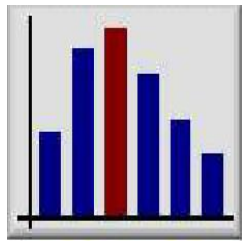
Chapter 2: Statistics with R

Johannes Hain

Chair of Mathematics VIII – Statistics



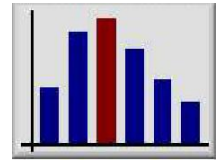
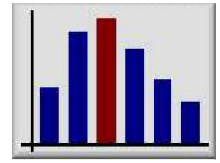


Table of Contents

1. **Foundations of Statistics:**
   Level of Measurements, Random Variables, Density Function, Moments of Random Variables, Moment Estimation
2. **Distribution Fitting:**
   Background, Graphical Tools, Testing Theory, Testing for Normality
3. **Metric Data:**
   One Metric Variable, Two Metric Samples – Correlation Hypothesis, Two Metric Samples – Difference Hypothesis
4. **Categorial Data:**
   Chi Square Test for Independence, Fisher's Exact Test
5. **Categorial and Metric Data:**
   Two Samples, More than Two Samples

Section 6: Foundations of Statistics



Level of Measurements

In statistics we distinct between the following two **data types**:

- **Categorial data:** Roughest level of measurement, classifies data only in categories without internal order.
  $\rightarrow$ examples: hair colors, car labels, sex
- **Metric data:** Measures that can be interpreted by numbers. We make the following sub-classification:
  - **Ordinal scaled data:** Data with an internal order so that a ranking is possible.
    $\rightarrow$ example: grades
  - **Interval scaled data:** Data with an uninterrupted range of values. Distance between observations is meaningful and interpretable.
    $\rightarrow$ examples: body height, body temperature, BMI

Random Variables

We consider now the case that a metric variable is of interest when running an experiment.

Examples:

- The natrium concentration of different soil samples is measured.
- The height of randomly chosen men of Germany is measured.

Such a quantity is called **random variable** X in theory.

We repeat the random experiment (measure the height of a man) multiple times, e.g. n times and say that we have n **repetitions** of X. A single measure is also called a **realization** of X.

Random Variables

Such random phenomenas are mostly not completely arbitrary. The chance is limited by external conditions.

$\Longrightarrow$ **Modeling randomness** with stochastic tools is possible.

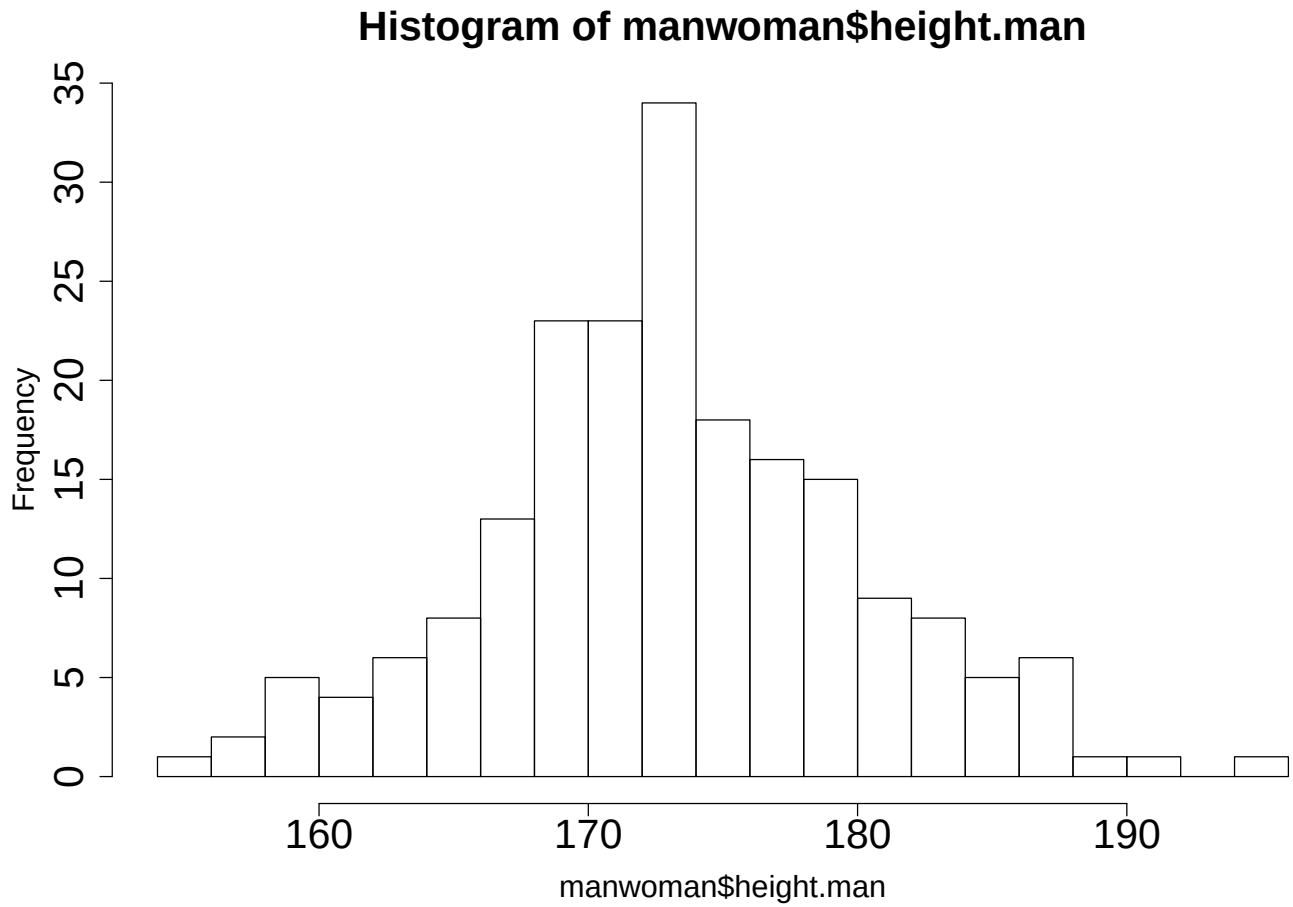
Example

The body height of German men is the realization of the random variable X. If we measure the height n times, we have n realizations of X. What do we know about X?

- We surely have $X > 0m$ and also $X < 3m$.
- Normally we rather have $1.50m < X < 2.10m$.

Random Variables

The distribution of the realizations of a random variable can be visualized with a histogram:



Density Function

Trying to fit a curve at the histogram leads to the **density function** of a distribution:

Definition

The **density function** of a distribution, f_X, is a function with which it is possible to calculate the probability that a repetition of a random variable X realizes itself in a certain interval.

Translation into math:

The function f_X is the density function of a random variable X, if

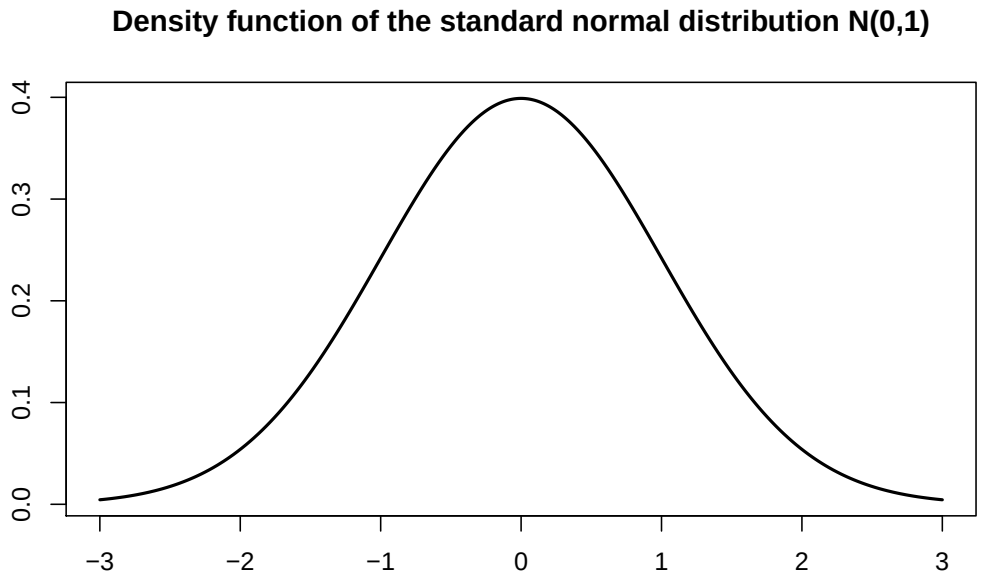
$$P(a < X < b) = \int_a^b f_X(t)\, dt.$$

Density Function

Example

The **density function of the normal distribution** $N(\mu, \sigma^2)$ is defined as:

$$f_{\mu,\sigma}(x) = \frac{1}{\sqrt{2\mu\sigma}} \cdot \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right), \quad x \in \mathbb{R}.$$

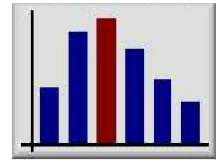


Density Function

There are other probability distributions in real life:

- the **Poisson distribution:** $f_\lambda(x) = e^{-\lambda}\frac{\lambda^x}{x!}$
  $\rightarrow$ Number of suicides, number of accidents in a nuclear power plant,...
- the **Exponential distribution:** $f_\lambda(x) = \lambda e^{-\lambda x}$
  $\rightarrow$ Time between two meteorite impacts, durability of electronic components,...
- the **Lognormal distribution:**
  $f_{\mu,\sigma}(x) = \frac{1}{\sqrt{2\pi}\sigma x}\exp\left(-\frac{(\log(x)-\mu)^2}{2\sigma^2}\right)$
  $\rightarrow$ stock prices, gross/net income of a population,...

However, the normal distribution will be the central point of interest.



Moments of Random Variables

Knowing which distribution the data follows is not enough, we also have to know something about the **expectation** and the **variance** of a distribution. These two parameters characterize the distribution and are also known as the **first** and the **second moments**.

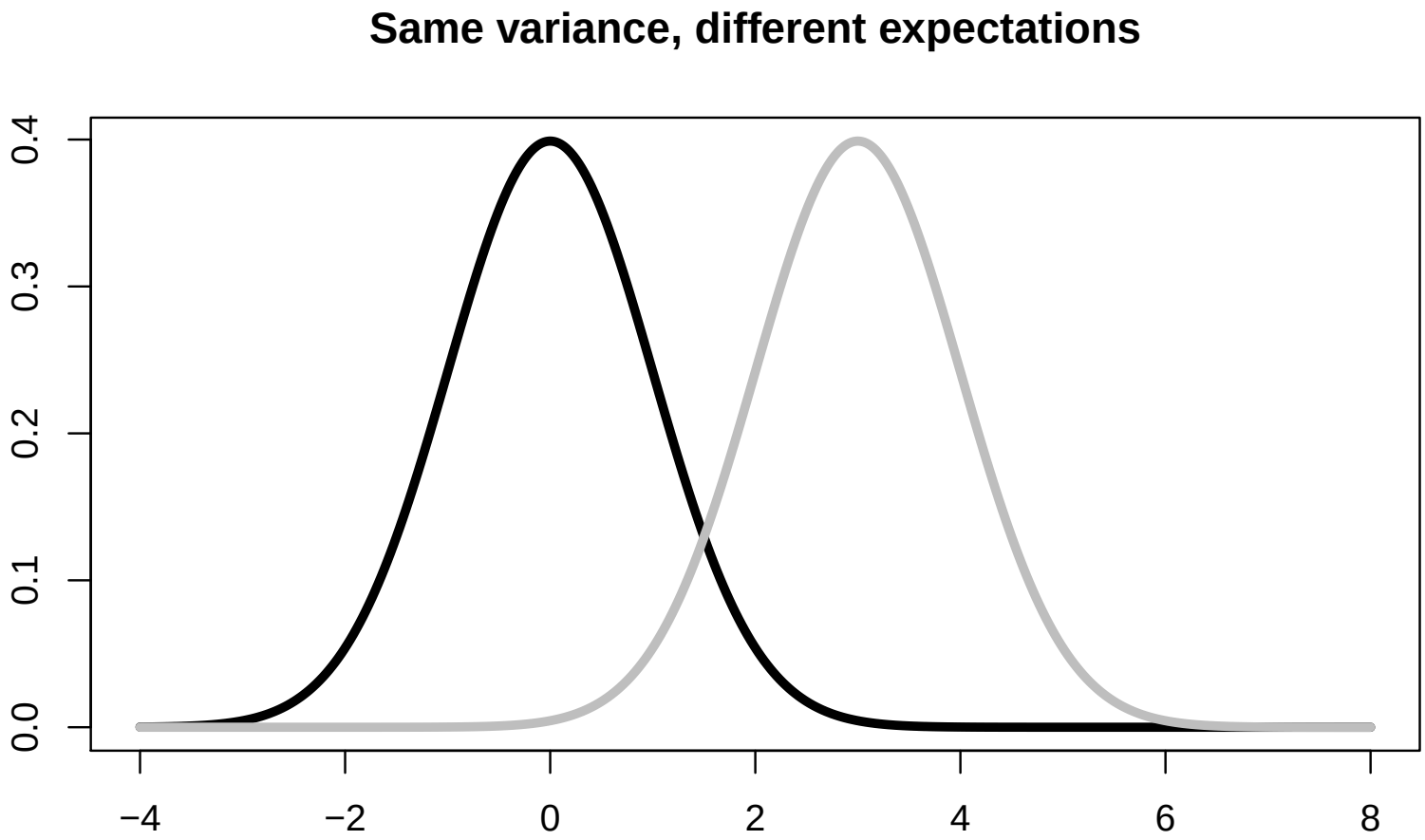
The expectation is defined as follows:

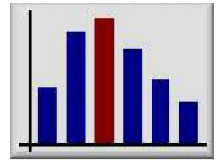
Definition

The **expectation** of a distribution describes the value that is observed on average when repeating the experiment X very often. (Also known as the **Law of Large Numbers**).

The following diagram shows why the expectation is also called **location parameter**.

Moments of Random Variables





Moments of Random Variables

The **variance** σ^2 of a random variable is defined as mean quadratic deviation of the mean, i.e.

$$\sigma^2 := Var(X) := E\left((X - E(X))^2\right).$$

The **standard deviation** σ is more commonly used than the variance and is simply the square root of the variance: $\sigma := \sqrt{Var(X)}$.

The standard deviation (and the variance) is a measure for the dispersion of the random variable X. The smaller σ, the higher is the probability for realizations of the random variable to lay near the expectation. This fact is demonstrated by the following diagram:

Moments of Random Variables

Moment Estimation

The dilemma in statistics is now that the moments of a random variable are of central importance for the understanding of the behavior of a random variable. However, for a random variable these moments are unknown!

To solve this problem we **estimate** the moments based on an available sample $x_1, \ldots, x_n$ of realizations of X and calculate **empirical moments**.

There are two ways to do so:

(i) Point estimation

(ii) Interval estimation

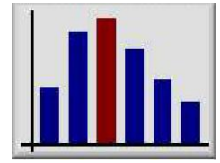
We present both methods in the following.

Moment Estimation

The aim of a **point estimator** is to summarize the information of a sample $x_1, \ldots, x_n$ and estimate a parameter in one single value. To become more mathematic, we have an unknown parameter λ (e.g. the expectation) and a sample of size n. The point estimator $\hat{\lambda}$ of the parameter is a function $T_\lambda(x_1, \ldots, x_n)$.

Example

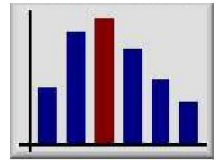
- The point estimator for the expectation μ is the arithmetic mean $\bar{x} := \frac{1}{n}\sum_{i=1}^{n} x_i$.
- The point estimator for the variance σ^2 is $S_n^2 := \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$. Hence the point estimator for the standard deviation σ is simply $S_n := \sqrt{S_n^2}$.



Moment Estimation

In R we can calculate these point estimators very easy. Take the age of the men and the women of `manwoman` as an example.

```
> # estimate the expectation of the men
> mean(manwoman$age.man)
[1] 42.62312
> # for the women we use na.rm due to the NA's
> mean(manwoman$age.woman, na.rm = T)
[1] 40.68235
> # estimate the standard deviation
> sd(manwoman$age.man)
[1] 11.64559
> sd(manwoman$age.woman, na.rm = T)
[1] 11.41442
```



Moment Estimation

The disadvantage of point estimators is that they do not deliver informations about the precision of the estimation. With an **interval estimator** we can calculate an interval around the estimator in which the "true" value of the parameters lies with a certain probability.

Expressed in math notation, we calculate an interval I_α with $P(\lambda \in I_\alpha) = 1 - \alpha$. I_α is then called **confidence interval** to the confidence level $1 - \alpha$. In most cases, $\alpha = 0.05$ so that we can say that with a probability of 95% the true value is in I_α.

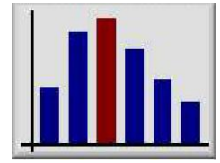
Moment Estimation

Example

Let $x_1, \ldots, x_n$ be a normal distributed sample. The confidence interval to the level $1 - \alpha$ for the expectation μ is then

$$I_\alpha = \left[\bar{x} - t_{1-\frac{\alpha}{2},n-1} \cdot \frac{S_n}{\sqrt{n}};\ \bar{x} + t_{1-\frac{\alpha}{2},n-1} \cdot \frac{S_n}{\sqrt{n}} \right],$$

where $t_{1-\frac{\alpha}{2},n-1}$ is the $1 - \frac{\alpha}{2}$ quantile of the t distribution with $n - 1$ degrees of freedom. It can be interpreted as some kind of a scaling factor.

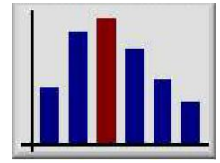


Moment Estimation

The confidence level for the height of the men in `manwoman` is calculated in the following R code in two steps:

```
> n     <- length(manwoman$height.man)
> m.val <- mean(manwoman$height.man)
> s.dev <- sd(manwoman$height.man)
> cl    <- qt(0.975, n - 1) * (s.dev / sqrt(n))
> # create a vector with lower and upper limit
> c(m.val - cl, m.val + cl)
[1] 172.2882 174.2103
```

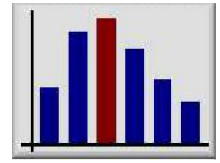
With a probability of 95%, the true expectation of the height of the man lies between 172.2882 and 174.2103.



Moment Estimation

Another disadvantage of point estimators (and of interval estimators) is that they are very sensible concerning outliers in the data. The data set `billionaire` demonstrates this very nice.

```
> # mean before and after
> mean(billionaire$municipal.before, na.rm = T)
[1] 22206.03
> mean(billionaire$municipal.after)
[1] 5022095
> # standard deviation before and after
> sd(billionaire$municipal.before, na.rm = T)
[1] 20678.71
> sd(billionaire$municipal.after)
[1] 70709111
```



Moment Estimation

The same effect can be observed for the confidence intervals for both variables.

```
> n     <- length(billionaire$municipal.before)
> m.val <- mean(billionaire$municipal.before, na.rm = T)
> s.dev <- sd(billionaire$municipal.before, na.rm = T)
> cl    <- qt(0.975, n - 1) * (s.dev / sqrt(n))
> c(m.val - cl, m.val + cl)
[1] 19322.62 25089.44
```

For the variable `municipal.before`, the 95% confidence interval is $(-4\,837\,469, 14\,881\,659)$.

Moment Estimation

→ One single value can have a very strong effect on the moment estimators.

→ We need other location and dispersion estimators that are not effected that strongly by outliers. These estimators are called **robust** estimators.

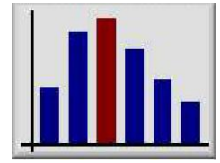
Robust moment estimators

- The **Median** is a measure for the center of the distribution. It separates the higher half of the sample from the lower half. Hence on each side of the median, 50% of the observations can be found.
- The **Interquartile range (IQR)** is a measure for the dispersion of the data. It is defined as the range of the area in which 50% of the data are located.

Moment Estimation

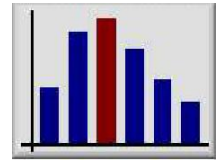
Robust moment estimation in R:

```
> # median before and after
> median(billionaire$municipal.before, na.rm = T)
[1] 15000
> median(billionaire$municipal.after)
[1] 15000
> # IQR before and after
> IQR(billionaire$municipal.before, na.rm = T)
[1] 15000
> IQR(billionaire$municipal.after)
[1] 15000
```



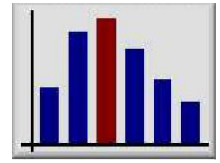
Exercises

1. Data set `manwoman`: Calculate the point estimators for expectation and standard deviation for the variables `age.man`, `height.woman` and `height.man`.
2. Data set `pisa`: Calculate the point estimators for expectation and standard deviation for the three performances parameters `reading`, `science` and `math`.
3. Data set `cinema`: Calculate the point estimators for expectation and standard deviation for the variables `age` and `visits`.

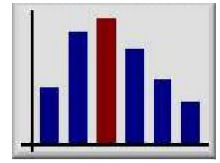


Exercises

4. Write a function `conf.interval()` that computes the confidence interval I_α for the expectation as given in the upper example. The input parameters should be the variable name and the confidence level. After that, determine the 95% confidence intervals of the tree variables given in task 1.
5. Calculate the robust estimators for location and dispersion for the three performance parameters in `pisa`. Compare them with the non-robust estimators in task 2.



Section 7: Distribution Fitting



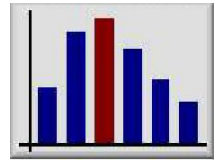
Background

As we just learned, the distribution behavior is of central importance. That is why we have to analyze which distribution we can assess to the observations before we can investigate the actual research hypothesizes. For our purposes we can simplify this question:

Issue of distribution fitting

Can the observed metric data be approximated by a normal distribution or not?

The reason is that many tests (e.g. the t test) demands that the data follows a normal distribution. If this assumption is injured and the test is conducted anyway, the results are worthless!



Background

To give an answer to this question, we have two different tools to our disposal: **graphical tools** and a **test for normality**. However, it is strongly recommended that not only one of these methods is used, but both.

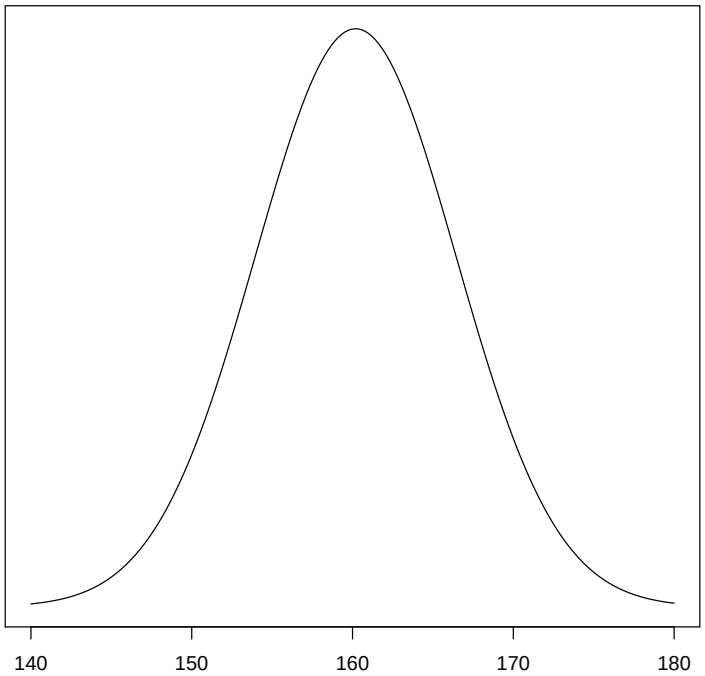
The most important graphical tools are

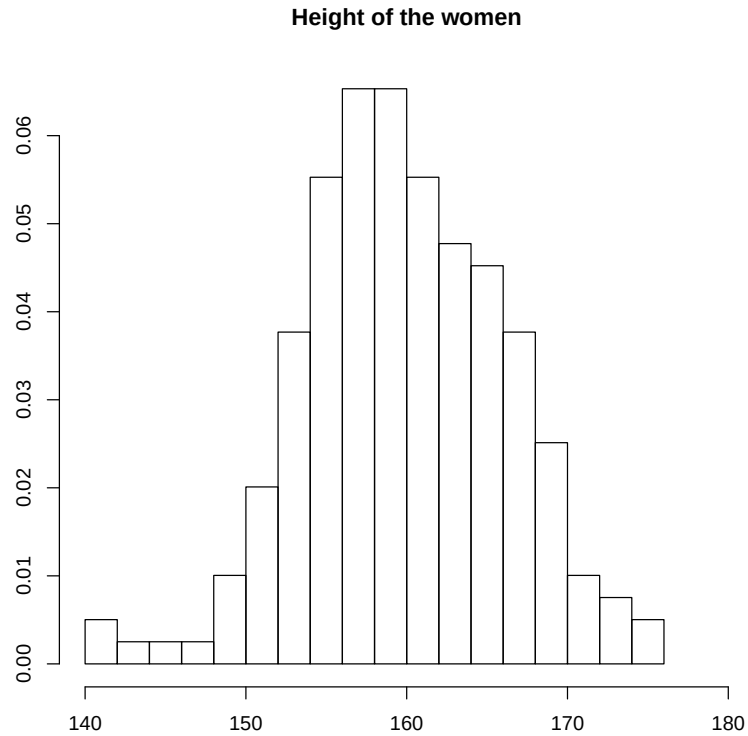
- **histograms**
- **boxplots**
- **normal probability plots**

There are other tools like for example the **stem-and-leaf plot** which are less important and not considered here.

Graphical Tools

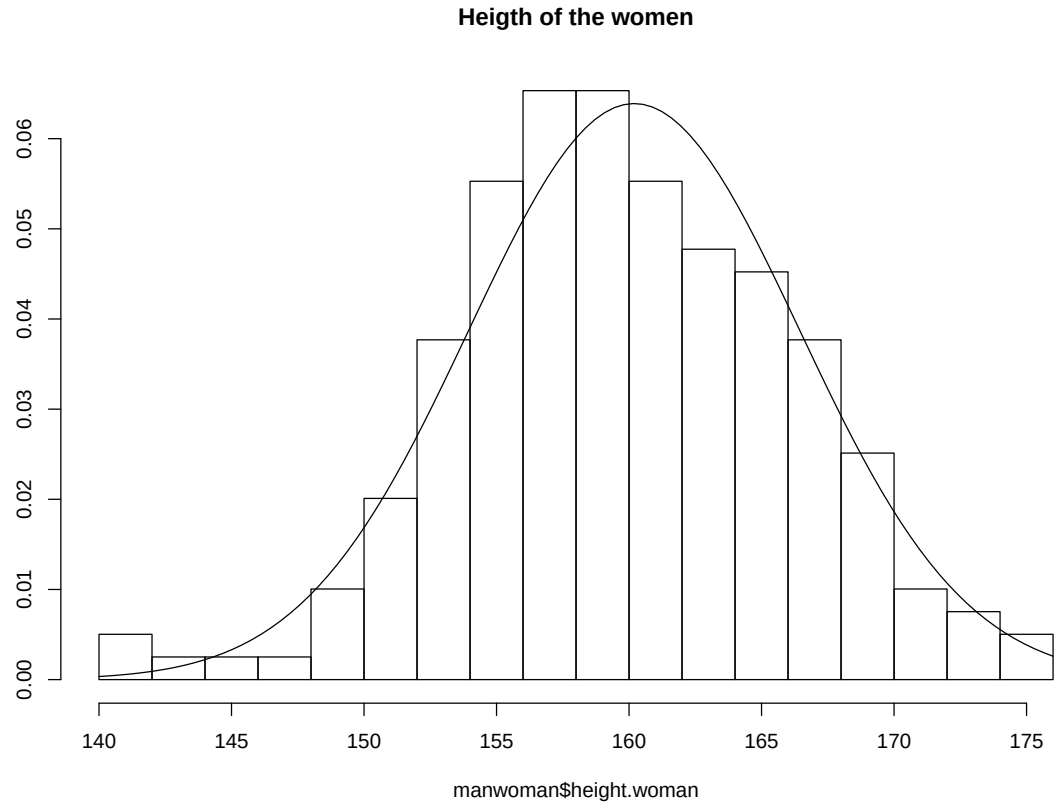
As we have already seen in chapter 1, a histogram plots the empirical distribution of metric data. We can now compare the shape of the histogram with the theoretical density of the normal distribution.

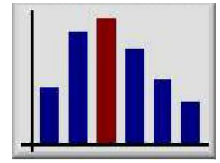




Graphical Tools

If we plot the two elements in one diagram we can make a decision about the goodness of fit. In this example it seems that the fit is quite good which is an evidence for normality.



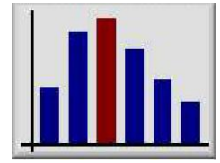


Graphical Tools

```
> # plot the histogram
> hist(manwoman$height.woman, freq = F, breaks = 20,
+ main = "Height of the women", ylab = "")
> # add the normal curve
> curve(dnorm(x, mean = mean(manwoman$height.woman),
+ sd = sd(manwoman$height.woman)), add = T)
```

→ We use the argument `freq = F` in `hist()` so that it can be interpreted as an empirical density.

→ The function `dnorm()` plots the values of the normal distribution. With the argument add, the curve is added to the histogram.



Graphical Tools

Another graphical tool to describe the distribution of a variable is the **boxplot** (or **box-and-whisker plot**).

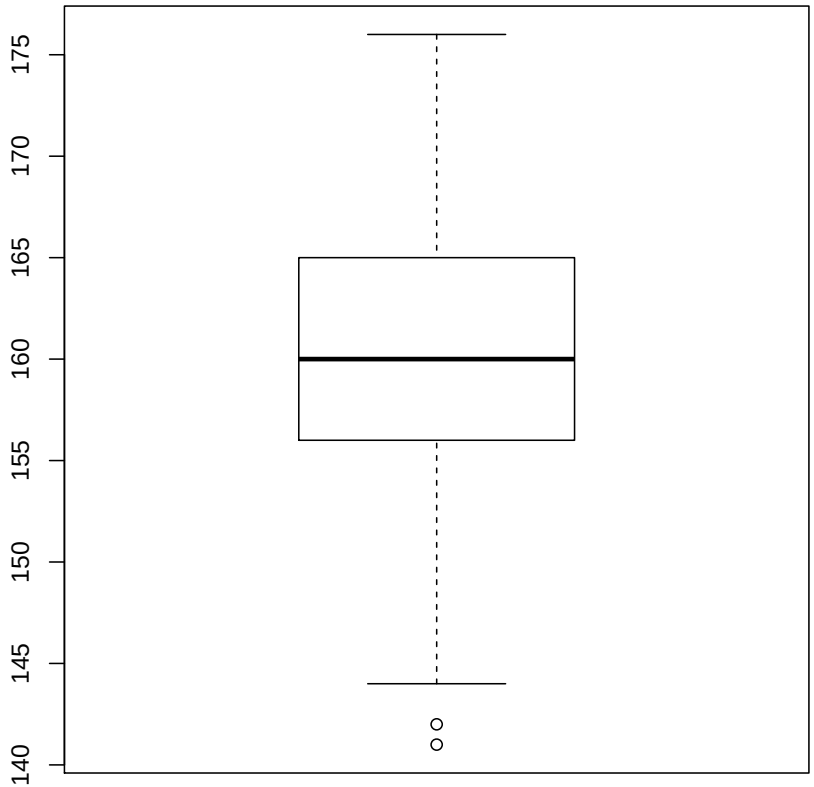
Construction of a boxplot

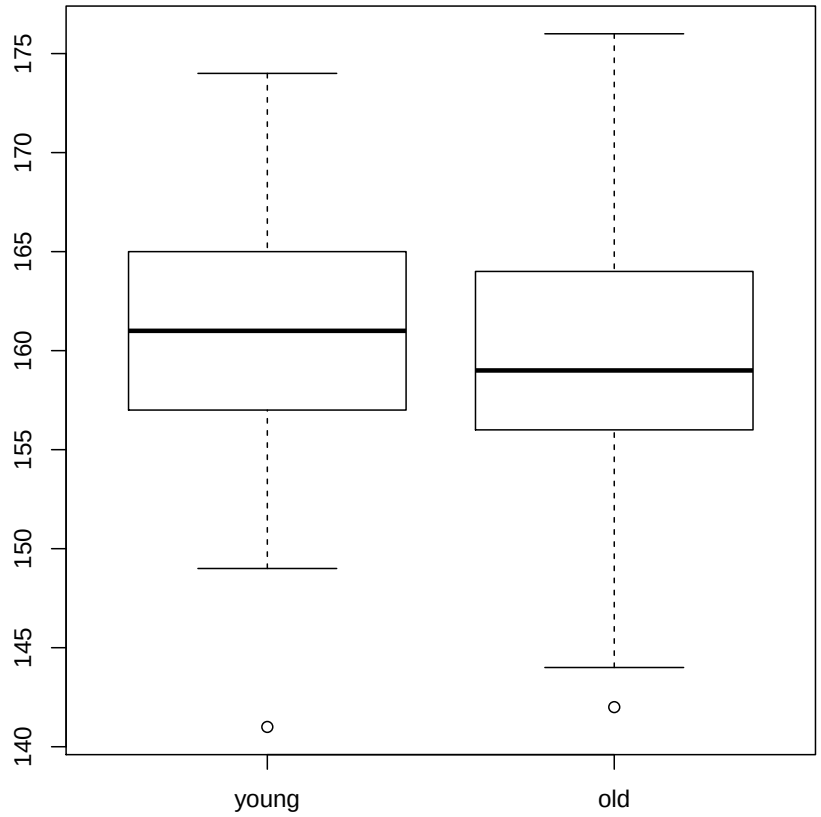
The boxplot is based on the interquartile range (IQR). The range of the "box" is the IQR, the line through the box is the median. The whiskers describe the location of the data in the exterior regions. Their end points are given by $\pm 1.5 \cdot IQR$ beginning at both ends of the box. All values outside the whiskers are labeled as outliers.

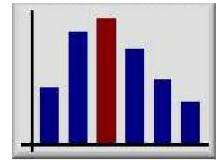
$\rightarrow$ With a boxplot we are able to see the location and the dispersion of the data at a glance.

$\rightarrow$ If the shape of the data is asymmetric, the length of the whiskers differ or the median is not located exactly in the middle of the box.

Graphical Tools







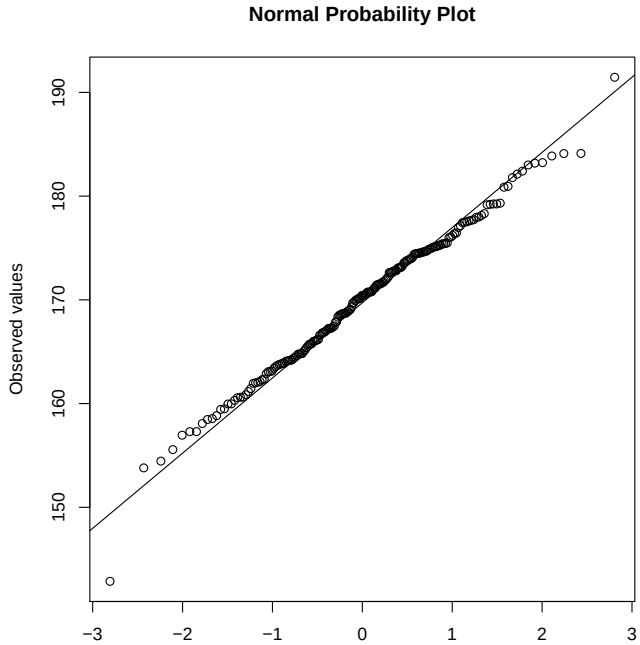
Graphical Tools

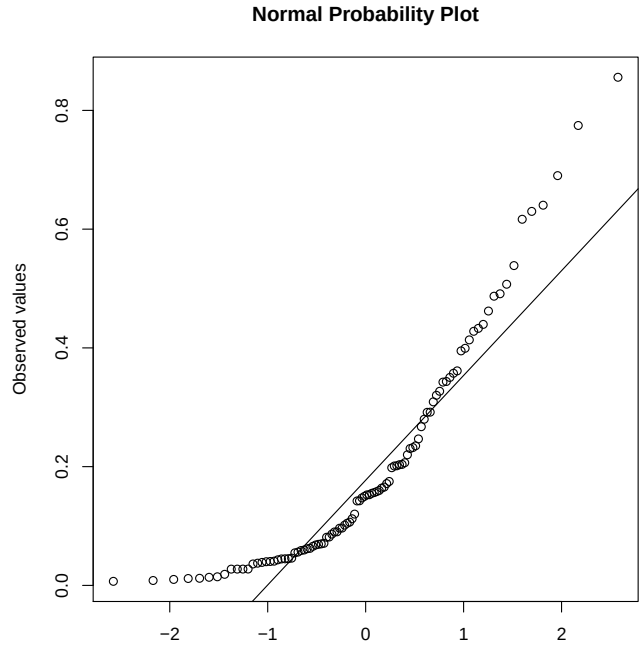
```
> # boxplot without grouping
> boxplot(manwoman$height.woman)
> # boxplot with grouping
> plot(manwoman$age.woman.recoded,
+ manwoman$height.woman)
```

The generic function `plot()` identifies the variable `age.woman.recoded` as a **factor** and `height.woman` as of the type **numeric**. In this case, a boxplot is plotted automatically.

Graphical Tools

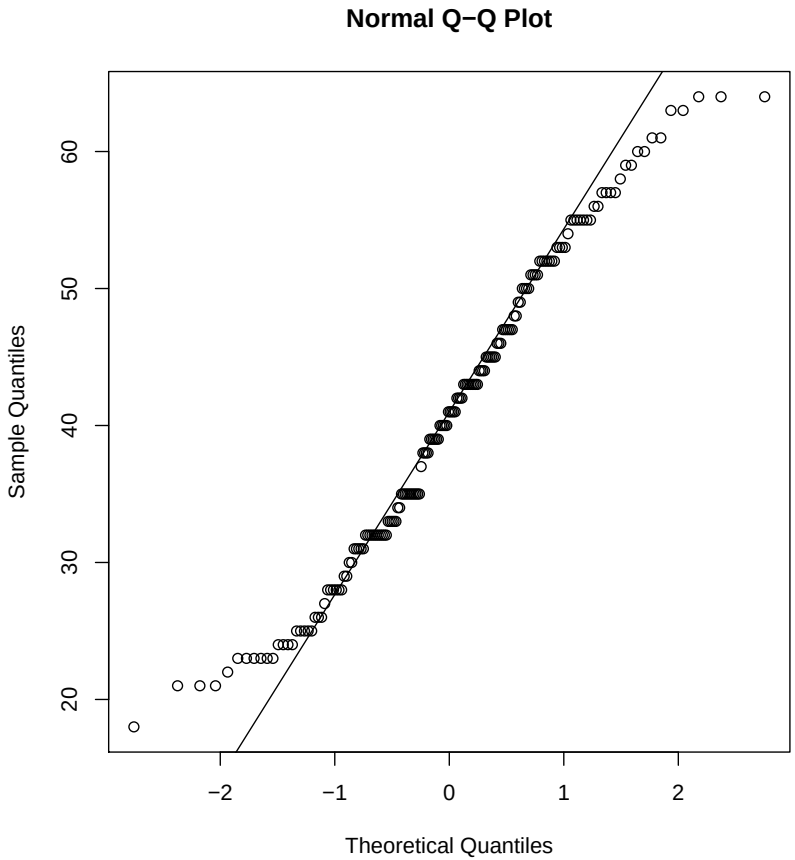
In a **normal probability plot** (or Q-Q plot) the empirical values are plotted against theoretical values in a diagram. The theoretical values are determined under the assumption of normality. If the hypothesis that the data can be approximated by a normal distribution is true, the points in the diagram should lie more or less on a line.

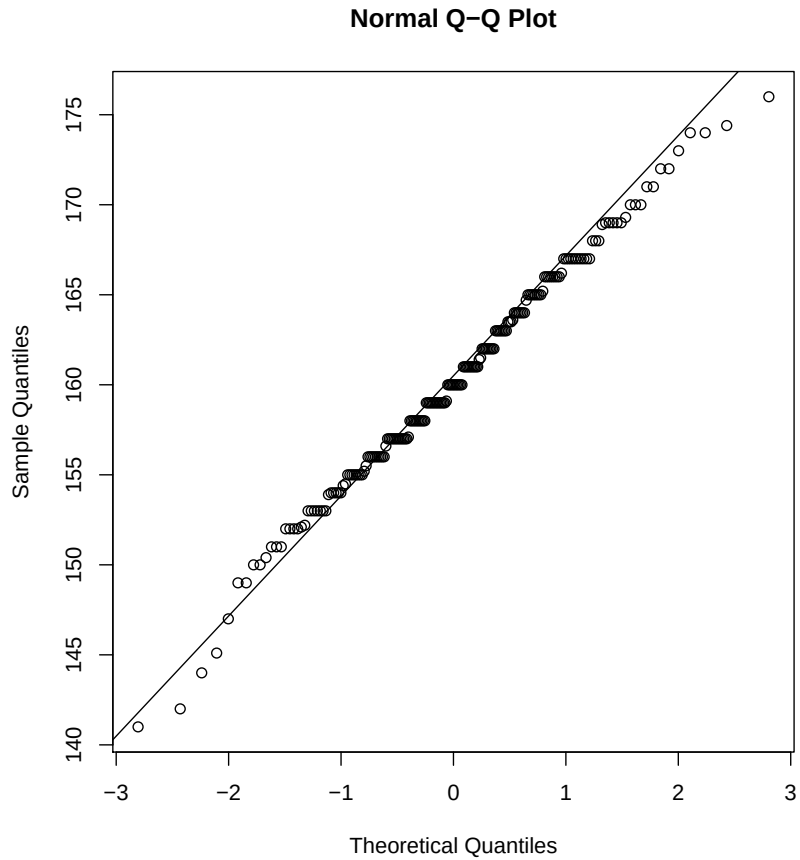




Graphical Tools

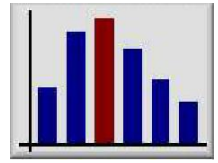
Example with the variables `age.woman` and `heigth.woman`.





Graphical Tools

```
> # Q-Q plot age.woman
> qqnorm(manwoman$age.woman)
> qqline(manwoman$age.woman)
> # Q-Q plot height.woman
> qqnorm(manwoman$height.woman)
> qqline(manwoman$height.woman)
```



Testing Theory

Besides the graphical tools to collect evidences for or against normality, we can also conduct a statistical test to check whether the data is normal or not. Before doing that, we have to make a short excursus into **inferential statistics**.

Purpose of inferential statistics

Based on a collected sample we try to draw conclusions about the properties of the basic population. This population is mostly much bigger than the actual sample size.

The tools of inferential statistics are also called **statistical hypothesis tests**. The topic in testing is to investigate the correctness of a certain hypothesis, also called **null hypothesis** (or H_0).

Testing Theory

Example

H_0 : The random variable X is normally $N(\mu, \sigma^2)$ distributed with arbitrary μ and σ^2.

H_1 : The random variable X is not normally distributed.

H_0 : Men and women have the same IQ.

H_1 : The IQ of men and women is not the same.

H_0 : In company X women earn at least as much money than men.

H_1 : In company X women earn less money than men.

Testing Theory

Summary

- For each H_0 we have to give an **alternative hypothesis** H_1 that is completely opposite to H_0.
- H_0 is the basis to decide whether we accept H_1 or not.

$\rightarrow$ The actual interesting hypothesis has to be formulated in H_1!!

Attention: The following formulation is wrong (why?)

H_0 : The reading ability of wealthy and lower-income children does not differ.

H_1 : Wealthy children can read better than lower-income children.

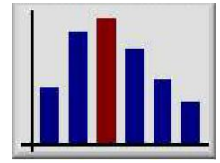
Testing Theory

$\rightarrow$ How to decide whether to reject H_0 or not?

Basic idea

Given a sample $x_1, \ldots, x_n$ of independent and identically distributed (iid) random variables we can calculate a specific value, also called **test statistic** $T = T(x_1, \ldots, x_n)$. Based on the value of T and its theoretical distribution we can make a decision about H_0.

The most popular method in deciding about H_0 based on T is to calculate the p value.



Testing Theory

The p value

The p value is the probability of obtaining a test statistic at least as extreme as the one that was actually observed, assuming that the null hypothesis is true.

- → The less probable the validity of H_0, the smaller gets the p value. If the probability falls below a certain limit, H_0 becomes so improbable that we cannot assume its validity any more. We decide to reject H_0 and say that we **significantly** accept H_1.
- → The most popular limit for the probability is 0.05, i.e. for p values smaller than 0.05 we reject H_0.
- → The p value can be interpreted as a measure of credibility for the null hypothesis.

Testing Theory

A statistical test allows **only one** of two decisions:

$$\textbf{rejection of } H_0 \;=\; \textbf{acceptation } H_1$$

or

$$\textbf{non-rejection of } H_0 \;\neq\; \textbf{acceptation of } H_0\textbf{.}$$

That means:

- $\rightarrow$ The non-rejection of H_0 can by no means be misinterpreted as a proof of the validity of H_0.
- $\rightarrow$ Strictly speaking, the non-rejection of H_0 can be interpreted as an abstention from voting, i.e. *the sample result is consistent with the null hypothesis*.

Testing Theory

"Statistics is never having to say you're certain" – making a decision in statistics always accompains with the risk of making the wrong decision:

	H_0 is true	H_0 not true
non-rejection for H_0	no error	type II error (β)
decision for H_1	type I error (α)	no error

- Significance tests can only control the type I error which is always ≤ 0.05.
- The type II error can become perhaps large.
- That is the reason why we have to put the proposition we want to proof into the alternative hypothesis.

Testing for Normality

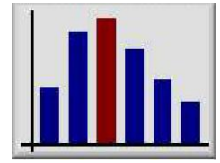
There are plenty of different tests for normality implemented in R. We will only use one of the most common tests, the **Shapiro-Wilk test**. Like all other tests for normality the null hypothesis of this test is

$$H_0 : \text{The sample is normally distributed}$$

Note that because of the formulation we are not interested in rejecting H_0. If the p value is under 0.05, we reject that the sample is normally distributed. Otherwise, we assume that the sample is normal even though this is not absolutely correct.

Testing for Normality

```
> # Shapiro-Wilk test for the height
> shapiro.test(manwoman$height.woman)

        Shapiro-Wilk normality test

data:  manwoman$height.woman
W = 0.9928, p-value = 0.4398
> # separated for the age groups (output omitted)
> tapply(manwoman$height.woman,
+ manwoman$age.woman.recoded, shapiro.test)
```



Exercises

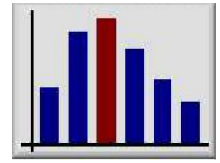
Examine whether the three variables `age.man`, `age.woman` and `height.man` can be approximated by a normal distribution. Take the following tools into your consideration:

(i) Histograms

(ii) Boxplots

(iii) Normal probability plots

(iv) Shapiro-Wilk test



Section 8: Metric Data

One Metric Variable

If the level of measurement of the data is metric, we distinct between the following scenarios for the statistical analysis:

- One metric variable
  - → One-sample t test
  - → Wilcoxon signed-rank test for one sample
- Two metric variables
  - Correlation hypothesis
    - → Pearson's correlation
    - → Spearman's correlation
  - Difference hypothesis
    - → Two-sample t test for dependent samples
    - → Wilcoxon signed-rank test for two samples

One Metric Variable

Assumptions

Given is a sample $x_1, \ldots, x_n$ of n independent and identically distributed (iid) observations of a $N(\mu, \sigma^2)$ distributed random variable with both unknown expectation μ and variance σ^2.

The corresponding null hypothesis is

$$H_0 : \mu = \mu_0$$

with a hypothetical value μ_0. The name of the test is the **one-sample t test**.

Example

A company that produces energy saving lamps affirms that the durability of their lamps averages 10.000 hours. In a long-term experiment with $n = 25$ lamps the burning time is taken.

One Metric Variable

Idea of the test

Under H_0 the average durability $\bar{x}_n$ should be near the hypothetical value μ_0. If the difference of both values is big, the validity of H_0 will be casted on doubt. If the difference is "too big", we reject H_0.

The test statistic of the one-sample t test is

$$T := \sqrt{n} \cdot \frac{\bar{x}_n - \mu_0}{S_n}.$$

Under H_0, T is t distributed with $(n-1)$ degrees of freedom.

Note that this is only true when the data is normally distributed. This has to be checked first!

One Metric Variable

Example

```
> # descriptive overview (output omitted)
> summary(lamps$burn.time)
> # One-sample t test (output omitted)
> t.test(lamps$burn.time, mu = 10000)
```

- The mean is with 9.744 hours smaller than than 10.000 (not good for the company).
- The p value is 0.2122, so H_0 is not rejected (good for the company).
- The 95% confidence interval of the expectation is $[9332.34, 10156.13]$. Since the value 10.000 is included H_0 is not rejected. However we get an impression about the "direction" of the deviation.

One Metric Variable

Assumptions

Given is a sample $x_1, \ldots, x_n$ of n iid observations with unknown median m.

- The assumption of normality is not demanded here.
- Tests that work without the assumption of a distribution are called **distribution free** or more often **nonparametric** tests.

The corresponding null hypothesis is

$$H_0 : m = m_0$$

with a hypothetical value m_0. The name of the test is **Wilcoxon signed-rank test**. Details for this test are given later.

One Metric Variable

Example

```
> # Wilcoxon signed-rank test
> wilcox.test(lamps$burn.time, mu = 10000)

        Wilcoxon signed rank test

data:  lamps$burn.time
V = 104, p-value = 0.1199
alternative hypothesis: true location is not ...
```

- The p value is 0.1199 and hence also larger than 0.05. The null hypothesis is not rejected as well.

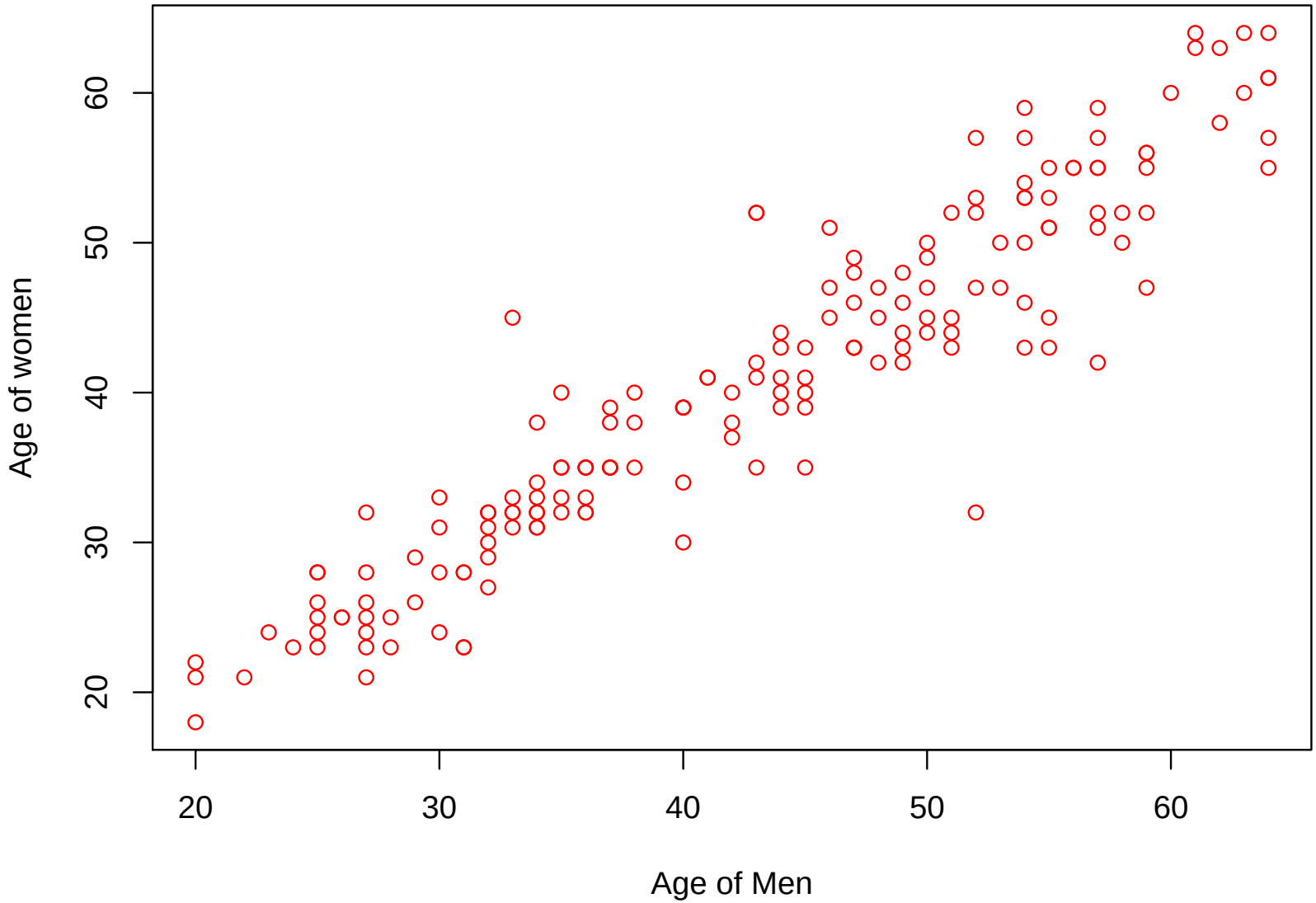
Two Metric Samples – Correlation Hypothesis

Assumptions

Given is a sample of two iid observations of two metric random variables that are arranged as couples $(x_1, y_1), \ldots, (x_n, y_n)$.

The structure of dependence between two random variables can be investigated graphically with a **scatterplot** where the two samples are plotted against each other in a diagram.

The higher the relationship between the two samples, the higher is the tendency for a structure in the scatterplot. As an example, take a look at the scatterplot of the age of men and women in the data set `manwoman` on the next slide.


Age of women
20
30
40
50
60
Age of Men
20
30
40
50
60

Two Metric Samples – Correlation Hypothesis

```
> # scatterplot of age.man and age.woman
> plot(manwoman$age.man, manwoman$age.woman,
+ xlab = "Age of Men", ylab = "Age of women",
+ col = "red")
```

→ With `col` the color of the points is changed to red.

→ The argument `pch` changes the point style in the plot. Try for example the setting `pch = 17` or other numbers.

Two Metric Samples – Correlation Hypothesis

In order to quantify the linear dependency between two variables, we have to get familiar with the **correlation** between two variables. The correlation between two random variables X and Y is defined as:

$$Corr(X, Y) = \frac{Cov(X, Y)}{\sigma_X \cdot \sigma_Y} \in [-1; 1].$$

- $\rightarrow$ The correlation is standardized on the interval $[-1; 1]$ which is why it can easliy be interpreted.
- $\rightarrow$ The standardization also allows the comparison of two correlation coefficients, no matter which measures are considered.

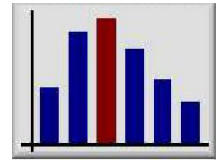
Two Metric Samples – Correlation Hypothesis

Interpretation of the correlation

A high positive (negative) correlation means that an above-average high value of X accompains with an above-average high (low) value of Y.

Guidelines for the strength of a correlation:

- $Corr(X, Y) \approx 0$: negligible linear relation between X and Y.
- $0.3 < |Corr(X, Y)| < 0.7$: weak linear relation between X and Y.
- $|Corr(X, Y)| > 0.7$: strong linear relation between X and Y.



Two Metric Samples – Correlation Hypothesis

We have

$$X \text{ and } Y \text{ independent} \quad \Rightarrow \quad X \text{ and } Y \text{ uncorrelated.}$$

But be careful:

$$X \text{ and } Y \text{ uncorrelated} \quad \Rightarrow \quad X \text{ and } Y \text{ independent.}$$

is NOT true in general!

Caution

The correlation measures only the **linear** relation. There are also other kinds of relation between variables, e.g. quadratic or logarithmic relation.

Two Metric Samples – Correlation Hypothesis

The correlation is a theoretical value that is unknown for the data. Like for expectation and variance, we have to estimate the correlation based on the available data. We do can this by determining the **Pearson correlation coefficient**

$$r_P := \frac{\widehat{Cov}(x,y)}{\hat{\sigma}_x \hat{\sigma}_y} = \frac{(\frac{1}{n}\sum_{i=1}^n x_i y_i) - (\frac{1}{n}\sum_{i=1}^n x_i)(\frac{1}{n}\sum_{i=1}^n y_i)}{\sqrt{(\frac{1}{n}\sum_{i=1}^n (x_i-\bar{x})^2)(\frac{1}{n}\sum_{i=1}^n (y_i-\bar{y})^2)}}.$$

Interpretation of r_P

If X increases in one unit, the value of Y changes in r_P units. Depending on the sign of r_P, the value of Y increases or decreases in r_P units.

Two Metric Samples – Correlation Hypothesis

Beside the calculation of r_P, we can also conduct a statistical test to check whether the relation between X and Y is significant.

Assumptions

Given is a sample of n iid observations of two metric random variables that are arranged as couples $(x_1, y_1), \ldots, (x_n, y_n)$. Furthermore, the two samples are normally distributed, i.e. $x_1, \ldots, x_n \sim N(\mu_X, \sigma^2)$ and $y_1, \ldots, y_n \sim N(\mu_Y, \sigma^2)$.

$\rightarrow$ Note that it is not enough for the data to be metrically scaled. They also have to be normally distributed!

Two Metric Samples – Correlation Hypothesis

The corresponding null hypothesis is

$$H_0 : r_P = 0,$$

i.e. it is investigated whether there is a dependence between X and Y at all. The test statistic

$$T := \frac{r_P}{\sqrt{1 - r_P^2}} \sqrt{n-2}$$

is under H_0 t distributed with $(n-2)$ degrees of freedom. If H_0 is rejected, we can derive the direction of relation by the sign of r_P or – even better – by the confidence interval.

Example

```
> cor.test(manwoman$age.man, manwoman$age.woman)

        Pearson's product-moment correlation

data:  manwoman$age.man and manwoman$age.woman
t = 35.2493, df = 168, p-value < 2.2e-16
alternative hypothesis: true correlation is not ...
95 percent confidence interval:
 0.9176836 0.9542678
```

- H_0 is rejected ($p < 0.001$), the correlation is significant.
- The confidence interval is $[0.92, 0.95]$ which indicates a positive relationship.

Two Metric Samples – Correlation Hypothesis

If the data is not normally or only ordinally scaled, we have to use another correlation coefficient:

Assumptions

Given is a sample of n iid observations of two metric random variables that are arranged as couples $(x_1, y_1), \ldots, (x_n, y_n)$. The data does not have to be normally distributed.

The Pearson correlation is not suitable in this case. We have to calculate the **Spearman rank correlation coefficient**, r_S.

Two Metric Samples – Correlation Hypothesis

Calculation of r_S

- Order the two samples $x_1, \ldots, x_n$ and $y_1, \ldots, y_n$ respectively according to their size in ascending order.
- Each value x_i and y_i is assigned to a rank $r_{x,i}$ and $r_{y,i}$ according to its order in the sample.
- The Spearman rank correlation coefficient is calculated as follows:

$$r_S := \frac{6 \sum_{i=1}^{n} (r_{x,i} - r_{y,i})^2}{n(n^2 - 1)} \in [-1; 1],$$

which is exactly the Pearson correlation coefficient for the rank values.

Two Metric Samples – Correlation Hypothesis

Similar to the Pearson correlation, it is also possible to test the null hypothesis

$$H_0 : r_S = 0$$

that investigates whether the two variables are dependent from each other – no matter in which direction.

The test statistic

$$T := \frac{r_S}{\sqrt{1 - r_S^2}} \sqrt{n - 2}$$

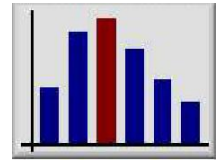
is for $n > 30$ approximately t distributed with $(n - 2)$ degrees of freedom. If $n \leq 30$ the p value is determined based on implemented tabluar values.

Two Metric Samples – Correlation Hypothesis

Example

```
> # Spearman correlation
> cor(manwoman$age.man, manwoman$age.woman,
+ use = "complete.obs", method = "spearman")
[1] 0.9399444
> # test for Spearman correlation (output omitted)
> cor.test(manwoman$age.man, manwoman$age.woman,
+ method = "spearman")
```

- For the function `cor()` we have to specify the argument `use` in case of missing values.
- H_0 is rejected ($p < 0.001$), the correlation is significant.
- The test result for the Spearman correlation delivers no confidence interval.



Two Metric Samples – Difference Hypothesis

If the aim is not to reveal a relationship between to paired measures but a **difference**, we have to use different test procedures.

Assumptions

Given is a sample of n iid observations of two metric random variables that are arranged as couples $(x_1, y_1), \ldots, (x_n, y_n)$. The **pairwise differences** $d_i = x_i - y_i, i = 1, \ldots, n$ are normally distributed.

$\rightarrow$ Note that it is not enough to show that the two samples are normally distributed. From the normality of each of the two variables does not necessary follow the normality of the difference!

Two Metric Samples – Difference Hypothesis

Example

The blood pressure of $n = 35$ patients is measured before and after the intake of a blood pressure decreasing drug. Aim is to check whether the blood pressure really decreases.

The null hypothesis is

$$H_0 : \mu_X = \mu_Y \quad \Longleftrightarrow \quad \mu_X - \mu_Y = 0.$$

For the example with the blood pressure it means that the drug has no effect. The test is called **two sample t test for paired samples**.

Two Metric Samples – Difference Hypothesis

Idea of the test

As already mentioned, the test is based on the pairwise differences d_i. Under H_0 the two variables should be approximately equal and hence the difference should be near 0. The arithmetic mean of the differences $\bar{d} = \frac{1}{n}\sum_{i=1}^{n} d_i$ should under H_0 consequently near 0 too. In other words, we have one sample so we can conduct the one-sample t test with the theoretical value $\mu_0 = 0$.

Hence, the test statistic is

$$T := \sqrt{n} \cdot \frac{\bar{d}}{\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(d_i - \bar{d})^2}}.$$

Two Metric Samples – Difference Hypothesis

Example

```
> # boxplot of the data
> par(mfrow = c(1,2))
> boxplot(manwoman$height.man)
> boxplot(manwoman$height.woman)
> par(mfrow = c(1,1))
> # test for normality (output committed)
> diff <- manwoman$height.man - manwoman$height.woman
> shapiro.test(diff)
> # t-Test (output committed) with t.test(diff) or
> t.test(manwoman$height.man, manwoman$height.woman,
+ paired = T)
```

There is a significant difference in the height of men and women ($p < 0.001$).

Two Metric Samples – Difference Hypothesis

If the data is not normally distributed, we have to do a nonparametric alternative test instead of the two sample t test for paired samples.

Assumptions

Given is a sample of n iid observations of two metric random variables that are arranged as couples $(x_1, y_1), \ldots, (x_n, y_n)$.

The test for this situation is the **Wilcoxon signed-rank test**. The null hypothesis is

$$H_0 : X_i - Y_i \text{ has the median } 0.$$

This is basically the same hypothesis than with the corresponding t test.

Two Metric Samples – Difference Hypothesis

Idea of the test

- Calculate $|d_i| = |x_i - y_i|$ and assign corresponding ranks R_i.
- Determine the sum of ranks R_+ of positive d_i and the sum of ranks R_- of negative d_i.
- Under H_0 the signs of the d_i should appear approximately with equal frequencies. If R_+ or R_- is "too big", H_0 is rejected.

The test statistic

$$Z := \min\{R_+, R_-\}$$

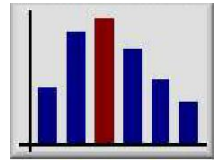
is under H_0 for $n > 25$ approximately normally distributed.

Two Metric Samples – Difference Hypothesis

Example

```
> # Wilcoxon signed-rank test
> diff <- manwoman$age.man - manwoman$age.woman
> wilcox.test(diff)

        Wilcoxon signed rank test with ...

data:  diff
V = 9460, p-value = 3.977e-12
alternative hypothesis: true location is not ...
```

- The difference between the age is significant ($p < 0.001$).
- No confidence intervall is given.



Exercises

1. data set `lamps`: What if the company affirms that the durability of their lamps averages 10.500 hours? Is this still in accordance with the null hypothesis?
2. data set `manwoman`: The average height of German women is 165 cm and the average height of German men is 180 cm. Do the average height values of the data set signifcantly differ from those numbers?
3. data set `sudan`: The average German BMI is 25.7. Do the patients of the dataset `sudan` have a siginficant lower BMI than German people?

Exercises

4. data set `pisa`:
   (i) Between which of the three performance parameters is the linear relation the highest? Before calculating the correlation coefficient, check which one is suitable in this case.
   (ii) Let μ_{math} and μ_{reading} be the expectation of the mathematic and the reading performance. Check the null hypothesis

$$H_0 : \mu_{\text{math}} = \mu_{\text{reading}}$$

   with a suitable testing procedure.

5. data set `cinema`: Is there a relationship between the age and the number of visits? For this purpose, visualize the data and conduct a significance test.

Exercises

6. data set `soccer`: The data set contains the number of scored goals for the first (`goals.ht1`) and the second (`goals.ht2`) half time of the 18 clubs in the "Bundesliga" (season 2009/2010).
   (i) Is there a (significant) difference in the number of scored goals between the first and the second half time?
   (ii) Is there a linear dependence between the points of each team and their budget? Investigate this graphically and with a suitable significance test.
7. data set `olympia`: The data set contains the results of the best athlets in decathlon of the Olympic Games 2014 in London. Is there a relationship between the time of the 100-meters race and the result of the long jump?

Exercises

8 data set `salary`: The data set contains information about the average salary (fixed and variable component) of the board members of German companies listed in the DAX. Further, the companies earings are shown.

(i) Is there a relationship between the salary and the earnings of the company for 2006 and 2007? Consider both salary parts separately and also the sum of them.

(i) Is there a relationship between the earnings of the board members in 2007 and the company's performance in 2006?

Section 9: Categorial Data

χ^2 Test for Independence

Assumptions

Given is a sample of n iid observations of two categorical random variables that are arranged as couples $(x_1, y_1), \ldots, (x_n, y_n)$.

Plotting a scatterplot for categorical data is of course not meaningful. Alternatively, we can try to summarize the information in the data with a **contingency table**.

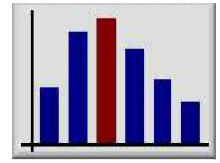
Besides the single cell frequency we are also interested in the total frequencies for each value of the variable.

χ^2 Test for Independence

Example

$X :=$ eye color: $X \in \{\text{blue, brown, green, beige}\}$
$Y :=$ hair color: $Y \in \{\text{blond, brown, red, black}\}$

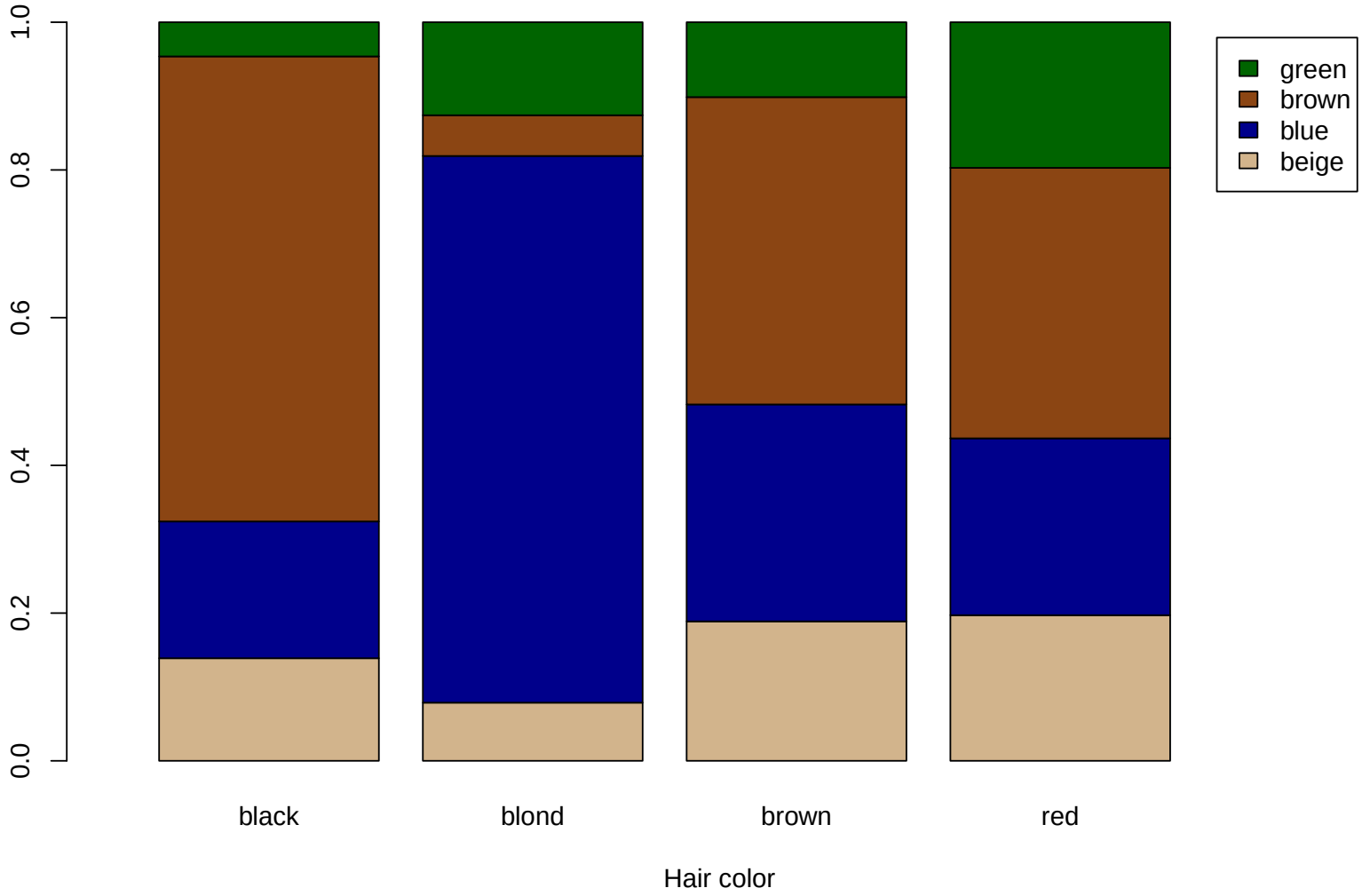
	blond	brown	red	black	total
blue	94	84	17	20	215
brown	7	119	26	68	220
green	16	29	14	5	64
beige	10	54	14	15	93
total	127	286	71	108	592

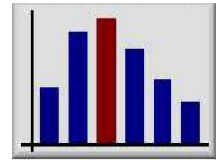


χ^2 Test for Independence

```
# save the table
freq.haireye <- table(haireye)
# add the row and column sums
freq.haireye <- addmargins(freq.haireye)
freq.haireye
```

χ^2 Test for Independence





χ^2 Test for Independence

```
> # table with row-wise relative frequencies
> graphic.haireye <- prop.table(table(haireye), 2)
> # barplot with relative frequencies
> barplot(graphic.haireye, xlab = "Hair color",
+ col = c("tan", "darkblue", "saddlebrown",
+ "darkgreen"),
+ legend = c("beige", "blue", "brown", "green"),
+ xlim = c(0, 5.5))
```

χ^2 Test for Independence

The test to investigate whether the two categorical variables X and Y are statistically related is called χ^2 **Test for Independence**. The null hypothesis is as follows:

$$H_0: \quad X \text{ are } Y \text{ independent from each other.}$$

Idea of the test

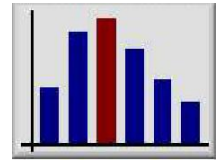
Under H_0 it is possible to calculate **expected frequencies** for each cell of the contingency table. The expected frequencies are compared with the **observed frequencies**. The bigger the deviation in the cells, the more H_0 is casted on doubt.

χ^2 Test for Independence

Let X be able to take on I categories and Y be able to take on J categories. Moreover, let O_{ij} be the observed frequency in the (i, j)-th cell and E_{ij} the corresponding expected frequency. The test statistic given by

$$X^2 := \sum_{i=1}^{I} \sum_{j=1}^{J} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}.$$

is approximately χ^2 distributed with $(I-1)(J-1)$ degrees of freedom.



χ^2 Test for Independence

Since X^2 is only approximately χ^2 distributed, the following rule should be fulfilled so that the approximation is sufficiently exact.

Chochran's rule

At least 80% of the expected cell count should be 5 or more and no expected cell count should be less than 1.

If this recommendation is injured we can:

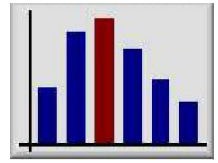
- either summarize two categories to one or
- ignore one or more categories with very low observations.

χ^2 Test for Independence

Example

```
> # chi-square test
> chisq.test(haireye$hair, haireye$eye)

        Pearson's Chi-squared test

data:  haireye$hair and haireye$eye
X-squared = 138.2898, df = 9, p-value < 2.2e-16

> # expected frequencies (output omitted)
> chisq.test(haireye$hair, haireye$eye)$expected
```

Since $p < 0.001$, there is a significant relationship between both variables.



χ^2 Test for Independence

If $p < 0.05$, the question arises how to correctly interpret this result. For a better understanding, we can look at the standardized residuals.

Example

```
> # interpretation of the test result
> # using standardized residuals (output omitted)
> chisq.test(haireye$hair, haireye$eye)$stdres
```

In a sloppy speech, we try to locate the significance more exactly by looking at the standardized residuals. The rule of thumb is that if $|\text{res}| \geq 2$, the combination is high and potentially co-responsible for the global signifiance.

Fisher's Exact Test

A special case is if both variables are **binary**, i.e. can both take on only two categories.

Assumptions

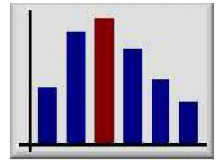
Given is a sample of n iid observations of two categorical random variables that are arranged as couples $(x_1, y_1), \ldots, (x_n, y_n)$. Both underlying random variables are binary, i.e. $I = J = 2$.

In this case we can avoid the approximative test procedure an do the **Fisher's exact test** instead. As its name implies, the test is not approximative but exact so we do not have to care about Chochran's rule. However, H_0 stays the same.

Fisher's Exact Test

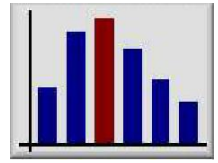
Example

```
> table(cinema$age.recoded, cinema$sex)

       female male
  young      6    2
  old        3    7
> fisher.test(cinema$age.recoded, cinema$sex)

        Fisher's Exact Test for Count Data

data:  cinema$age.recoded and cinema$sex
p-value = 0.1534
alternative hypothesis: true odds ratio is not...
```



Exercises

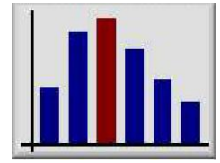
1. data set `titanic`: The data set contains informations about the class membership, the gender and the age of all passengers of the titanic.
   - Is there a relationship between the survival and the class membership?
   - Is there a relationship between the survival and the gender?
   - Is there a relationship between the survival and the age? To answer this question recode the age into two groups: passengers until 14 years and passenger over 14 years.
2. data set `suicide`: The two variables are the gender of people who committed suicide and the way of suicide. Is there a relationship between both measures?



Exercises

3. data set `sudan`: Is there a relationship between the membership to the three groups „malaria“, „diabetes“ and „malaria & diabetes“ in the variable `group` and the gender of the patients? Is there a relationship between one of the two diseases alone and the gender of the patients?
4. data set `income`: The data set contains information of a large survey of the U.S. population. The question of interest is whether people living in an urban region do have a statistically lower or higher income than people living in a rural region. Conduct a statistical test to examine this question.

Section 10: Categorial and Metric Data



Two Samples

Depending on the number of possible values of the categorical variable, we distinct between the following scenarios:

- Difference hypothesis with two samples
  - → Two-sample t test for independent samples
  - → Wilcoxon rank-sum test
- Difference hypothesis with more than two samples
  - → Oneway ANOVA
  - → Kruskal-Wallis test

Two Samples

Assumptions

Given are two independently measured samples $x_1, \ldots, x_m$ and $y_1, \ldots, y_n$ of iid observations. The x_i follow a $N(\mu_X, \sigma^2)$ distribution and the y_i a $N(\mu_Y, \sigma^2)$ distribution with equal variances σ^2.

The null hypothesis is that the expectations are equal, i.e.

$$H_0 : \mu_X = \mu_Y.$$

The test is called **Two-sample t test for independent samples**.

Example

The coagulation time of two different medicines is measured for two different patient groups. The question is if one medicine has a lower coagulation time.

Two Samples

Idea of the test

Under H_0, the two arithmetic means $\bar{x}_m$ and $\bar{y}_n$ of the two samples should be approximately equal. If the difference of both values is big, the validity of H_0 will be casted on doubt. If the difference is "too big", we reject H_0.

The test statistic of the two-sample t test for independent samples is

$$T := \frac{\bar{x}_m - \bar{y}_n}{\sqrt{\left(\frac{1}{m} + \frac{1}{n}\right) \cdot \left(\frac{(m-1)S^2_{X,m} + (n-1)S^2_{Y,n}}{m+n-2}\right)}},$$

under H_0, T is t distributed with $(m + n - 2)$ degrees of freedom.

Two Samples

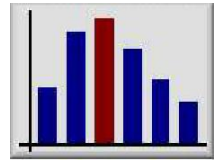
Before we can apply the t test to the data, we have to make sure that the variances of both samples are equal. This can be done with the F **test**. The null hypothesis of the test is

$$H_0 : \text{Both samples have the same variance.}$$

The test statistics is as follows

$$Q = \frac{\max\{S^2_{X,m}, S^2_{Y,n}\}}{\min\{S^2_{X,m}, S^2_{Y,n}\}},$$

where $S^2_{X,n}$ and $S^2_{Y,n}$ are the sample variances of the two samples.



Two Samples

Example

```
> # groupwise mean and standard deviation
> tapply(coagulation$time, coagulation$med, mean)
       A        B
8.750000 9.742857
> tapply(coagulation$time, coagulation$med, sd)
        A         B
0.5822371 0.8182443
> # F test for equality of variances (no output)
> var.test(coagulation$time ~ coagulation$med)
```

- The result is not significant ($p = 0.472$).
- Note the **model formula** with the tilde (~) in the argument of `var.test()`.

Two Samples

If the F test does not reject the equality of variances, the t test for independent samples is conducted.

Otherwise we cannot use the t test. In this case there is only an approximative test, called the **Welch test**. The test statistic is based on the T statistic with some modifications which are of less interest for us.

Note that even if the variances for this test do not have to be equal, the samples still have to be normally distributed!

Two Samples

Example

```
> # equal variances: t test (no output)
> t.test(coagulation$time ~ coagulation$med,
+ var.equal = T)
> # assume unequal variances: Welch test (no output)
> t.test(coagulation$time ~ coagulation$med)
```

- According to the p value of 0.031 there are significant differences in the coagulation times.
- As the group means indicate the time in group A is lower which also shows the confidence interval of $[-1.88, -0.11]$.
- The result of the Welch test is similar ($p = 0.028$).

Two Samples

If the data is non-normal, we have to do a nonparametric testing procedure.

Assumptions

Given are two independently measured samples $x_1, \ldots, x_m$ and $y_1, \ldots, y_n$ of iid observations.

The test to conduct is called **Wilcoxon rank-sum test**. Its null hypothesis is given by:

H_0 : The distribution functions of both samples are equal

This null hypothesis is basically the same than for the t test, hence the interpretation of the test results stays the same.

Two Samples

Idea of the test

- Combine the two samples to **one** sample and assign a rank to each observation.
- Calculate for each group the rank sum R_x and R_y as well as $U_x = mn + \frac{m(m+1)}{2} - R_x$ and $U_y = mn + \frac{n(n+1)}{2} - R_y$.
- Under H_0 the rank sums R_x and R_y should be approximately equal. If R_x or R_y is "too" big, H_0 is rejected.

The test statistic

$$U := \min\{U_x, U_y\}$$

is under H_0 for sufficient large n approximately normally distributed.

Two Samples

Example

```
> # Wilcoxon rank-sum test
> wilcox.test(coagulation$time ~ coagulation$med)

        Wilcoxon rank sum test ...

data:  coagulation$time by coagulation$med
W = 7, p-value = 0.05313
```

- The null hypothesis is not rejected in this case, even though the t test revealed differences for the two groups!
- This example shows that the parametric test (t tests) are to prefer to the nonparametric tests because the former have a higher power, as seen in this example.

More Than Two Samples

Assumptions

Given are $I > 2$ independently measured samples $x_{i1}, \ldots, x_{iJ_i}$ of iid observations. The observations of the i-th sample $x_{ij},\ j = 1, \ldots, J_i$ follow a $N(\mu_i, \sigma^2)$ distribution. The variances are equal for each sample.

$\rightarrow$ With the two-sample t test we could only examine the difference of two samples.

$\rightarrow$ The now presented procedure is the generalization of the t test and is called **Analysis of Variance (ANOVA)**.

More Than Two Samples

The independent variable that appears in I categories is also called **factor**, the single categories are called **factor levels**. Since only the influence of one factor on the dependent (metric) variable is investigated, the procedure is also called **oneway ANOVA**.

The null hypothesis is as follows:

$$\mu_1 = \mu_2 = \ldots = \mu_I.$$

That means that there are no differences in the expectation between the I factor levels.

More Than Two Samples

Example

The influence on the performance of pupils for four different education methods is investigated. To this end, four classes are educated with four different methods for one year. At the end each pupil writes a final exam. The achieved points in the test are documented for each pupil:

authoritarian	corporal punishment	group work	laissez faire
25.5	25	27	22
23.5	19	26.5	26
⋮	⋮	⋮	⋮

More Than Two Samples

In order to use the ANOVA correctly, the assumptions have to be fulfilled:

Assumptions of the ANOVA

1. Samples have to be independent from each other.
2. The i-th sample ($i = 1, \ldots, I$) follows a $N(\mu_i, \sigma^2)$ distribution. This can be checked with the usual methods (graphical tools, Shapiro-Wilk test)
3. The variance is equal for all samples, which can be tested by the **Bartlett test** that is a generalization of the F test. The null hypothesis is that the variances of all samples are equal.

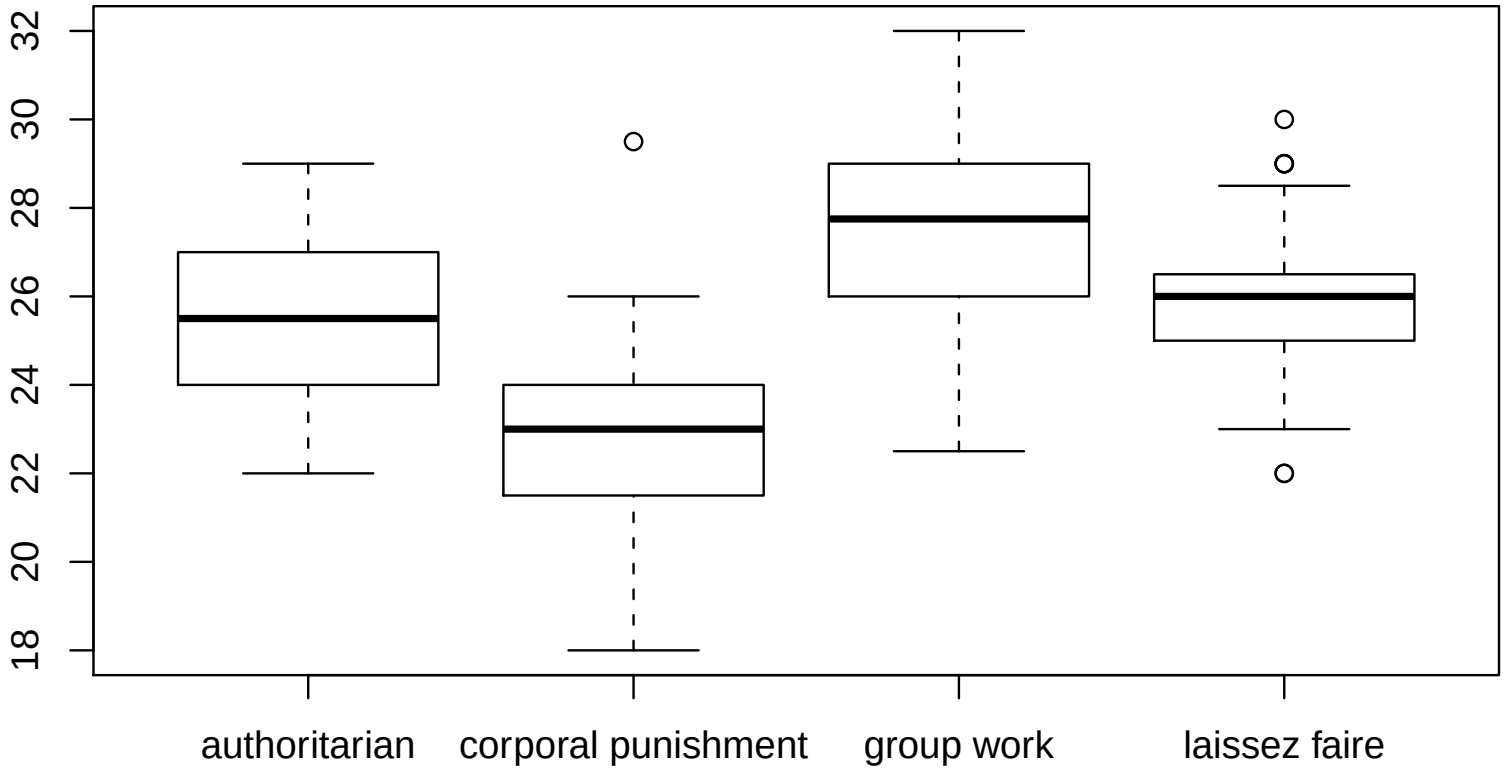
More Than Two Samples

Example

```
> # Boxplot of the data
> plot(education$method, education$test.result)
> # Shapiro-Wilk test (output omitted)
> tapply(education$test.result, education$method,
+ shapiro.test)
> # Bartlett test for homogeneity of variances
> bartlett.test(education$test.result ~
+ education$method)
```

- Samples seems to be normally distributed. No p value of the Shapiro-Wilk test is bigger than 0.05
- Assumption of heterogeneity of variances is fulfilled. Bartlett test not significant ($p = 0.55$)

More Than Two Samples



More Than Two Samples

Idea of the Test

Let x_{ij} be the j-th observation of the i-th sample and $\bar{x}$ the **overall mean** and $\bar{x}_i$ the i-**th group mean**. Then we have:

$$x_{ij} = \bar{x} + \underbrace{(\bar{x}_i - \bar{x})}_{\substack{\text{deviation group mean} \\ \text{of overall mean}}} + \underbrace{(x_{ij} - \bar{x}_i)}_{\substack{\text{deviation observation} \\ \text{of overall mean}}}$$

If H_0 is not true the deviation of the group mean will be relatively large compared to the deviation of the observations to the group mean.

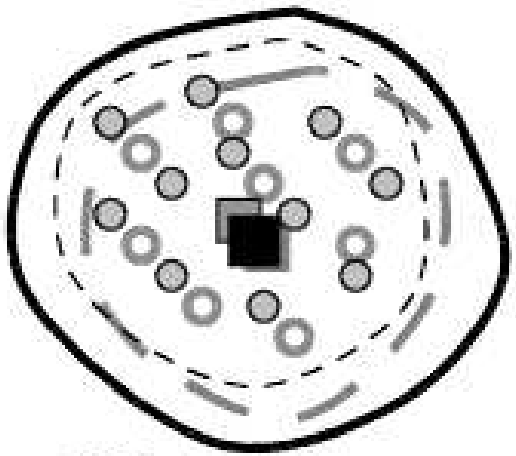
More Than Two Samples

The test statistic is also based on these considerations:

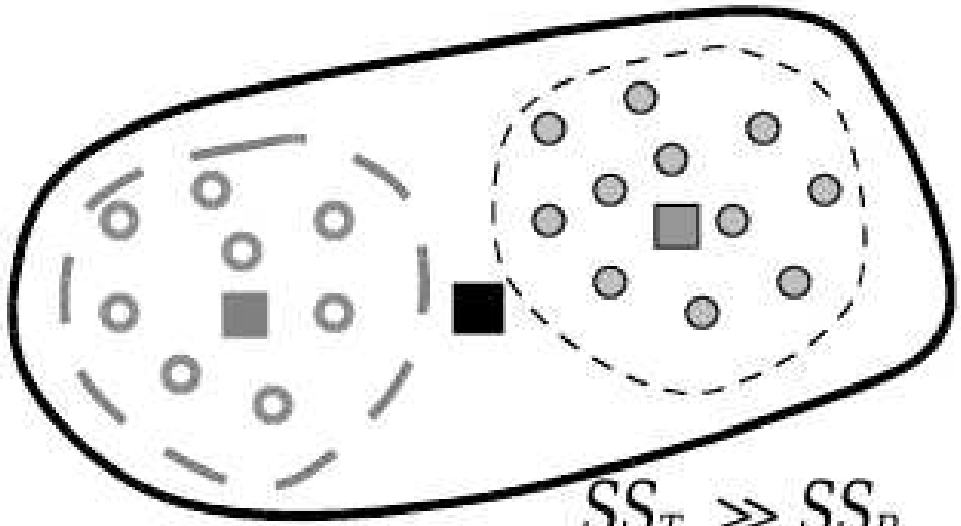
$$F_{0,\alpha} := \frac{\frac{1}{I-1} \cdot SS_A}{\frac{1}{n-1} \cdot SS_R} = \frac{\frac{1}{I-1} \cdot J \sum_{i=1}^{J} (\bar{x}_i - \bar{x})^2}{\frac{1}{n-1} \cdot \sum_{i=1}^{I} \sum_{j=1}^{J} (x_{ij} - \bar{x}_i)^2}.$$

The more the means of the single factor levels deviate from the overall mean, the bigger is getting SS_A compared to SS_R. Thus under H_0 the ratio of $\frac{SS_A}{SS_R}$ should be near 0. The bigger SS_A gets – and hence the bigger the ratio gets – the less probable is the validity of H_0. If F gets "too big", H_0 is rejected.

More Than Two Samples



$SS_T \approx SS_R$



$SS_T \gg SS_R$

More Than Two Samples

Example

```
> # ANOVA for homogeneity of variances
> anova(lm(education$test.result ~ education$method))
Analysis of Variance Table

Response: education$test.result
                  Df Sum Sq Mean Sq F value    Pr(>F)
education$method   3 358.58 119.528  26.986 2.593e-13
Residuals        116 513.78   4.429
```

- Significant result ($p < 0.001$): difference (somewhere) between the expectation of the education methods.

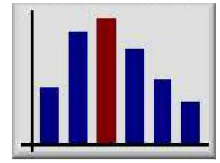
More Than Two Samples

If the samples do not have the same variance we use the generalization of the Welch test:

Example

```
> oneway.test(education$test.result ~
+ education$method)

        One-way analysis of means (not assuming ...)

data:  education$test.result and education$method
F = 22.1985, num df = 3.00, denom df = 64.19,
p-value = 5.624e-10
```

- Significant result ($p < 0.001$): difference (somewhere) between the expectation of the education methods.



More Than Two Samples

The ANOVA is only a **global test** which means that we can only find out whether there is **any** difference in the means. It is not possible the say where these differences are exactly.

→ Question: How to localize the significant differences between the factor levels (if any)?

→ Answer: **Posthoc analysis**.

Definition

Posthoc analyses are pairwise comparisons between the means of the different factor levels. In case of a significant result of the ANOVA, it is possible to examine the factor levels for differences via posthoc procedures.

More Than Two Samples

An intuitive approach would be to conduct two-sample t tests for each combination of pairwise comparisons.

$\rightarrow$ Caution: Doing so, the type I error increases very fast!

$\rightarrow$ Solution: **Adjustment of the significance level**.

Bonferroni-Holm correction

For m pairwise comparisons the adjusted p value is given by $p_i^{\text{Holm}} = p_{(i)} \cdot (m - i + 1)$, where $p_{(1)}, \ldots, p_{(m)}$ are the ordered p values of the m test. The adjustment is called **Bonferroni-Holm correction**.

More Than Two Samples

Example

```
> # Bonferroni-Holm correction (output omitted)
> pairwise.t.test(education$test.result,
+ education$method)
```

- The output is given in a matrix with all factor level combinations and the adjusted p values
- Every difference in the mean between the education methods is significant, except of the difference between the methods `authoritarian` and `laissez faire` ($p = 0.561$).

More Than Two Samples

A disadvantage of the adjustment procedure is that the test becomes very **conservative**, i.e. the differences have to be relatively large to be revealed.

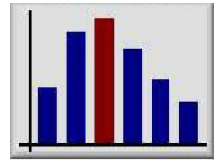
Another posthoc test is the **Tukey test** that is less conservative and thus should be preferred to the Bonferroni-Holm correction. The Tukey test also keeps the type I error under 5%. However the Tukey test should only be used if the design is **balanced**, i.e. the number of observations is the same for all factor levels. Additionally the assumption of homogeneity of variances should be fulfilled.

More Than Two Samples

Example

```
> # Tukey test (output omitted)
> TukeyHSD(aov(education$test.result ~
+ education$method))
```

- The output is given in a list, the adjusted p value is in the last column.
- Confirms the results that the only non-significant difference is between `authoritarian` and `laissez faire` ($p = 0.937$).
- If the variances are unequal the suitable call is `pairwise.t.test(education$test.result, education$method, pool.sd = F)`.



More Than Two Samples

A special situation in pairwise comparisons appears, if only one group should be compared to all other but no other comparisons. This design typically appears for example when different treatment groups and one control group is measured and only the comparison to the control group is of interest. Here, the **Dunnett test** can be performed that compares a reference group to all other groups.

The test is not contained in the base packages of R but in the package **multcomp**. We have to load the package with `library()` (and install it first). Next step is the definition of the linear model like in the usual ANOVA. The difference is that the constant term must not be considered which is why we substract the model formular with 1.

More Than Two Samples

```
> library(multcomp)
> wine.anova <- lm(wine$ajudgement ~ wine$rater - 1)
```

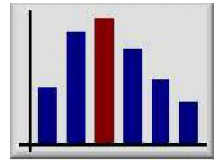
For the testing, we also have to create a matrix `K` that contains the information which of the factor levels is considered as the control group. In the next example, we define the first factor level as the control group (value -1).

```
> K <- rbind(c(-1, 1, 0, 0, 0),
+            c(-1, 0, 1, 0, 0),
+            c(-1, 0, 0, 1, 0),
+            c(-1, 0, 0, 0, 1))
```

More Than Two Samples

For a better readability of the output, we define row and columnnames for `K`

```
> rownames(K) <- c("Amthor vs. Muck",
+  "Amthor vs. Munter", "Amthor vs. Reiss",
+  "Amthor vs. Stahl")
> colnames(K) <- names(coef(wine.anova))
```



More Than Two Samples

The Dunnett test is finally performed using the function `glht()` and the results can be shown with `summary()`

```
> # Dunnett test (output omitted)
> summary(glht(wine.anova, linfct = K))
```

We get for every comparison a p-value that is alread adjusted. The result of taster "Amthor" differs only significant to the result of taster "Muck" ($p < 0.001$).

More Than Two Samples

If not all factor levels are normally distributed, we have to use again a nonparametric alternative.

Assumptions

Given are $I > 2$ independently measured samples $x_{i1}, \ldots, x_{iJ_i}$ of iid observations. The observations do not necessarily have to be normally distributed.

The corresponding test is called **Kruskal-Wallis test**. Its null hypothesis is:

$$H_0: \text{ All samples have the same distribution function}$$

This means in particular that the expectations are equal which was H_0 in the parametric case.

More Than Two Samples

Idea of the test

- Combine all samples to **one** sample and assign a rank to each observation.
- Calculate the **rank overall mean** $\bar{R}$ and the i **rank group mean** $\bar{R}_i$.
- Under H_0 the difference of $\bar{R}$ and the $\bar{R}_i$ should be small. If the sum of the differences

$$SRS_A = \sum_{i=1}^{J} J(\bar{R}_i - \bar{R})^2$$

gets big, H_0 is casted on doubt.

More Than Two Samples

The test statistic

$$\frac{12}{n(n+1)} SRS_A$$

is under H_0 approximately χ^2 distributed with $I-1$ degrees of freedom.

The Kruskal-Wallis test is not an exact test but works only approximative. In order to keep the error small, the sample size has to fulfill the following rule:

Rule for the sample size of the Kruskal-Wallis test

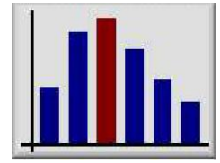
- If $I = 3$: $n_1, n_2, n_3 \geq 5$.
- If $I \geq 4$: $n_1, \ldots, n_I \geq 4$.

More Than Two Samples

Example

```
> # Kruskal-Wallis test
> kruskal.test(education$test.result ~
+ education$method)

        Kruskal-Wallis rank sum test

data:  education$test.result by education$method
Kruskal-Wallis chi-squared = 48.9163, df = 3,
p-value = 1.359e-10
```

- The Kruskal-Wallis test is significant ($p < 0.001$).



More Than Two Samples

The Kruskal-Wallis test is only a global test. In case of a significant test result, we have to run additional posthoc analysis.

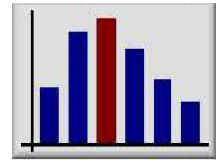
Unfortunately, in the nonparametric case, we do not have the Tukey test available. Hence we have to run the Wilcoxon rank-sum test with a suitable significance correction.

More Than Two Samples

Example

```
> # Wilcoxon test as posthoc procedure
> pairwise.wilcox.test(education$test.result,
+ education$method)
```

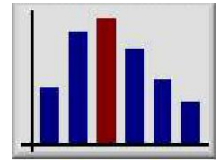
- The output is a matrix like for the function `pairwise.t.test()`.
- The only nonsignificant comparison is between `authoritarian` and `laissez faire` ($p = 0.583$).
- The warning messages are not of importance.



Exercises

1. data set `ph`: To examine the influence of chalking on the pH value of soil, soil samples of chalked and unchalked forest soil are taken. Is there a significant difference in the pH value for chalked and unchalked forest soil?
2. data set `cinema`: Check the following null hypothesis:
   (i) H_0 : Men and woman are of the same age
   (ii) H_0 : Men and woman visited the cinema equally often
3. data set `urine`: The data set contains the ph and the ca value of patient with and without crystals in their urine. Is there a significant difference in
   (i) the ph values
   (ii) the ca values

   of the observed patients?



Exercises

4. data set `sudan`:
   (i) Is there a (significant) difference between male and female patients according to their age? What about BMI and blood glucose?
   (ii) Investigate the same questions from part (i) with the independet variable `group`.
5. data set `ph`: Besides the influence of chalking it is also of interest whether different watering with the three levels `none`, `acidic` and `normally` has an influence on the pH value. Does this factor has a significant effect on the ph value?

Exercises

6 data set `wine`: The data set shows the assessments of wine tasters for different wine types. The final assessment of each wine type is the mean between the three criterions aroma, taste and harmony. Some of the tasters are under the suspicion of giving systematically lower rates than the other raters. Can you confirm this suspicion running a suitable statistical procedure?