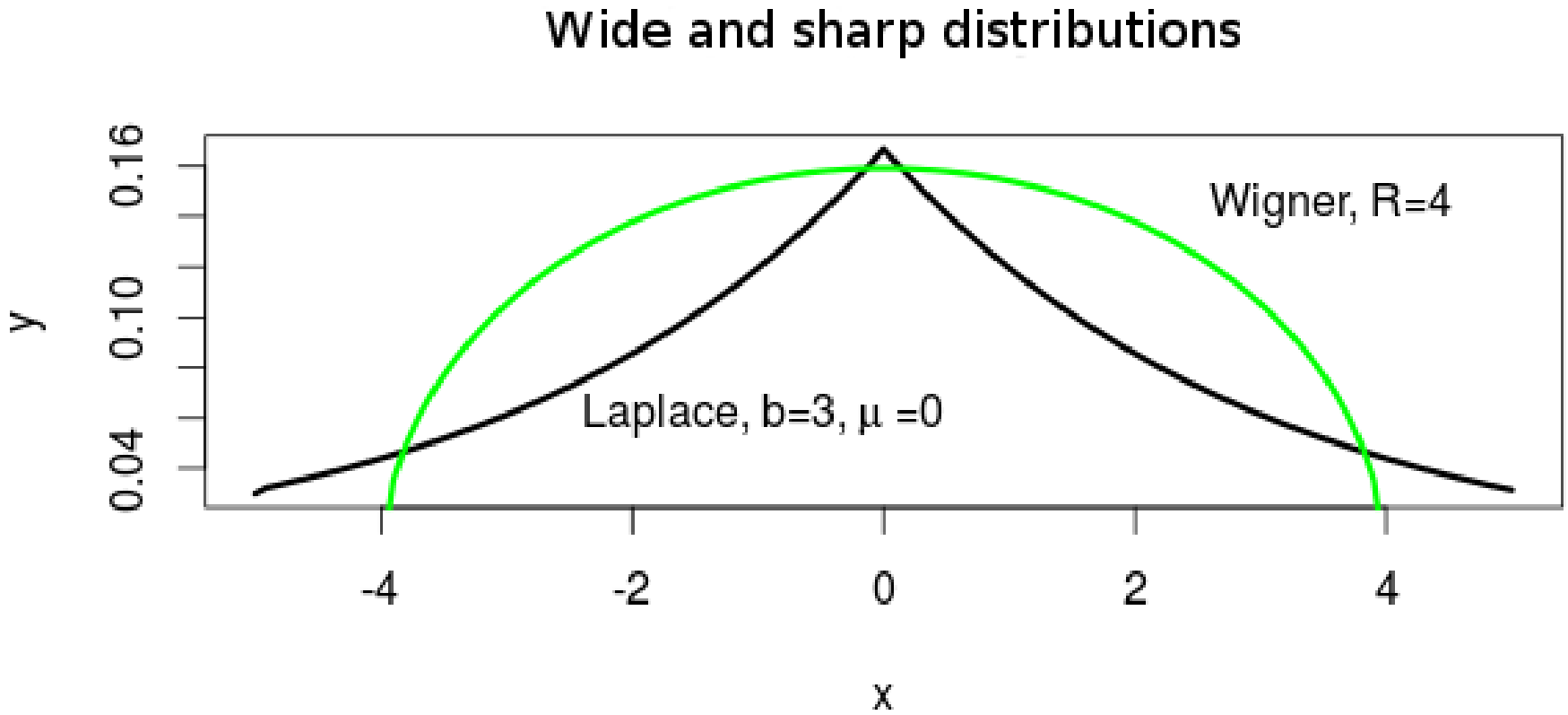
- A random variable with normal distribution has kurtosis value 0.
- If the kurtosis value is $\kappa > 0$, then the distribution has heavier tails than the normal distribution.
- If $\kappa < 0$, then the distribution ahs lighter tails than the normal distribution.

Zero kurtosis

Two normal distributions with different parameters. Kurtosis is the same for both.



Small and large kurtosis



Descriptive statistics for multivariate data

Numerous different descriptive statistics exist also for multivariate data.

The measures are commonly only defined for bivariate data and then computed for all possible pairs of variables.

The most common bivariate descriptive statistic is the *correlation*.

Correlation

Let $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$ be i.i.d. observations of a bivariate random variable (x, y).

- The **sample covariance**

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

  estimates the population covariance $\sigma_{xy} = \mathrm{E}[(x - \mathrm{E}[x])(y - \mathrm{E}[y])]$.

- The **sample correlation**

$$\hat{\rho}(x, y) = \frac{s_{xy}}{s_x s_y} = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n} (x_i - \bar{x})^2 \sum_{i=1}^{n} (y_i - \bar{y})^2}}$$

  estimates the population correlation $\rho(x, y) = \sigma_{xy}/(\sigma_x \sigma_y)$.

Correlation measures the linear dependence between two random variables. The coefficient is always in the interval $[-1, 1]$