

Prof. Dr. Hendrik Godbersen

---

Workshop

Statistics & R

---

## 1 Installation of R, R Studio & Relevant Packages

### 2 Structure of R Studio

### 3 R Commands

### 4 Exercise Data

### 5 Descriptive Statistics

#### 5.1 Scale of Measurement

#### 5.2 Frequencies

#### 5.3 Measures of Central Tendency & Dispersion

## 6 Inferential Statistics

### 6.1 Principle & Overview

### 6.2 Chi-squared Test

### 6.3 t-test

### 6.4 ANOVA

### 6.5 Shapiro-Wilk Test

### 6.6 Wilcoxon Test

### 6.7 Correlation

### 6.8 Regression

### 6.9 Principal Component Analysis

# Installation of R, R Studio & Relevant Packages

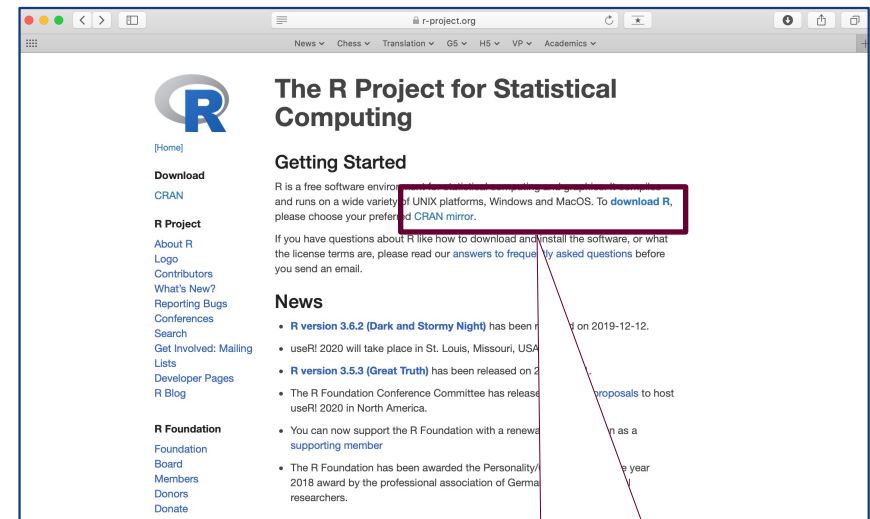
## 1) Downloading & Installing R

- 1) Go to: <https://www.r-project.org/>
- 2) Choose a for downloading R
  - Beware that R matches your system
- 3) Download R
- 4) Install R on your computer

## 2) Download & Installation von R Studio

- 1) Go to: <https://posit.co/download/rstudio-desktop/>
- 2) Download R Studio
- 3) Install R Studio on your computer

## 3) Installing additional R packages



Downloading R on  
[www.r-project.org](https://www.r-project.org)

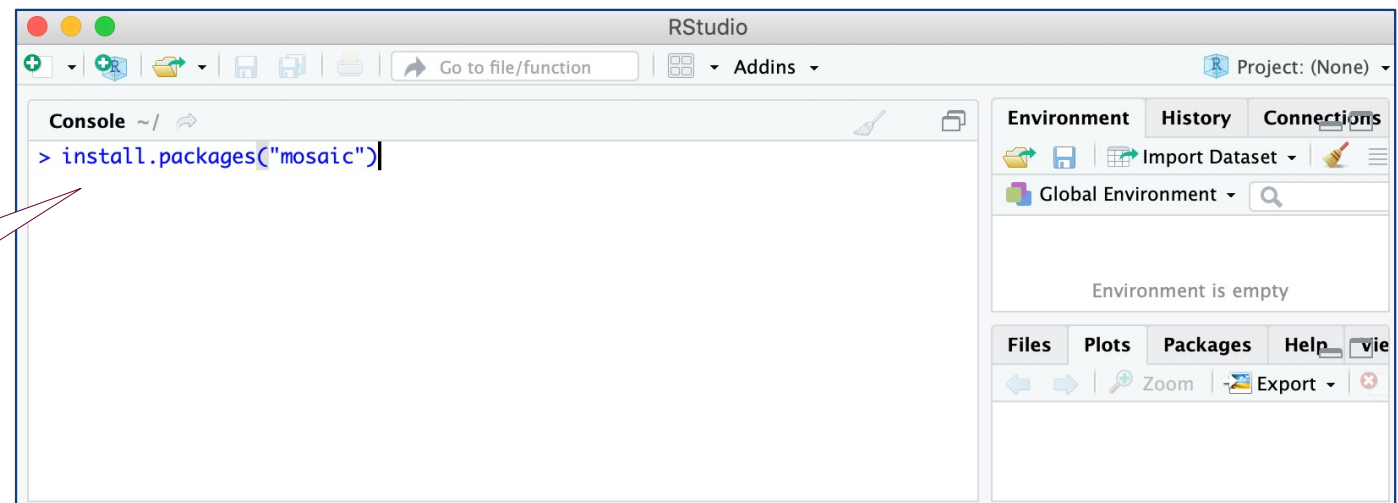
## Installation of R, R Studio & Relevant Packages

- 1) Downloading & Installing R ✓
- 2) Download & Installation von R Studio ✓
- 3) **Installing additional R packages**
  - 1) Make sure that your computer is connected to the internet
  - 2) Enter `install.packages("mosaic")` in Console
  - 3) Press the return button on your keyboard
  - 4) Repeat this process for all of the required packages

### Required R packages:

- mosaic
- psy
- psych

Enter `install.packages("mosaic")`  
& press return



1 Installation of R, R Studio & Relevant Packages

2 Structure of R Studio

3 R Commands

4 Exercise Data

5 Descriptive Statistics

5.1 Scale of Measurement

5.2 Frequencies

5.3 Measures of Central Tendency & Dispersion

6 Inferential Statistics

6.1 Principle & Overview

6.2 Chi-squared Test

6.3 t-test

6.4 ANOVA

6.5 Shapiro-Wilk Test

6.6 Wilcoxon Test

6.7 Correlation

6.8 Regression

6.9 Principal Component Analysis

## Structure of RStudio

The screenshot shows the RStudio environment with several callouts pointing to specific features:

- Script with R commands:** Points to the source editor where R code is written.
- Opening, saving etc. of files through the menu (the same as with other applications):** Points to the top menu bar.
- Button to run commands (lines in script must be marked):** Points to the 'Run' button in the toolbar.
- Loaded dataset & generated objects:** Points to the 'Environment' pane showing loaded objects like 'frame\_Gesc...' and 'radio'.
- Display of dataset (Beware: display ≠ loaded dataset):** Points to the 'Console' pane showing the output of commands.
- Generated graphs:** Points to the 'Plots' pane showing a bar chart of frequency by gender.
- Console with results (entering commands possible; however, commands will not be saved):** Points to the 'Console' pane.

**Script with R commands:**

```
radio$SenderCluster2[radio$SenderCluster=="Privat jung"] <- "Privat"
radio$SenderCluster2[radio$SenderCluster=="Privat alt"] <- "Privat"
radio$SenderCluster2[radio$SenderCluster=="SWR"] <- "SWR"
radio$Ge <- (radio$Comedy + radio$Gewinnspiele) / 2
radio_weiblich <- radio[radio$Geschlecht == "weiblich", ]
tally(~radio_weiblich)

# Häufigkeiten, Lageparameter & Streuungsmasse
# Häufigkeiten
tally(~Geschlecht, data = radio)
tally(~Geschlecht, format = "proportion", data = radio)
tally(~Geschlecht, format = "percent", data = radio)
tally(~SenderCluster, format = "percent", data = radio)
tally(SenderCluster~Geschlecht, format = "percent", data = radio)
```

**Environment pane:**

Object	Class	Size
frame_Gesc...	data.frame	2 obs. of 2 variables
radio	data.frame	275 obs. of 26 variables

**Console:**

```
> # Setup
> require(mosaic)
> require(psy)
> require(psych)
> require(openxlsx)
> load("~/Desktop/Rworkshop/radio.RData")
> View(radio)
> # Häufigkeiten, Lageparameter & Streuungsmasse
> # Häufigkeiten
> tally(~Geschlecht, data = radio)
Geschlecht
maennlich weiblich
129 146
> # Graphiken - Häufigkeiten
```

**Plots pane:**

Bar chart showing frequency (Freq) by gender (Geschlecht):

Geschlecht	Freq
maennlich	46.91
weiblich	53.09

1 Installation of R, R Studio & Relevant Packages

2 Structure of R Studio

3 R Commands

4 Exercise Data

5 Descriptive Statistics

5.1 Scale of Measurement

5.2 Frequencies

5.3 Measures of Central Tendency & Dispersion

6 Inferential Statistics

6.1 Principle & Overview

6.2 Chi-squared Test

6.3 t-test

6.4 ANOVA

6.5 Shapiro-Wilk Test

6.6 Wilcoxon Test

6.7 Correlation

6.8 Regression

6.9 Principal Component Analysis

- [illegible]



Installing & loading  
(additional) packages

Computing variables  
& datasets

Frequencies

Measures of central  
tendency & dispersion

Inferential statistics

Principal component  
analysis  
& Cronbach's alpha

## R (mosaic) – Key Commands

### Get going

```
install.packages("xyz")
require(xyz) ← starting mosaic, psych, psy, ggplot2, openxlsx, plspm (required for each session)
```

### Recoding, computing, formatting variables & creating new datasets:

Recoding (new) variables:  
dataset\$new\_variable[dataset\$old\_variable=="XXX"] <- "new\_value"  
→ Please note: categorical values require quotes (""), numerical values not

### Calculating new variables:

```
dataset$new_variable <- dataset$variable1 + dataset$variable2
```

### Changing the variable formats:

```
dataset$new_variable <- as.numeric(dataset$variable)
→ factor to numeric
(1): dataset$variable <- as.factor(dataset$variable)
(2): levels(dataset$variable) <- c("attribute1", "attribute1")
→ numeric to factor
```

### Creating new datasets (is equal, is not equal etc.):

```
new_dataset <- dataset[dataset$variable == "value",]
new_dataset <- dataset[dataset$variable != "value",]
new_dataset <- dataset[dataset$variable > value,]
```

### General command structure

```
command(y~x, z)
```

### Frequencies

```
tally(~variable1, data = dataset)
tally(~variable1, format = "percent", data = dataset)
tally(variable1~variable2, data = dataset)
tally(variable1~variable2, format = "percent", data = dataset)
```

### Median, mean, variance, standard deviation, maximum, minimum

```
median(~variable, data = dataset) / mean(~variable, data = dataset)
var(~variable, data = dataset) / sd(~variable, data = dataset)
max(~variable, data = dataset) / min(~variable, data = dataset)
favstats(~variable, data = dataset)
→ result: Min, Q1, median, Q3, Max, mean, standard deviation, n, missing values
```

Prof. Dr. Hendrik Godbersen: R (mosaic) – Key Commands

1

## Chi² test

```
xchisq.test(variable1~variable2, data = dataset)
```

## t-test

```
t.test(variable~grouping_variable, data = dataset)
```

## ANOVA

```
summary(aov(variable~grouping_variable, data = dataset))
→ alternative way:
modelxy <- aov(variable~grouping_variable, data = dataset)
summary(modelxy)
```

## Shapiro-Wilk test

```
shapiro.test(dataset$variable)
```

## Mann-Whitney & Wilcoxon test

```
wilcox.test(variable~grouping_variable, data = dataset)
(→ Mann-Whitney test)
wilcox.test(dataset$variable1, dataset$variable2, paired = T)
(→ Wilcoxon signed rank test (paired sample))
```

## Correlation

```
cor.test(variable1~variable2, data = dataset)
```

## Linear regression

```
summary(lm(dependent_variable~independent_variable, data = dataset))
```

## Multiple regression

```
summary(lm(dependent_variable~ind_variable1 + ind_variable2, data = dataset))
```

## Principal Component analysis (package: psych)

```
dataset_new <- cbind(dataset$variable1, dataset$variable2, dataset$variable3...)
colnames(dataset_new) <- c("item_1", "item_2")
KMO(dataset_new)
pcaX <- princomp(dataset_new, scores = TRUE, cor = TRUE)
summary(pcaX) plot(pcaX)
principal(dataset_new, nfactors = x, rotate = "varimax")
```

## Cronbach's Alpha (package: psy)

```
factorX <- cbind(dataset$variable1, dataset$variable2, dataset$variable3...)
cronbach(factorX)
```

Prof. Dr. Hendrik Godbersen: R (mosaic) – Key Commands

# R Commands: Command Structure

## R (mosaic) – Key Commands

### Get going

```
install.packages("xyz")
require(xyz) ← starting mosaic, psych, psy, ggplot2, openxlsx, plspm (required for each)
```

### Recoding, computing, formatting variables & creating new datasets:

#### Recoding (new) variables:

```
dataset$new_variable[dataset$old_variable=="XXX"] <- "new_value"
→ Please note: categorical values require quotes (""), numerical values
```

#### Calculating new variables:

```
dataset$new_variable <- dataset$variable1 + dataset$variable2
```

#### Changing the variable formats:

```
dataset$new_variable <- as.numeric(dataset$variable)
```

→ factor to numeric

```
(1): dataset$variable <- as.factor(dataset$variable)
```

```
(2): levels(dataset$variable) <- c("attribute1", "attribute1")
```

→ numeric to factor

#### Creating new datasets (is equal, is not equal etc.):

```
new_dataset <- dataset[dataset$variable == "value",]
```

```
new_dataset <- dataset[dataset$variable != "value",]
```

```
new_dataset <- dataset[dataset$variable > value,]
```

### General command structure

```
command(y~x, z)
```

### Frequencies

```
tally(~variable1, data = dataset)
```

```
tally(~variable1, format = "percent", data = dataset)
```

```
tally(variable1~variable2, data = dataset)
```

```
tally(variable1~variable2, format = "percent", data = dataset)
```

### Median, mean, variance, standard deviation, maximum, minimum

```
median(~variable, data = dataset) / mean(~variable, data = dataset)
```

```
var(~variable, data = dataset) / sd(~variable, data = dataset)
```

```
max(~variable, data = dataset) / min(~variable, data = dataset)
```

```
favstats(~variable, data = dataset)
```

→ result: Min, Q1, median, Q3, Max, mean, standard deviation, n, missing values)

Prof. Dr. Hendrik Godbersen: R (mosaic) – Key Commands

## General command structure

command(y~x, z)

## Frequencies

tally(~variable1, data = dataset)

tally(~variable1, format = "percent", data = dataset)

tally(variable1~variable2, data = dataset)

tally(variable1~variable2, format = "percent", data = dataset)

## Please note:

red R command (must not be changed)

blue names of variables or dataset, which must be individually adjusted

y dependent variable

x independent variable

z dataset

1 Installation of R, R Studio & Relevant Packages

2 Structure of R Studio

3 R Commands

4 Exercise Data

5 Descriptive Statistics

5.1 Scale of Measurement

5.2 Frequencies

5.3 Measures of Central Tendency & Dispersion

6 Inferential Statistics

6.1 Principle & Overview

6.2 Chi-squared Test

6.3 t-test

6.4 ANOVA

6.5 Shapiro-Wilk Test

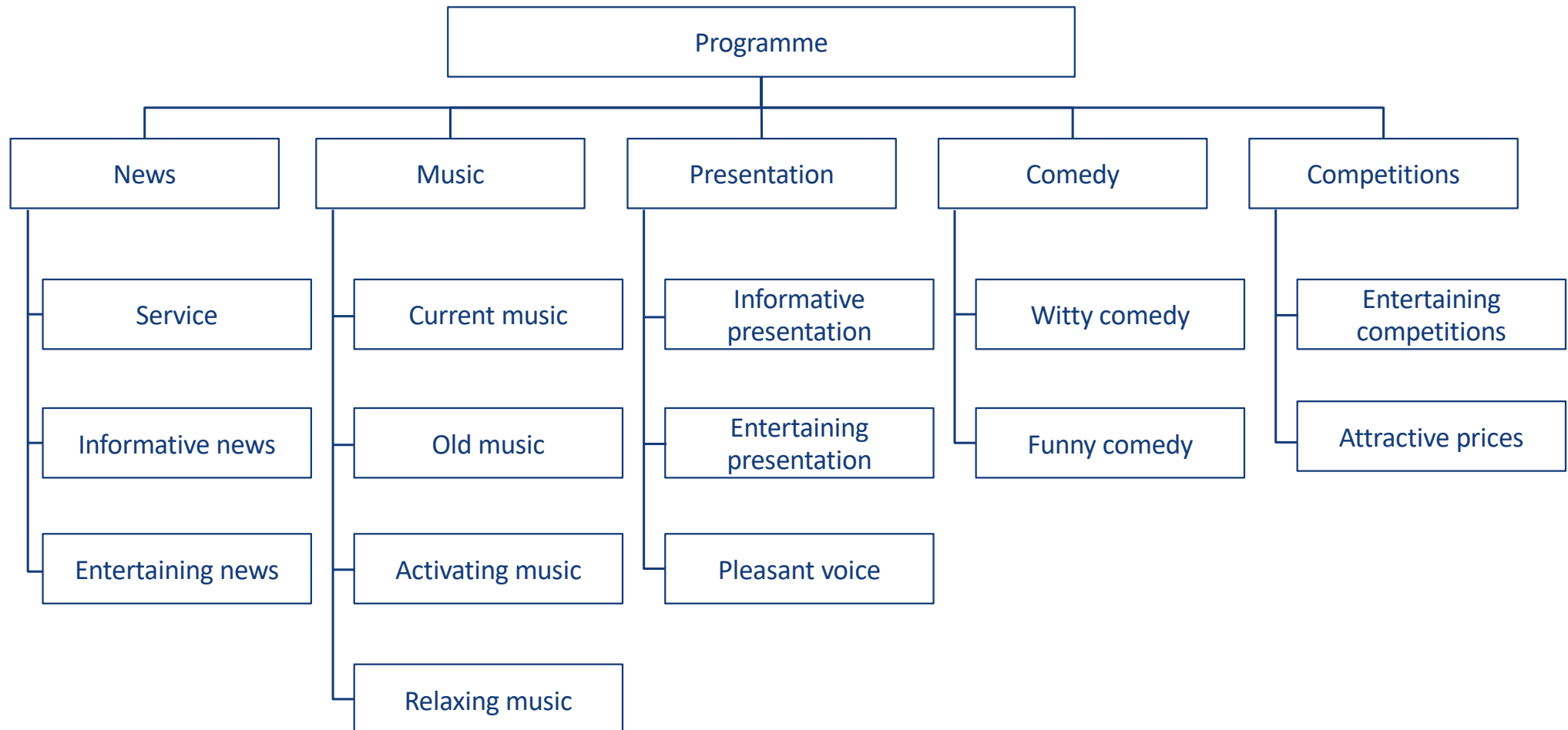
6.6 Wilcoxon Test

6.7 Correlation

6.8 Regression

6.9 Principal Component Analysis

## Exercise Data: Hypothesised Model



## Exercise Data: Questionnaire & Variables

- station
  - Measurement: main station in the last 2 weeks
- station\_cluster
  - Measurement: calculated variable with three attributes
  - Attributes: SWR / private young / private old
- Programme attributes
  - Measurement: Continuous Rating Scale – 0 (not good) to 100 (very good)
  - 14 items
    - service
    - informative\_news
    - entertaining\_news
    - current\_music
    - old\_music
    - activating\_music
    - relaxing\_music
    - informative\_presentation
    - entertaining\_presentation
    - pleasant\_voice
    - witty\_comedy
    - funny\_comedy
    - entertaining\_competitions
    - attractive\_prices
- Programme elements
  - Measurement: Continuous Rating Scale – 0 (not good) to 100 (very good)
  - 5 items
    - news
    - music
    - presentation
    - comedy
    - competitions
- programme
  - Measurement: Continuous Rating Scale – 0 (not good) to 100 (very good)
- gender
  - Measurement: single choice question
- age
  - Measurement: open question
- age\_groups (old vs. young, 2 groups)
- current\_music\_A (affinity, 2 groups)

Please note: Variable names in red

1 Installation of R, R Studio & Relevant Packages

2 Structure of R Studio

3 R Commands

4 Exercise Data

5 Descriptive Statistics

5.1 Scale of Measurement

5.2 Frequencies

5.3 Measures of Central Tendency & Dispersion

6 Inferential Statistics

6.1 Principle & Overview

6.2 Chi-squared Test

6.3 t-test

6.4 ANOVA

6.5 Shapiro-Wilk Test

6.6 Wilcoxon Test

6.7 Correlation

6.8 Regression

6.9 Principal Component Analysis

## Scale of Measurement

Scale of measurement		Mathematical characteristics	Characteristics of values	Example	Measure of central tendency
Categorical (dichotomous)	Nominal	$\neq$	Values are equal or not.	Gender	<ul style="list-style-type: none"> <li>• Mode</li> </ul>
	Ordinal	$\neq$ ; $</>$	Values are larger, smaller or equal.	Olympic ranks	<ul style="list-style-type: none"> <li>• Mode</li> <li>• Median</li> </ul>
Metric (continuous)	Interval	$\neq$ ; $</>$ ; - ; +	The distance between values can be determined.	Temperature	<ul style="list-style-type: none"> <li>• Mode</li> <li>• Median</li> <li>• Arithmetic mean</li> </ul>
	Ratio	$\neq$ ; $</>$ ; + / - ; * / ÷	The distance and ratio between values can be determined.	Height	

1 Installation of R, R Studio & Relevant Packages

2 Structure of R Studio

3 R Commands

4 Exercise Data

5 Descriptive Statistics

5.1 Scale of Measurement

5.2 Frequencies

5.3 Measures of Central Tendency & Dispersion

6 Inferential Statistics

6.1 Principle & Overview

6.2 Chi-squared Test

6.3 t-test

6.4 ANOVA

6.5 Shapiro-Wilk Test

6.6 Wilcoxon Test

6.7 Correlation

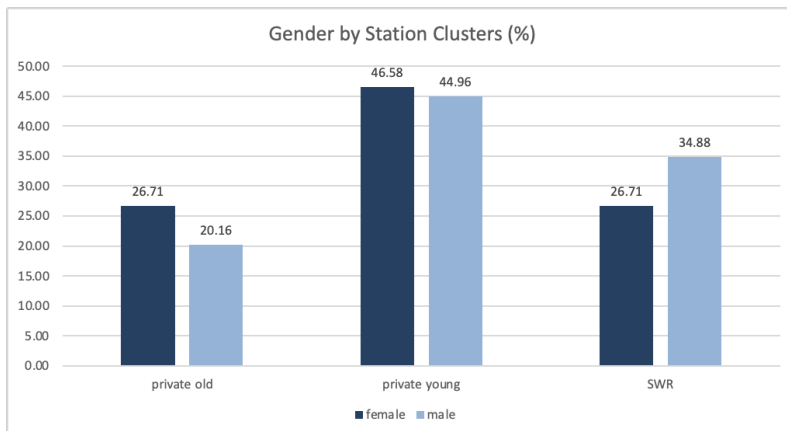
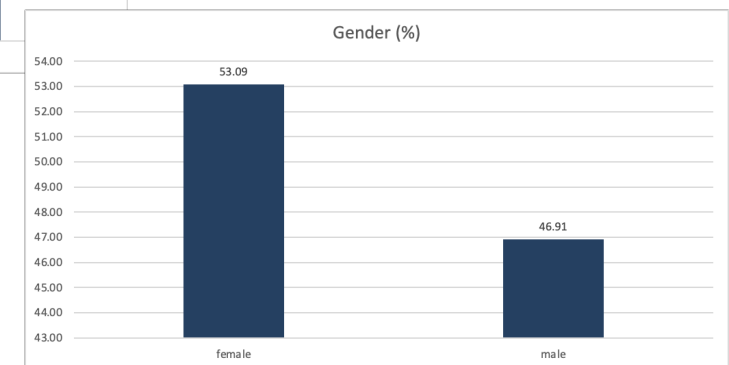
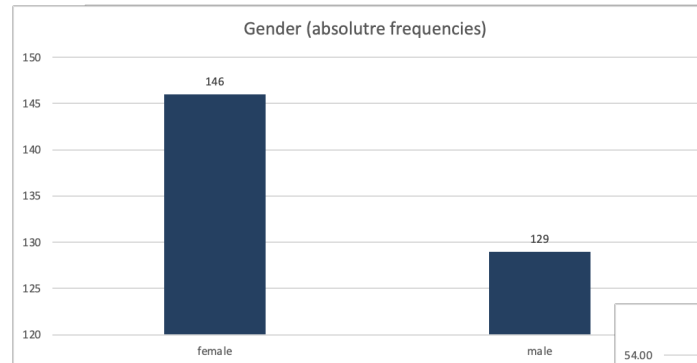
6.8 Regression

6.9 Principal Component Analysis



## Frequencies: Brief

- Absolute frequency:
  - Number of times a value of a variable occurred
- Relative frequency:
  - Ratio of an absolute frequency of a value to the total number of values for a variable



- Cross tabulation (contingency table):
  - Combination of the distribution of two variables

## Frequencies: Exercise & R Commands (1)

### Exercise:

- Of how many men and women does the sample consist?
- What is the proportion of men and women (%)?
- What is the proportion of station clusters (%)?

### Frequencies

`tally(~variable1, data = dataset)`

`tally(~variable1, format = "percent", data = dataset)`

`tally(variable1~variable2, data = dataset)`

`tally(variable1~variable2, format = "percent", data = dataset)`

→ numeric to factor

Creating new datasets (is.equal is not equal etc.):

```
new_dataset <- dataset[dataset$variable == "value",]
new_dataset <- dataset[dataset$variable != "value",]
new_dataset <- dataset[dataset$variable > "value",]
```

General command structure

```
command(y~x, z)
```

**Frequencies**

```
tally(~variable1, data = dataset)
tally(~variable1, format = "percent", data = dataset)
tally(variable1~variable2, data = dataset)
tally(variable1~variable2, format = "percent", data = dataset)
```

**Median, mean, variance, standard deviation, maximum, minimum**

```
median(~variable, data = dataset) / mean(~variable, data = dataset)
var(~variable, data = dataset) / sd(~variable, data = dataset)
max(~variable, data = dataset) / min(~variable, data = dataset)
favstats(~variable, data = dataset)
→ result: Min, Q1, median, Q3, Max, mean, standard deviation, n, missing values
```

**Wilcoxon test**

```
wilcox.test(dataset$variable1, dataset$variable2, paired = T)
(→ Wilcoxon signed rank test (paired sample))
```

**Correlation**

```
cor.test(variable1~variable2, data = dataset)
```

**Linear regression**

```
summary(lm(dependent_variable~independent_variable, data = dataset))
```

**Multiple regression**

```
summary(lm(dependent_variable~ind_variable1 + ind_variable2, data = dataset))
```

**Principal Component analysis** (package: psych)

```
dataset_new <- cbind(dataset$variable1, dataset$variable2, dataset$variable3...)
colnames(dataset_new) <- c("item_1", "item_2")
KMO(dataset_new)
pcaX <- princomp(dataset_new, scores = TRUE, cor = TRUE)
summary(pcaX) plot(pcaX)
principal(dataset_new, nfactors = 2, rotate = "varimax")
```

**Cronbach's Alpha** (package: psy)

```
factorX <- cbind(dataset$variable1, dataset$variable2, dataset$variable3...)
cronbach(factorX)
```

### Variables

- station
- **station\_cluster**
- service
- informative\_news
- entertaining\_news
- current\_music
- old\_music
- activating\_music
- relaxing\_music
- informative\_presentation
- entertaining\_presentation
- pleasant\_voice
- witty\_comedy
- funny\_comedy
- entertaining\_competitions
- attractive\_prices
- news
- music
- presentation
- comedy
- competitions
- programme
- **gender**
- age
- age\_groups
- current\_music\_A

## Frequencies: Exercise & R Commands (2)

### Exercise:

- How are men and women distributed over the station clusters (%)?

### Variables

- station
- station\_cluster**
- service
- informative\_news
- entertaining\_news
- current\_music
- old\_music
- activating\_music
- relaxing\_music
- informative\_presentation
- entertaining\_presentation
- pleasant\_voice
- witty\_comedy
- funny\_comedy
- entertaining\_competitions
- attractive\_prices
- news
- music
- presentation
- comedy
- competitions
- programme
- gender**
- age
- age\_groups
- current\_music\_A

### Frequencies

`tally(~variable1, data = dataset)`

`tally(~variable1, format = "percent", data = dataset)`

`tally(variable1~variable2, data = dataset)`

`tally(variable1~variable2, format = "percent", data = dataset)`

→ numeric to factor

Creating new datasets (is equal, is not equal etc.):

```
new_dataset <- dataset[dataset$variable == "value"]
new_dataset <- dataset[dataset$variable != "value"]
new_dataset <- dataset[dataset$variable > value]
```

General command structure

```
command(y~x, z)
```

**Frequencies**

```
tally(~variable1, data = dataset)
tally(~variable1, format = "percent", data = dataset)
tally(variable1~variable2, data = dataset)
tally(variable1~variable2, format = "percent", data = dataset)
```

**Median, mean, variance, standard deviation, maximum, minimum**

```
median(~variable, data = dataset) / mean(~variable, data = dataset)
var(~variable, data = dataset) / sd(~variable, data = dataset)
max(~variable, data = dataset) / min(~variable, data = dataset)
favstats(~variable, data = dataset)
→ result: Min, Q1, median, Q3, Max, mean, standard deviation, n, missing values
```

**Wilcoxon test**

```
wilcox.test(dataset$variable1, dataset$variable2, paired = T)
(→ Wilcoxon signed rank test (paired sample))
```

**Correlation**

```
cor.test(variable1~variable2, data = dataset)
```

**Linear regression**

```
summary(lm(dependent_variable~independent_variable, data = dataset))
```

**Multiple regression**

```
summary(lm(dependent_variable~ind_variable1 + ind_variable2, data = dataset))
```

**Principal Component analysis** (package: psych)

```
dataset_new <- cbind(dataset$variable1, dataset$variable2, dataset$variable3...)
colnames(dataset_new) <- c("item_1", "item_2")
KMO(dataset_new)
pcaX <- princomp(dataset_new, scores = TRUE, cor = TRUE)
summary(pcaX) plot(pcaX)
principal(dataset_new, nfactors = 2, rotate = "varimax")
```

**Cronbach's Alpha** (package: psy)

```
factorX <- cbind(dataset$variable1, dataset$variable2, dataset$variable3...)
cronbach(factorX)
```

1 Installation of R, R Studio & Relevant Packages

2 Structure of R Studio

3 R Commands

4 Exercise Data

5 Descriptive Statistics

5.1 Scale of Measurement

5.2 Frequencies

5.3 Measures of Central Tendency & Dispersion

6 Inferential Statistics

6.1 Principle & Overview

6.2 Chi-squared Test

6.3 t-test

6.4 ANOVA

6.5 Shapiro-Wilk Test

6.6 Wilcoxon Test

6.7 Correlation

6.8 Regression

6.9 Principal Component Analysis

- **Measures of central tendency:**

- Arithmetic mean:
  - Sum of all values divided by the number of all values; metric measurement level required
- Median:
  - Value that separates the higher and lower half of a distribution; ordinal measurement level required
- Mode:
  - Most frequent value of frequency distribution; nominal measurement level required

- **Measures of dispersion:**

- Variance
  - Average squared difference of values from their arithmetic mean (please note: empirical variance  $\rightarrow /n$ ; sample variance  $\rightarrow /n-1$ )
- Standard deviation
  - Average difference of values from their arithmetic mean

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

$$\sigma = \sqrt{\sigma^2}$$

## Measures of Central Tendency & Dispersion: Brief

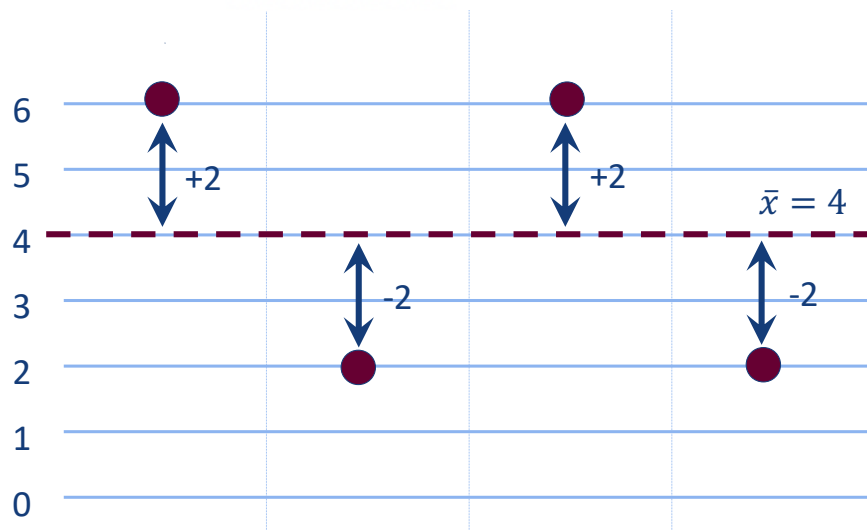
- Measures of dispersion**

- Variance

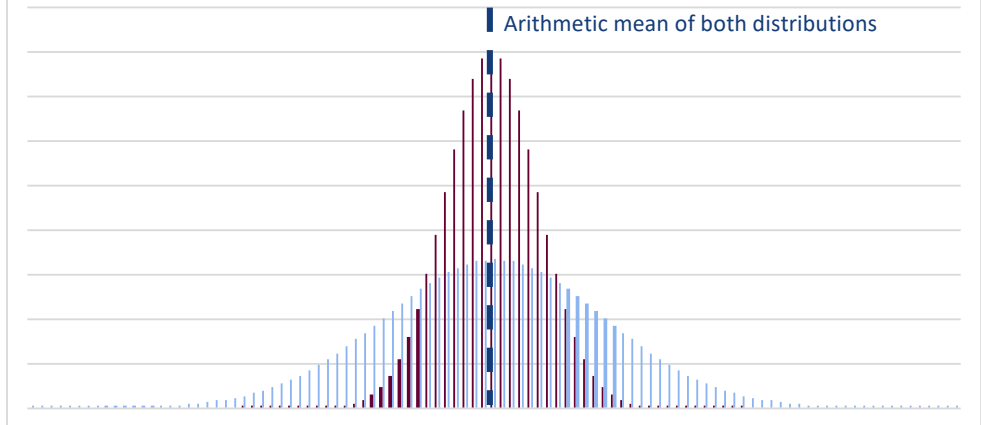
$$\sigma^2 = \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n}$$

- Standard deviation

$$\sigma = \sqrt{\sigma^2}$$



Two Normal Distributions with Different Variances



$\sum_{i=1}^n (x_i - \bar{x})$	{	$(+2) + (-2) + (+2) + (-2) = 0$	}
$\sum_{i=1}^n (x_i - \bar{x})^2$	{	$(+2)^2 + (-2)^2 + (+2)^2 + (-2)^2 =$	}
		$4 + 4 + 4 + 4 = 16$	
$\sigma^2 = \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n}$	{	$16 / 4 = 4$	}
$\sigma = \sqrt{\sigma^2}$	{	$\sqrt{4} = 2$	}

# Measures of Central Tendency & Dispersion: Exercise & R Commands (1)

## Exercise:

- What is the average age of the participants (arithmetic mean)?
- What is the median age of the participants?
- What is the variance and standard deviation of the age?

## Median, mean, variance, standard deviation, maximum, minimum

`median(~variable, data = dataset) / mean(~variable, data = dataset)`

`var(~variable, data = dataset) / sd(~variable, data = dataset)`

`max(~variable, data = dataset) / min(~variable, data = dataset)`

`favstats(~variable, data = dataset)`

```
new_dataset <- dataset[dataset$variable > value.]

General command structure
command1~x, z)

Frequencies
tally(~variable1, data = dataset)
tally(~variable1, format = "percent", data = dataset)
tally(variable1~variable2, data = dataset)
tally(variable1~variable2, format = "percent", data = dataset)

Median, mean, variance, standard deviation, maximum, minimum
median(~variable, data = dataset) / mean(~variable, data = dataset)
var(~variable, data = dataset) / sd(~variable, data = dataset)
max(~variable, data = dataset) / min(~variable, data = dataset)
favstats(~variable, data = dataset)
→ result: Min, Q1, median, Q3, Max, mean, standard deviation, n, missing values
```

Prof. Dr. Hendrik Godbersen: R (mosaic) - Key Commands

1

```
Correlation
cor.test(variable1~variable2, data = dataset)

Linear regression
summary(lm(dependent_variable~independent_variable, data = dataset))

Multiple regression
summary(lm(dependent_variable~ind_variable1 + ind_variable2, data = dataset))

Principal Component analysis (package: psych)
dataset_new <- cbind(dataset$variable1, dataset$variable2, dataset$variable3...)
colnames(dataset_new) <- c("item_1", "item_2")
KMO(dataset_new)
pcaX <- princomp(dataset_new, scores = TRUE, cor = TRUE)
summary(pcaX) plot(pcaX)
principal(dataset_new, nfactors = 2, rotate = "varimax")

Cronbach's Alpha (package: psy)
factorX <- cbind(dataset$variable1, dataset$variable2, dataset$variable3...)
cronbach(factorX)
```

Prof. Dr. Hendrik Godbersen: R (mosaic) - Key Commands

2

## Variables

- station
- station\_cluster
- service
- informative\_news
- entertaining\_news
- current\_music
- old\_music
- activating\_music
- relaxing\_music
- informative\_presentation
- entertaining\_presentation
- pleasant\_voice
- witty\_comedy
- funny\_comedy
- entertaining\_competitions
- attractive\_prices
- news
- music
- presentation
- comedy
- competitions
- programme
- gender
- **age**
- age\_groups
- current\_music\_A

## Measures of Central Tendency & Dispersion: Exercise & R Commands (2)

### Exercise:

- What is the average programme evaluation of the participants (arithmetic mean)?
- What is the median programme evaluation of the participants?
- What is the variance and standard deviation of the programme evaluation?

### Median, mean, variance, standard deviation, maximum, minimum

`median(~variable, data = dataset) / mean(~variable, data = dataset)`

`var(~variable, data = dataset) / sd(~variable, data = dataset)`

`max(~variable, data = dataset) / min(~variable, data = dataset)`

`favstats(~variable, data = dataset)`

```
new_dataset <- dataset[dataset$variable > value.]

General command structure
command1~x, z)

Frequencies
tally(~variable1, data = dataset)
tally(~variable1, format = "percent", data = dataset)
tally(variable1~variable2, data = dataset)
tally(variable1~variable2, format = "percent", data = dataset)

Median, mean, variance, standard deviation, maximum, minimum
median(~variable, data = dataset) / mean(~variable, data = dataset)
var(~variable, data = dataset) / sd(~variable, data = dataset)
max(~variable, data = dataset) / min(~variable, data = dataset)
favstats(~variable, data = dataset)
  -> result: Min, Q1, median, Q3, Max, mean, standard deviation, n, missing values
```

Prof. Dr. Hendrik Godbersen: R (mosaic) - Key Commands

1

```
Correlation
cor.test(variable1~variable2, data = dataset)

Linear regression
summary(lm(dependent_variable~independent_variable, data = dataset))

Multiple regression
summary(lm(dependent_variable~ind_variable1 + ind_variable2, data = dataset))

Principal Component analysis (package: psych)
dataset_new <- cbind(dataset$variable1, dataset$variable2, dataset$variable3...)
colnames(dataset_new) <- c("item_1", "item_2")
KMO(dataset_new)
pcaX <- princomp(dataset_new, scores = TRUE, cor = TRUE)
summary(pcaX) plot(pcaX)
principal(dataset_new, nfactors = 2, rotate = "varimax")

Cronbach's Alpha (package: psy)
factorX <- cbind(dataset$variable1, dataset$variable2, dataset$variable3...)
cronbach(factorX)
```

Prof. Dr. Hendrik Godbersen: R (mosaic) - Key Commands

2

### Variables

- station
- station\_cluster
- service
- informative\_news
- entertaining\_news
- current\_music
- old\_music
- activating\_music
- relaxing\_music
- informative\_presentation
- entertaining\_presentation
- pleasant\_voice
- witty\_comedy
- funny\_comedy
- entertaining\_competitions
- attractive\_prices
- news
- music
- presentation
- comedy
- competitions
- **programme**
- gender
- age
- age\_groups
- current\_music\_A



## Measures of Central Tendency & Dispersion: Exercise & R Commands (3)

### Exercise:

- Determine arithmetic mean, median and standard deviation for age with only one R command.
- Determine arithmetic mean, median and standard deviation for programme with only one R command.

### Median, mean, variance, standard deviation, maximum, minimum

`median(~variable, data = dataset) / mean(~variable, data = dataset)`

`var(~variable, data = dataset) / sd(~variable, data = dataset)`

`max(~variable, data = dataset) / min(~variable, data = dataset)`

`favstats(~variable, data = dataset)`

`new_dataset <- dataset[dataset$variable > value,]`

#### General command structure

`command1~x, z)`

#### Frequencies

`tally(~variable1, data = dataset)`

`tally(~variable1, format = "percent", data = dataset)`

`ply(variable1~variable2, data = dataset)`

`tally(variable1~variable2, format = "percent", data = dataset)`

#### Median, mean, variance, standard deviation, maximum, minimum

`median(~variable, data = dataset) / mean(~variable, data = dataset)`

`var(~variable, data = dataset) / sd(~variable, data = dataset)`

`max(~variable, data = dataset) / min(~variable, data = dataset)`

`favstats(~variable, data = dataset)`

→ result: Min, Q1, median, Q3, Max, mean, standard deviation, n, missing values

Prof. Dr. Hendrik Godbersen: R (mosaic) - Key Commands

1

#### Correlation

`cor.test(variable1~variable2, data = dataset)`

#### Linear regression

`summary(lm(dependent_variable~independent_variable, data = dataset))`

#### Multiple regression

`summary(lm(dependent_variable~ind_variable1 + ind_variable2, data = dataset))`

#### Principal Component analysis (package: psych)

`dataset_new <- cbind(dataset$variable1, dataset$variable2, dataset$variable3...)`

`colnames(dataset_new) <- c("item_1", "item_2")`

`KMO(dataset_new)`

`pcaX <- princomp(dataset_new, scores = TRUE, cor = TRUE)`

`summary(pcaX)`

`plot(pcaX)`

`principal(dataset_new, nfactors = x, rotate = "varimax")`

#### Cronbach's Alpha (package: psy)

`factorX <- cbind(dataset$variable1, dataset$variable2, dataset$variable3...)`

`cronbach(factorX)`

Prof. Dr. Hendrik Godbersen: R (mosaic) - Key Commands

2

### Variables

- station
- station\_cluster

- service
- informative\_news
- entertaining\_news
- current\_music
- old\_music
- activating\_music
- relaxing\_music
- informative\_presentation
- entertaining\_presentation
- pleasant\_voice
- witty\_comedy
- funny\_comedy
- entertaining\_competitions
- attractive\_prices

- news
- music
- presentation
- comedy
- competitions

### programme

- gender
- age

- age\_groups
- current\_music\_A

## Measures of Central Tendency & Dispersion: Exercise & R Commands (4)

### Exercise:

- Compare the average age of men and women (arithmetic means).
- Compare the average programme evaluation of men and women (arithmetic means).

### Median, mean, variance, standard deviation, maximum, minimum

`median(~variable, data = dataset) / mean(~variable, data = dataset)`

`var(~variable, data = dataset) / sd(~variable, data = dataset)`

`max(~variable, data = dataset) / min(~variable, data = dataset)`

`favstats(~variable, data = dataset)`

`new_dataset <- dataset[dataset$variable > value,]`

#### General command structure

`command1~x, z)`

#### Frequencies

`tally(~variable1, data = dataset)`

`tally(~variable1, format = "percent", data = dataset)`

`ply(variable1~variable2, data = dataset)`

`tally(variable1~variable2, format = "percent", data = dataset)`

#### Median, mean, variance, standard deviation, maximum, minimum

`median(~variable, data = dataset) / mean(~variable, data = dataset)`

`var(~variable, data = dataset) / sd(~variable, data = dataset)`

`max(~variable, data = dataset) / min(~variable, data = dataset)`

`favstats(~variable, data = dataset)`

→ result: Min, Q1, median, Q3, Max, mean, standard deviation, n, missing values

Prof. Dr. Hendrik Godbersen: R (mosaic) - Key Commands

1

#### Correlation

`cor.test(variable1~variable2, data = dataset)`

#### Linear regression

`summary(lm(dependent_variable~independent_variable, data = dataset))`

#### Multiple regression

`summary(lm(dependent_variable~ind_variable1 + ind_variable2, data = dataset))`

#### Principal Component analysis (package: psych)

`dataset_new <- cbind(dataset$variable1, dataset$variable2, dataset$variable3...)`

`colnames(dataset_new) <- c("item_1", "item_2")`

`KMO(dataset_new)`

`pcaX <- princomp(dataset_new, scores = TRUE, cor = TRUE)`

`summary(pcaX)`

`plot(pcaX)`

`principal(dataset_new, nfactors = 3, rotate = "varimax")`

#### Cronbach's Alpha (package: psy)

`factorX <- cbind(dataset$variable1, dataset$variable2, dataset$variable3...)`

`cronbach(factorX)`

Prof. Dr. Hendrik Godbersen: R (mosaic) - Key Commands

2

### Variables

- station
- station\_cluster
- service
- informative\_news
- entertaining\_news
- current\_music
- old\_music
- activating\_music
- relaxing\_music
- informative\_presentation
- entertaining\_presentation
- pleasant\_voice
- witty\_comedy
- funny\_comedy
- entertaining\_competitions
- attractive\_prices

- news
- music
- presentation
- comedy
- competitions
- **programme**
- **gender**
- **age**
- age\_groups
- current\_music\_A

1 Installation of R, R Studio & Relevant Packages

2 Structure of R Studio

3 R Commands

4 Exercise Data

5 Descriptive Statistics

5.1 Scale of Measurement

5.2 Frequencies

5.3 Measures of Central Tendency & Dispersion

6 Inferential Statistics

6.1 Principle & Overview

6.2 Chi-squared Test

6.3 t-test

6.4 ANOVA

6.5 Shapiro-Wilk Test

6.6 Wilcoxon Test

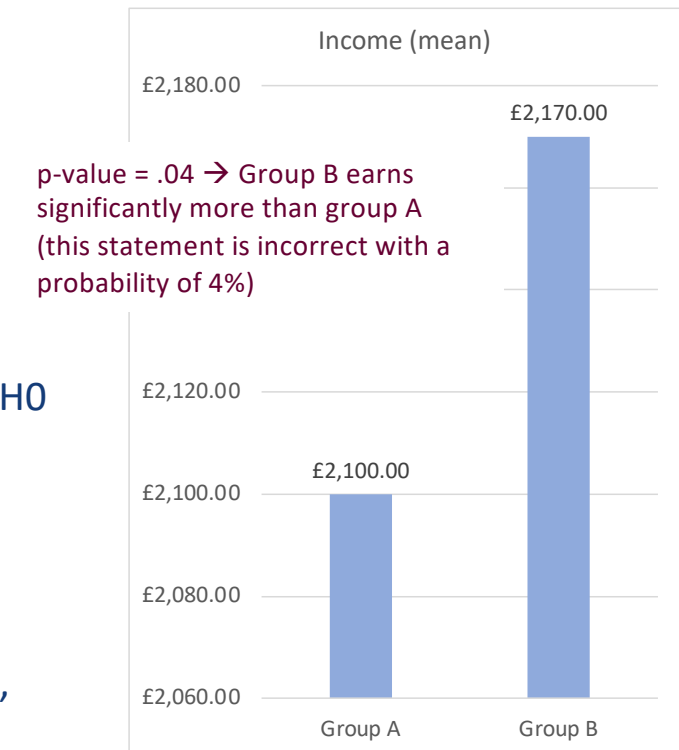
6.7 Correlation

6.8 Regression

6.9 Principal Component Analysis

## Inferential Statistics: Principle

- Inferential statistics determines if effects are significant
  - Significant = not coincidental (common language: an effect “really” exists)
  - Non-significant = coincidental (common language: there is no effect)
- The quantitative “black and white world:
  - H1: Effect XY exists.
  - H0: Effect XY does not exist.
- „Logic“ of inferential statistical procedures
  - The inferential statistical procedures try to reject H0.
  - The results is the p-value with values between 0 and 1.
  - The p-value indicates the probability that H0 holds true if though H0 was rejected.
- Steps of the analysis
  - (1) p-value → Significance:  $p < .05$  = significant (higher degrees of significance at  $p < .01$  and  $p < .001$ )
  - (2) If p-value  $< .05$ : interpreting the measured values (arith. mean, frequencies etc.) → direction & strength of the effect



# Inferential Statistics: Overview

Prof. Dr.  
Godbersen

Scale of measurement		Mathematical characteristics	Characteristics of values	Example	Measure of central tendency
Categorical (dichotomous)	Nominal	$\neq$	Values are equal or not.	Gender	• Mode
	Ordinal	$\neq ; </>$	Values are larger, smaller or equal.	Olympic ranks	• Mode • Median
Metric (continuous)	Interval	$\neq ; </> ; - ; +$	The distance between values can be determined.	Temperature	• Mode • Median • Arithmetic mean
	Ratio	$\neq ; </> ; +/- ; * / \div$	The distance and ratio between values can be determined.	Height	
Exploratory Data Analysis		<ul style="list-style-type: none"> <li>• Principal component analysis</li> <li>• Cluster analysis</li> </ul>			

Inferential Statistical Procedures		Independent variable	
		categorical (dichotomous)	Metric (continuous)
Dependent variable	categorical	Chi <sup>2</sup> test	Discriminant analysis
	metric	t-test*** (2 groups) & Analysis of variances (≥ 3 groups)	Regression (dependence) & correlation (interdependence)

\*\*\* „Additional“ tests:

- Shapiro-Wilk test  
(tests if a variable is normally distributed; precondition for t-tests at small sample sizes,  $n \leq 50$ )
- Mann-Whitney test/Wilcoxon test  
(tests if 2 groups differ on an at least ordinally scaled variable, without requiring a normal distribution)

## Exploratory Data Analysis

- Principal component analysis
- Cluster analysis

1 Installation of R, R Studio & Relevant Packages

2 Structure of R Studio

3 R Commands

4 Exercise Data

5 Descriptive Statistics

5.1 Scale of Measurement

5.2 Frequencies

5.3 Measures of Central Tendency & Dispersion

6 Inferential Statistics

6.1 Principle & Overview

6.2 Chi-squared Test

6.3 t-test

6.4 ANOVA

6.5 Shapiro-Wilk Test

6.6 Wilcoxon Test

6.7 Correlation

6.8 Regression

6.9 Principal Component Analysis

- **Application:**
  - Testing the (in)dependence of two categorical (nominal) variables
- **Leading question:**
  - Is there an association of two variables with a nominal measurement level? / Are two categorical variables independent from each other?
- **Steps of the analysis:**
  - 1) Is the p-value  $< .05$  ( $< .01$ ,  $< .001$ )?  $\rightarrow$  Significance?
  - 2) How do the frequencies compare?
- **Statistical Analysis („in the background“):**
  - 1) Calculation of the expected frequencies (e) which would indicate no difference between the variables (column sum \* row sum / number of observations)
  - 2) Test if the observed values (h) significantly deviate from the expected values  $\epsilon$

	Female	Male	Sum
Blue	h = 35 e = 25	h = 15 e = 25	50
Red	h = 15 e = 25	h = 35 e = 25	50
Sum	50	50	100



## Chi-squared Test: Brief

- **Statistical Analysis („in the background“):**
  - 1) Calculation of the expected frequencies (e) which would indicate no difference between the variables (column sum \* row sum / number of observations)
  - 2) Test if the observed values (h) significantly deviate from the expected values €

V1	Female	Male	Sum
Blue	h = 15 e = 25	h = 35 e = 25	50
Red	h = 35 e = 25	h = 15 e = 25	50
Sum	50	50	100

p-value = 0.0001447

V2	Female	Male	Sum
Bleu	h = 24 e = 25	h = 26 e = 25	50
Red	h = 26 e = 25	h = 24 e = 25	50
Sum	50	50	100

p-value = 0.8415

# Chi-squared Test: Exercise & R Commands

## Exercise:

- Do men and women differ in their preferences for station clusters?
- Do the old and young age segment differ in their affinity to current music? If so, how does this difference look like?

**R (mosaic) – Key Commands**

Get going  
install.packages("xyz")  
require(xyz) <- starting mosaic, psych, ggplot2, openssl, plspm (required for each session)

Recoding, computing, formatting variables & creating new datasets:

**Chi<sup>2</sup> test**  
**xchisq.test(variable1~variable2, data = dataset)**

(2): levels(dataset\$variable) <- c("attribute1", "attribute1")  
→ numeric to factor

Creating new datasets (is equal, is not equal etc.):  
new\_dataset <- dataset[dataset\$variable == "value",]  
new\_dataset <- dataset[dataset\$variable != "value",]  
new\_dataset <- dataset[dataset\$variable > value,]

General command structure  
command1~x, z)

Frequencies  
tally(~variable1, data = dataset)  
tally(~variable1, format = "percent", data = dataset)  
tally(variable1~variable2, data = dataset)  
tally(variable1~variable2, format = "percent", data = dataset)

Median, mean, variance, standard deviation, maximum, minimum  
median(~variable, data = dataset) / mean(~variable, data = dataset)  
var(~variable, data = dataset) / sd(~variable, data = dataset)  
max(~variable, data = dataset) / min(~variable, data = dataset)  
favstats(~variable, data = dataset)  
→ result: Min, Q1, median, Q3, Max, mean, standard deviation, n, missing values

**Chi<sup>2</sup> test**  
xchisq.test(variable1~variable2, data = dataset)

**t-test**  
t.test(variable~grouping\_variable, data = dataset)

**ANOVA**  
summary(aov(variable~grouping\_variable, data = dataset))

**wilcox.test(variable~grouping\_variable, data = dataset)**  
(→ Mann-Whitney test)  
wilcox.test(dataset\$variable1, dataset\$variable2, paired = T)  
(→ Wilcoxon signed rank test (paired sample))

**Correlation**  
cor.test(variable1~variable2, data = dataset)

**Linear regression**  
summary(lm(dependent\_variable~independent\_variable, data = dataset))

**Multiple regression**  
summary(lm(dependent\_variable~ind\_variable1 + ind\_variable2, data = dataset))

**Principal Component analysis** (package: psych)  
dataset\_new <- cbind(dataset\$variable1, dataset\$variable2, dataset\$variable3...)  
colnames(dataset\_new) <- c("item\_1", "item\_2")  
KMO(dataset\_new)  
pcaX <- princomp(dataset\_new, scores = TRUE, cor = TRUE)  
summary(pcaX) plot(pcaX)  
principal(dataset\_new, nfactors = 2, rotate = "varimax")

**Cronbach's Alpha** (package: psy)  
factorX <- cbind(dataset\$variable1, dataset\$variable2, dataset\$variable3...)  
cronbach(factorX)

## Variables

- station
- station\_cluster**
- service
- informative\_news
- entertaining\_news
- current\_music
- old\_music
- activating\_music
- relaxing\_music
- informative\_presentation
- entertaining\_presentation
- pleasant\_voice
- witty\_comedy
- funny\_comedy
- entertaining\_competitions
- attractive\_prices
- news
- music
- presentation
- comedy
- competitions
- programme
- gender**
- age
- age\_groups**
- current\_music\_A**

1 Installation of R, R Studio & Relevant Packages	
2 Structure of R Studio	
3 R Commands	
4 Exercise Data	
5 Descriptive Statistics	
5.1 Scale of Measurement	
5.2 Frequencies	
5.3 Measures of Central Tendency & Dispersion	
6 Inferential Statistics	
6.1 Principle & Overview	
6.2 Chi-squared Test	
6.3 t-test	
6.4 ANOVA	
6.5 Shapiro-Wilk Test	
6.6 Wilcoxon Test	
6.7 Correlation	
6.8 Regression	
6.9 Principal Component Analysis	

## t-Test: Brief

---

- **Application:**
  - Testing if a metric variable differs between two groups (independent sample t-test)
- **Leading question:**
  - Do two groups (samples) differ in a metric variable?
  - Does a categorical variable (two groups) effect a metric variable?
- **Types of t-tests:**
  - Independent sample t-test – measurements from two groups
  - Paired sample t-test – two measurement in one group (e.g., before and after)
  - One-sample t-test – comparison of a group's arithmetic mean with a fixed value
- **Steps of the analysis:**
  - 1) Is the p-value  $< .05$  ( $< .01$ ,  $< .001$ )?  $\rightarrow$  Significance?
  - 2) How do the arithmetic means compare?

# t-Test: Exercise & R Commands

## Exercise:

- Do men and women evaluate the programme of radio stations differently?
- Do men and women evaluate the music of radio stations differently?

### R (mosaic) – Key Commands

**Get going**  
install.packages("xyz")  
require(xyz) ← starting mosaic, psych, ggplot2, openssl, plspm (required for each session)

**Recoding, computing, formatting variables & creating new datasets:**  
**Recoding (new) variables:**  
dataset\$new\_variable[dataset\$old\_variable=="XXX"] <- "new\_value"  
→ Please note: categorical values require quotes (""), numerical values not

**Calculating new variables:**  
dataset\$new\_variable <- dataset\$variable1 + dataset\$variable2

**Changing the the variable format:**  
dataset\$new\_variable <- as.numeric(dataset\$variable)

**Chi<sup>2</sup> test**  
xchisq.test(variable1~variable2, data = dataset)

**t-test**  
t.test(variable~grouping\_variable, data = dataset)

**ANOVA**  
summary(aov(variable~grouping\_variable, data = dataset))  
→ alternative way:  
modelxy <- aov(variable~grouping\_variable, data = dataset)  
summary(modelxy)

**Shapiro-Wilk test**  
shapiro.test(dataset\$variable)

## t-test

**t.test(variable~grouping\_variable, data = dataset)**

**General command structure**  
command1~x, z)

**Frequencies**  
tally(~variable1, data = dataset)  
tally(~variable1, format = "percent", data = dataset)  
tally(variable1~variable2, data = dataset)  
tally(variable1~variable2, format = "percent", data = dataset)

**Median, mean, variance, standard deviation, maximum, minimum**  
median(~variable, data = dataset) / mean(~variable, data = dataset)  
var(~variable, data = dataset) / sd(~variable, data = dataset)  
max(~variable, data = dataset) / min(~variable, data = dataset)  
favstats(~variable, data = dataset)  
→ result: Min, Q1, median, Q3, Max, mean, standard deviation, n, missing values)

**Linear regression**  
summary(lm(dependent\_variable~independent\_variable, data = dataset))

**Multiple regression**  
summary(lm(dependent\_variable~ind\_variable1 + ind\_variable2, data = dataset))

**Principal Component analysis** (package: psych)  
dataset\_new <- cbind(dataset\$variable1, dataset\$variable2, dataset\$variable3...)  
colnames(dataset\_new) <- c("item\_1", "item\_2")  
KMO(dataset\_new)  
pcaX <- princomp(dataset\_new, scores = TRUE, cor = TRUE)  
summary(pcaX) plot(pcaX)  
principal(dataset\_new, nfactors = x, rotate = "varimax")

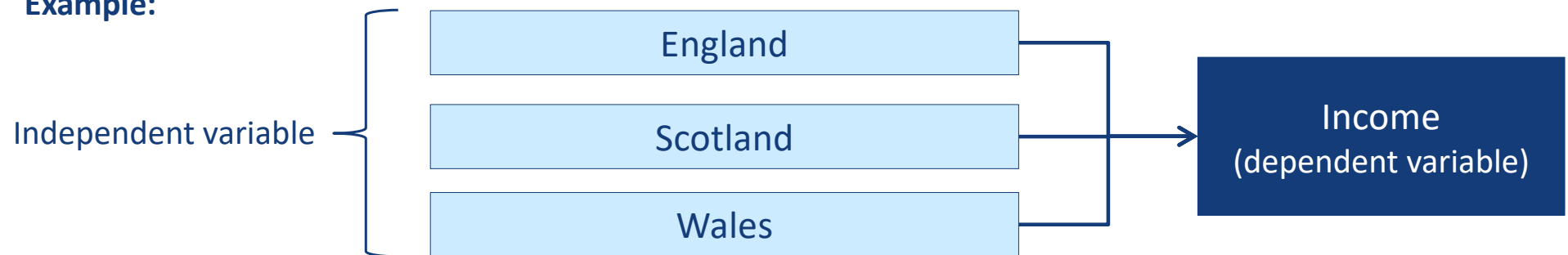
**Cronbach's Alpha** (package: psy)  
factorX <- cbind(dataset\$variable1, dataset\$variable2, dataset\$variable3...)  
cronbach(factorX)

## Variables

- station
- station\_cluster
- service
- informative\_news
- entertaining\_news
- current\_music
- old\_music
- activating\_music
- relaxing\_music
- informative\_presentation
- entertaining\_presentation
- pleasant\_voice
- witty\_comedy
- funny\_comedy
- entertaining\_competitions
- attractive\_prices
- news
- music**
- presentation
- comedy
- competitions
- programme**
- gender**
- age
- age\_groups
- current\_music\_A

1 Installation of R, R Studio & Relevant Packages	
2 Structure of R Studio	
3 R Commands	
4 Exercise Data	
5 Descriptive Statistics	
5.1 Scale of Measurement	
5.2 Frequencies	
5.3 Measures of Central Tendency & Dispersion	
	6 Inferential Statistics
	6.1 Principle & Overview
	6.2 Chi-squared Test
	6.3 t-test
	6.4 ANOVA
	6.5 Shapiro-Wilk Test
	6.6 Wilcoxon Test
	6.7 Correlation
	6.8 Regression
	6.9 Principal Component Analysis

- **Application:**
  - The ANOVA (analysis of variances) tests if the arithmetic means of three or more groups (samples) systematically differ.
- **Leading question:**
  - Do the arithmetic means of three or more groups (samples) differ in a metric variable?
- **Distinction from t-test:**
  - Whilst the t-test examines only two groups (samples), the ANOVA analyses three or more groups (samples)
- **Example:**



# ANOVA: Exercise & R Commands

## Exercise:

- Do the listeners of the different station clusters evaluate the programme of radio stations differently?
- Do the listeners of the different station clusters evaluate the presentation of radio stations differently?

- station
- station\_cluster**

## Variables

- service
- informative\_news
- entertaining\_news
- current\_music
- old\_music
- activating\_music
- relaxing\_music
- informative\_presentation
- entertaining\_presentation
- pleasant\_voice
- witty\_comedy
- funny\_comedy
- entertaining\_competitions
- attractive\_prices

- news
- music
- presentation**
- comedy
- competitions
- programme**
- gender
- age
- age\_groups
- current\_music\_A

### R (mosaic) – Key Commands

**Get going**  
install.packages("xyz")  
require(xyz) <- starting mosaic, psych, psy, ggplot2, openssl, plspm (required for each session)

**Recoding, computing, formatting variables & creating new datasets:**  
**Recoding (new) variables:**  
dataset\$new\_variable[dataset\$old\_variable=="XXX"] <- "new\_value"  
→ Please note: categorical values require quotes (""), numerical values not

**Calculating new variables:**  
dataset\$new\_variable <- dataset\$variable1 + dataset\$variable2

**Changing the variable format:**  
dataset\$new\_variable <- as.numeric(dataset\$variable)  
→ factor to numeric

**Chi<sup>2</sup> test**  
xchisq.test(variable1~variable2, data = dataset)

**t-test**  
t.test(variable~grouping\_variable, data = dataset)

**ANOVA**  
summary(aov(variable~grouping\_variable, data = dataset))  
→ alternative way:  
model <- aov(variable~grouping\_variable, data = dataset)  
summary(model)

**Shapiro-Wilk test**  
shapiro.test(dataset\$variable)

**Mann-Whitney & Wilcoxon test**

## ANOVA

**summary(aov(variable~grouping\_variable, data = dataset))**

command[~<x, z]

**Frequencies**  
tally(~variable1, data = dataset)  
tally(~variable1, format = "percent", data = dataset)  
tally(variable1~variable2, data = dataset)  
tally(variable1~variable2, format = "percent", data = dataset)

**Median, mean, variance, standard deviation, maximum, minimum**  
median(~variable, data = dataset) / mean(~variable, data = dataset)  
var(~variable, data = dataset) / sd(~variable, data = dataset)  
max(~variable, data = dataset) / min(~variable, data = dataset)  
favstats(~variable, data = dataset)  
→ result: Min, Q1, median, Q3, Max, mean, standard deviation, n, missing values

**Linear regression**  
summary(lm(dependent\_variable~independent\_variable, data = dataset))

**Multiple regression**  
summary(lm(dependent\_variable~ind\_variable1 + ind\_variable2, data = dataset))

**Principal Component analysis** (package: psych)  
dataset\_new <- cbind(dataset\$variable1, dataset\$variable2, dataset\$variable3...)  
colnames(dataset\_new) <- c("item\_1", "item\_2")  
KMO(dataset\_new)  
pcaX <- princomp(dataset\_new, scores = TRUE, cor = TRUE)  
summary(pcaX) plot(pcaX)  
principal(dataset\_new, nfactors = x, rotate = "varimax")

**Cronbach's Alpha** (package: psy)  
factorX <- cbind(dataset\$variable1, dataset\$variable2, dataset\$variable3...)  
cronbach(factorX)



1 Installation of R, R Studio & Relevant Packages

2 Structure of R Studio

3 R Commands

4 Exercise Data

5 Descriptive Statistics

5.1 Scale of Measurement

5.2 Frequencies

5.3 Measures of Central Tendency & Dispersion

6 Inferential Statistics

6.1 Principle & Overview

6.2 Chi-squared Test

6.3 t-test

6.4 ANOVA

6.5 Shapiro-Wilk Test

6.6 Wilcoxon Test

6.7 Correlation

6.8 Regression

6.9 Principal Component Analysis

- **Application:**
  - The Shapiro-Wilk tests compares the values of a variable with normally distributed values to determine if the variable is normally distributed.
  - Please note: **A normal distribution is a precondition for the application of a t-test if the sample is small (approx.  $n \leq 50$ ).**
- **Leading question & results:**
  - Are the values of a metric variable normally distributed?
    - If  $p > .10$  (not significant), the distribution does not differ from a normal distribution → **Normal distribution**
    - If  $p \leq .10$  (significant), the distribution differs from a normal distribution → **no normal distribution**
- **Consequence of lack of normal distribution:**
  - Instead of the t-test, Wilcoxon or Mann-Whitney test is applied (small samples)
  - Please note: It normally makes sense to not only conduct a Shapiro-Wilk test but also graphically examine the distribution

```
Shapiro-Wilk normality test
data:  radio$age
W = 0.87431, p-value = 3.079e-14
```

The values are not normally distributed.

# Shapiro-Wilk Test: Exercise & R Commands

## Exercise:

- Is the variable age normally distributed?
- Is the variable programme normally distributed?

**R (mosaic) – Key Commands**

Get going  
install.packages("xyz")  
require(xyz) <- starting mosaic, psych, ggplot2, openssl, plspm (required for each session)

Recoding, computing, formatting variables & creating new datasets:  
Recoding (new) variables:  
dataset\$new\_variable[dataset\$old\_variable=="XXX"] <- "new\_value"  
→ Please note: categorical values require quotes (""), numerical values not

Calculating new variables:  
dataset\$new\_variable <- dataset\$variable1 + dataset\$variable2

Changing the the variable formats:  
dataset\$new\_variable <- as.numeric(dataset\$variable)  
→ factor to numeric  
(1): dataset\$variable <- as.factor(dataset\$variable)  
(2): levels(dataset\$variable) <- c("attribute1", "attribute1")  
→ numeric to factor  
dataset\$new\_variable <- as.factor(dataset\$variable)

Chi<sup>2</sup> test  
xchisq.test(variable1~variable2, data = dataset)

t-test  
t.test(variable~grouping\_variable, data = dataset)

ANOVA  
summary(aov(variable~grouping\_variable, data = dataset))  
→ alternative way:  
modelxy <- aov(variable~grouping\_variable, data = dataset)  
summary(modelxy)

Shapiro-Wilk test  
shapiro.test(dataset\$variable)

Mann-Whitney & Wilcoxon test  
wilcox.test(variable~grouping\_variable, data = dataset)  
(→ Mann-Whitney test)  
wilcox.test(dataset\$variable1, dataset\$variable2, paired = T)

## Shapiro-Wilk test

shapiro.test(dataset\$variable)

tally(-variable1, data = dataset)  
tally(-variable1, format = "percent", data = dataset)  
tally(variable1~variable2, data = dataset)  
tally(variable1~variable2, format = "percent", data = dataset)

Median, mean, variance, standard deviation, maximum, minimum  
median(~variable, data = dataset) / mean(~variable, data = dataset)  
var(~variable, data = dataset) / sd(~variable, data = dataset)  
max(~variable, data = dataset) / min(~variable, data = dataset)  
fvstats(~variable, data = dataset)  
→ result: Min, Q1, median, Q3, Max, mean, standard deviation, n, missing values

multivariate correlation  
summary(lm(dependent\_variable~ind\_variable1 + ind\_variable2, data = dataset))

Principal Component analysis (package: psych)  
dataset\_new <- cbind(dataset\$variable1, dataset\$variable2, dataset\$variable3...)  
colnames(dataset\_new) <- c("item\_1", "item\_2")  
KMO(dataset\_new)  
pcaX <- princomp(dataset\_new, scores = TRUE, cor = TRUE)  
summary(pcaX) plot(pcaX)  
principal(dataset\_new, nfactors = 2, rotate = "varimax")

Cronbach's Alpha (package: psy)  
factorX <- cbind(dataset\$variable1, dataset\$variable2, dataset\$variable3...)  
cronbach(factorX)

## Variables

- station
- station\_cluster
- service
- informative\_news
- entertaining\_news
- current\_music
- old\_music
- activating\_music
- relaxing\_music
- informative\_presentation
- entertaining\_presentation
- pleasant\_voice
- witty\_comedy
- funny\_comedy
- entertaining\_competitions
- attractive\_prices
- news
- music
- presentation
- comedy
- competitions
- programme
- gender
- age
- age\_groups
- current\_music\_A

1 Installation of R, R Studio & Relevant Packages	
2 Structure of R Studio	
3 R Commands	
4 Exercise Data	
5 Descriptive Statistics	
5.1 Scale of Measurement	
5.2 Frequencies	
5.3 Measures of Central Tendency & Dispersion	
	6 Inferential Statistics
	6.1 Principle & Overview
	6.2 Chi-squared Test
	6.3 t-test
	6.4 ANOVA
	6.5 Shapiro-Wilk Test
	6.6 Wilcoxon Test
	6.7 Correlation
	6.8 Regression
	6.9 Principal Component Analysis

## Wilcoxon Test: Brief

- **Application:**
  - The Wilcoxon test is a non-parametric procedure that compares the ranks of a variable between two groups (a normal distribution is not required)
- **Leading question:**
  - Do two groups differ with regard to a variable on an at least ordinal scale of measurement?
- **Example:**

Wilcoxon rank sum test with continuity correction

data: programme by gender

W = 10782, p-value = 0.03801

alternative hypothesis: true location shift is not equal to 0

The groups show differences.

# Wilcoxon Test: Exercise & R Commands

## Exercise:

- According to the Wilcoxon test, do men and women evaluate the programme of radio stations differently?
- According to the Wilcoxon test, do men and women evaluate the music of radio stations differently?

**R (mosaic) – Key Commands**

Get going  
install.packages("xyz")  
require(xyz) <- starting mosaic, psych, ggplot2, openxlsx, plspm (required for each session)

Recoding, computing, formatting variables & creating new datasets:  
Recoding (new) variables:  
dataset\$new\_variable[dataset\$old\_variable=="XXX"] <- "new\_value"  
→ Please note: categorical values require quotes (""), numerical values not

Calculating new variables:  
dataset\$new\_variable <- dataset\$variable1 + dataset\$variable2

Changing the variable format:  
dataset\$new\_variable <- as.numeric(dataset\$variable)  
→ factor to numeric  
(1): dataset\$variable <- as.factor(dataset\$variable)  
(2): levels(dataset\$variable) <- c("attribute1", "attribute1")  
→ numeric to factor

Creating new datasets (is equal, is not equal etc.):  
new\_dataset <- dataset[dataset\$variable == "value",]  
new\_dataset <- dataset[dataset\$variable != "value",]  
new\_dataset <- dataset[dataset\$variable > value,]

Chi<sup>2</sup> test  
xchisq.test(variable1~variable2, data = dataset)

t-test  
t.test(variable~grouping\_variable, data = dataset)

ANOVA  
summary(aov(variable~grouping\_variable, data = dataset))  
→ alternative way:  
modelxy <- aov(variable~grouping\_variable, data = dataset)  
summary(modelxy)

Shapiro-Wilk test  
shapiro.test(dataset\$variable)

Mann-Whitney & Wilcoxon test  
wilcox.test(variable~grouping\_variable, data = dataset)  
(→ Mann-Whitney test)  
wilcox.test(dataset\$variable1, dataset\$variable2, paired = T)  
(→ Wilcoxon signed rank test (paired sample))

Correlation  
cor.test(variable1~variable2, data = dataset)

### Mann-Whitney & Wilcoxon test

**wilcox.test(variable~grouping\_variable, data = dataset)**

(→Mann-Whitney test)

favstats()~variable, data = dataset)  
→ result: Min, Q1, median, Q3, Max, mean, standard deviation, n, missing values)

principal()(dataset\_new, n.factors = x, rotate = "varimax")

Cronbach's Alpha (package: psych)  
factorX <- cbind(dataset\$variable1, dataset\$variable2, dataset\$variable3...)  
cronbach(factorX)

## Variables

- station
- station\_cluster
- service
- informative\_news
- entertaining\_news
- current\_music
- old\_music
- activating\_music
- relaxing\_music
- informative\_presentation
- entertaining\_presentation
- pleasant\_voice
- witty\_comedy
- funny\_comedy
- entertaining\_competitions
- attractive\_prices
- news
- **music**
- presentation
- comedy
- competitions
- **programme**
- **gender**
- age
- age\_groups
- current\_music\_A

1 Installation of R, R Studio & Relevant Packages	
2 Structure of R Studio	
3 R Commands	
4 Exercise Data	
5 Descriptive Statistics	
5.1 Scale of Measurement	
5.2 Frequencies	
5.3 Measures of Central Tendency & Dispersion	
	6 Inferential Statistics
	6.1 Principle & Overview
	6.2 Chi-squared Test
	6.3 t-test
	6.4 ANOVA
	6.5 Shapiro-Wilk Test
	6.6 Wilcoxon Test
	6.7 Correlation
	6.8 Regression
	6.9 Principal Component Analysis

## Correlation: Brief

- **Application:**
  - A correlation analysis determines the linear association of two metric variables.
- **Leading question:**
  - How strong and in what direction is the linear association of two metric variables?  
→ Does the value of variable A increase or decrease if the value of variable B increases or decreases (et vice versa)?
- **Steps of the analysis:**
  - 1) Is the p-value  $< .05$  ( $< .01$ ,  $< .001$ )? → Significance?
  - 2) Which value does the correlation coefficient have?
- **Possible interpretation of the correlation coefficient:**

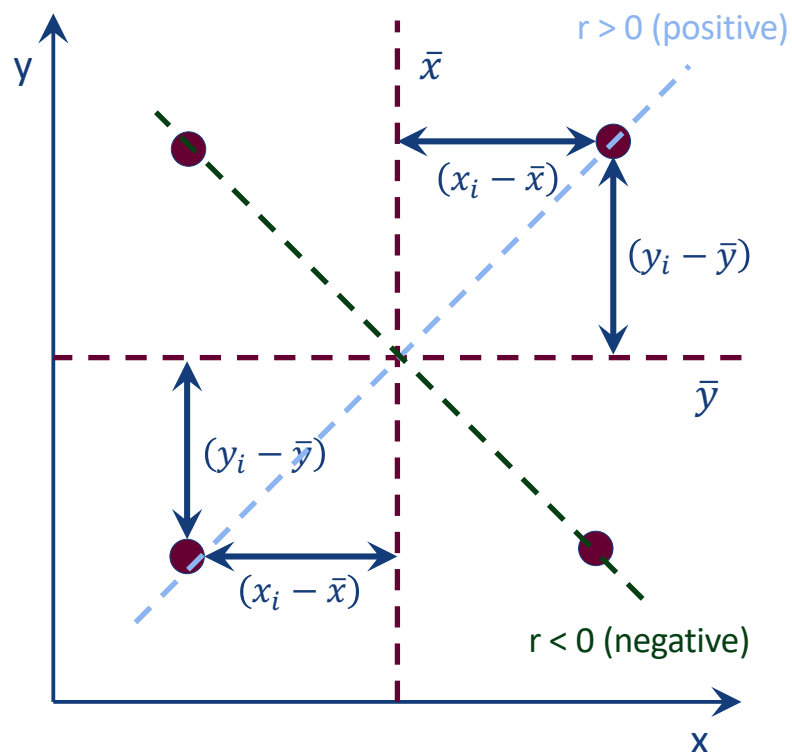
$0 < r \leq +1.0$	Correlation in the same direction
$-1.0 \geq r < 0$	Correlation in the opposite direction

$0.0 <  r  \leq 0.2$	weak to no correlation
$0.2 <  r  \leq 0.4$	weak correlation
$0.4 <  r  \leq 0.6$	moderate correlation
$0.6 <  r  \leq 0.8$	strong correlation
$0.8 <  r  \leq 1.0$	very strong correlation



## Correlation: Brief

- Statistical explanation of the correlation coefficient



**Covariance** (shared dispersion of two variables):

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

**Variance** (dispersion of a variable):

$$\text{var} = \sigma_x = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2$$

**Correlation coefficient** (ratio of the shared dispersion to the total dispersion):

$$r = \frac{\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\text{var}_x * \text{var}_y}}$$

# Correlation: Exercise & R Commands

## Exercise:

- Is there an association between the evaluation of music and presentation?
- Is there an association between the evaluation of comedy and competitions?

- |   |  |
|---|--|
| <ul style="list-style-type: none"> <li>station</li> <li>station_cluster</li> <li>service</li> <li>informative_news</li> <li>entertaining_news</li> <li>current_music</li> <li>old_music</li> <li>activating_music</li> <li>relaxing_music</li> <li>informative_presentation</li> <li>entertaining_presentation</li> <li>pleasant_voice</li> <li>witty_comedy</li> <li>funny_comedy</li> <li>entertaining_competitions</li> <li>attractive_prices</li> </ul> | <p><b>Variables</b></p> <ul style="list-style-type: none"> <li>news</li> <li><b>music</b></li> <li><b>presentation</b></li> <li><b>comedy</b></li> <li><b>competitions</b></li> <li>programme</li> <li>gender</li> <li>age</li> <li>age_groups</li> <li>current_music_A</li> </ul> |
|---|--|

Prof. Dr. Godbersen
Chi² test
Prof. Dr. Godbersen

## Correlation

**cor.test(variable1~variable2, data = dataset)**

→ Please note: categorical values require quotes (""), numerical values not

**Calculating new variables:**

```
dataset$new_variable <- dataset$variable1 + dataset$variable2
```

**Changing the variable format:**

```
dataset$new_variable <- as.numeric(dataset$variable)
```

→ factor to numeric

```
(1) dataset$variable <- as.factor(dataset$variable)
(2) levels(dataset$variable) <- c("attribute1", "attribute1")
```

→ numeric to factor

**Creating new datasets (is equal, is not equal etc.):**

```
new_dataset <- dataset[dataset$variable == "value",]
new_dataset <- dataset[dataset$variable != "value",]
new_dataset <- dataset[dataset$variable > value,]
```

**General command structure**

```
command1~x, z)
```

**Frequencies**

```
tally(~variable1, data = dataset)
tally(~variable1, format = "percent", data = dataset)
tally(variable1~variable2, data = dataset)
tally(variable1~variable2, format = "percent", data = dataset)
```

**Median, mean, variance, standard deviation, maximum, minimum**

```
median(~variable, data = dataset) / mean(~variable, data = dataset)
var(~variable, data = dataset) / sd(~variable, data = dataset)
max(~variable, data = dataset) / min(~variable, data = dataset)
favstats(~variable, data = dataset)
```

→ result: Min, Q1, median, Q3, Max, mean, standard deviation, n, missing values

**Shapiro-Wilk test**

```
shapiro.test(dataset$variable)
```

**Mann-Whitney & Wilcoxon test**

```
wilcox.test(variable~grouping_variable, data = dataset)
(→ Mann-Whitney test)
wilcox.test(dataset$variable1, dataset$variable2, paired = T)
(→ Wilcoxon signed rank test (paired sample))
```

**Correlation**

```
cor.test(variable1~variable2, data = dataset)
```

**Linear regression**

```
summary(lm(dependent_variable~independent_variable, data = dataset))
```

**Multiple regression**

```
summary(lm(dependent_variable~ind_variable1 + ind_variable2, data = dataset))
```

**Principal Component analysis** (package: psych)

```
dataset_new <- cbind(dataset$variable1, dataset$variable2, dataset$variable3...)
colnames(dataset_new) <- c("item_1", "item_2")
KMO(dataset_new)
pcaX <- princomp(dataset_new, scores = TRUE, cor = TRUE)
summary(pcaX) plot(pcaX)
principal(dataset_new, nfactors = x, rotate = "varimax")
```

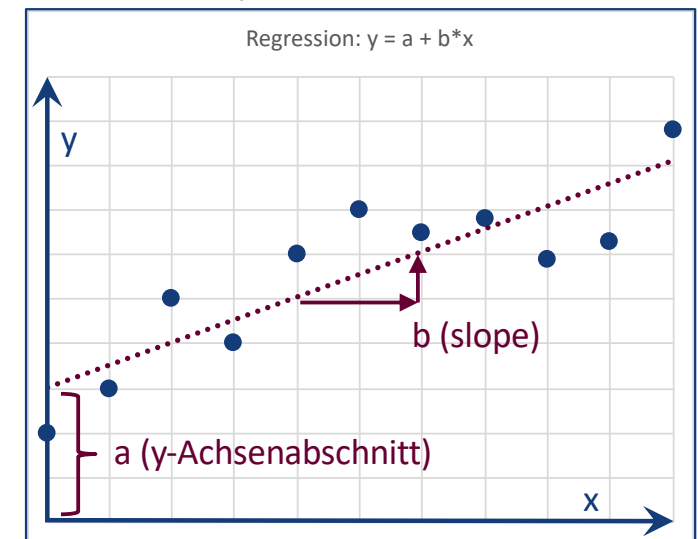
**Cronbach's Alpha** (package: psy)

```
factorX <- cbind(dataset$variable1, dataset$variable2, dataset$variable3...)
cronbach(factorX)
```

Prof. Dr. Hendrik Godbersen: R (mosaic) - Key Commands
1
Prof. Dr. Hendrik Godbersen: R (mosaic) - Key Commands
2

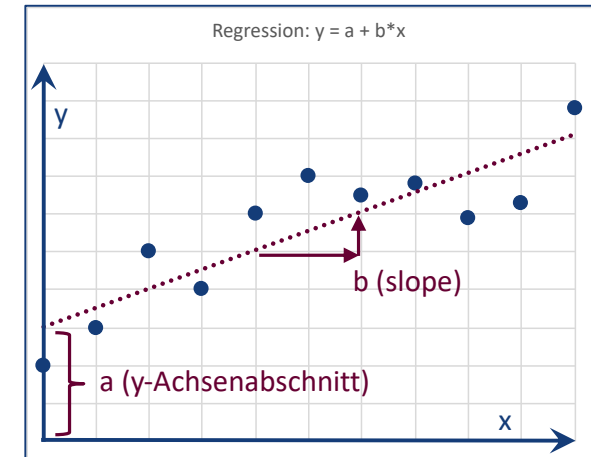
1 Installation of R, R Studio & Relevant Packages	
2 Structure of R Studio	
3 R Commands	
4 Exercise Data	
5 Descriptive Statistics	
5.1 Scale of Measurement	
5.2 Frequencies	
5.3 Measures of Central Tendency & Dispersion	
	6 Inferential Statistics
	6.1 Principle & Overview
	6.2 Chi-squared Test
	6.3 t-test
	6.4 ANOVA
	6.5 Shapiro-Wilk Test
	6.6 Wilcoxon Test
	6.7 Correlation
	6.8 Regression
	6.9 Principal Component Analysis

- **Application:**
  - The linear regression examines the linear effect of a metric variable (simple regression) or multiple metric variables (multiple regression) on a metric variable (dependent variable).
- **Leading question:**
  - How strong is the effect of a metric variable (simple regression) or multiple metric variables (multiple regression) on another metric variable?
- **Regression model/line (result of an analysis):**
  - $y = a + b \cdot x$
  - y: dependent variable („effect“)
  - x: independent variable („cause“)
  - a: y-intercept
  - b: regression coefficient (“much y increases – if x increases by one “)
- **Statistical analysis (“in the background“):**
  - The regression analysis determines a line whose squared deviations from the observed values are the least (method of least squares)



- Steps of the analysis:

- Are the p-values  $< .05$  ( $< .01$ ,  $< .001$ )?
  - Does  $x_i$  significantly effects  $y$  (p-values)?
- How are the regression coefficients (b)?
  - How strong is the effect of  $x_i$  on  $y$ ?
- What is the value for  $R^2$  (coefficient of determination)?
  - What is the explanatory power of the regression model?
  - $R^2$  = proportion of the variance of the dependent variable that can be explained by the independent variables



```
Call:
lm(formula = programme ~ presentation, data = radio)

Residuals:
    Min       1Q   Median       3Q      Max
-48.169  -6.168   1.293   8.611  50.672

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  49.32790    2.26833   21.75  <2e-16 ***
presentation  0.42062    0.03162   13.30  <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 13.36 on 273 degrees of freedom
Multiple R-squared:  0.3933, Adjusted R-squared:  0.3911
F-statistic: 177 on 1 and 273 DF, p-value: < 2.2e-16
```

1) p-value

2) regression coefficient

3)  $R^2$

## Regression: Exercise & R Commands

### Exercise:

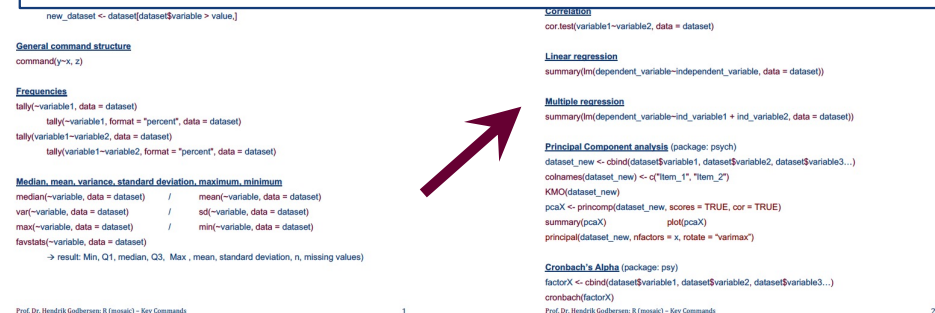
- Which effect does music have on the overall programme?
- Which effect does comedy have on the overall programme?
- Which effect do the five programme elements have on the overall programme?

### Linear regression

```
summary(lm(dependent_variable~independent_variable, data = dataset))
```

### Multiple regression

```
summary(lm(dependent_variable~ind_variable1 + ind_variable2, data = dataset))
```



The screenshot shows two pages of R commands. Page 1 (left) includes commands for creating a new dataset, general command structure, frequencies, median/mean/variance/standard deviation/maximum/minimum, and factor statistics. Page 2 (right) includes commands for correlation, linear regression, multiple regression, principal component analysis, and Cronbach's Alpha. A red arrow points from the 'Multiple regression' section on page 1 to the 'Correlation' section on page 2.

### Variables

- station
- station\_cluster
- service
- informative\_news
- entertaining\_news
- current\_music
- old\_music
- activating\_music
- relaxing\_music
- informative\_presentation
- entertaining\_presentation
- pleasant\_voice
- witty\_comedy
- funny\_comedy
- entertaining\_competitions
- attractive\_prices
- news
- music
- presentation
- comedy
- competitions
- programme
- gender
- age
- age\_groups
- current\_music\_A

1 Installation of R, R Studio & Relevant Packages

2 Structure of R Studio

3 R Commands

4 Exercise Data

5 Descriptive Statistics

5.1 Scale of Measurement

5.2 Frequencies

5.3 Measures of Central Tendency & Dispersion

6 Inferential Statistics

6.1 Principle & Overview

6.2 Chi-squared Test

6.3 t-test

6.4 ANOVA

6.5 Shapiro-Wilk Test

6.6 Wilcoxon Test

6.7 Correlation

6.8 Regression

6.9 Principal Component Analysis

# Principal Component Analysis (& Cronbach's $\alpha$ ): Brief

- **Application:**
  - The principal component analysis is applied to summarise variables/indicators/items on fewer dimensions.
- **Leading question:**
  - How can multi-dimensional metric variables be allocated to fewer principal components?
- **Principle idea:**
  - The items of a component should strongly correlate with each other.
  - The items from different components should not or only weakly correlate with each other.

Item	Perfectionistic and quality conscious buyer	Confused and poorly informed buyer	Fast and careless buyer	Novelty and innovation seeking buyer	Habitual and loyal buyer	Price conscious buyer	Brand and reputation conscious buyer	Cronbach's alpha
Procuring very good quality is very important for my company.	.78	-.10	-.08	.11	.15	.06	.13	.88
When purchasing products or services, my company attempts to get the very best or perfect choice.	.69				.17	-.13	.06	
My company normally attempts to buy the best overall quality.	.82				.07	.02	.09	
My company puts special efforts into choosing the very best quality products or services.	.79				.04	.00	.08	
My company's standards and expectations for product or service purchases are very high.	.74	-.21	-.11	-.07	.02	-.02	.16	
My company takes the time to carefully purchase the best option.	.67	.00	-.14	.35	-.07	-.17	-.11	
Suppliers that do the most advertising are usually very good choices for my company.	.03	.59	.27	.27	.12	.18	.03	.81
There are so many suppliers to choose from that the buying agents in my company get often confused.	-.21	.73	.27	.06	.04	-.01	.09	
Sometimes it is hard for my company to choose which suppliers to buy from.	.03			.10	.07	-.07	.02	
The more information my company receives about products or services, the harder it is for the buying agent to choose the best option	-.06			.3	-.01	.16	.10	
All the information my company gets on different products or services confuses the buying agents.	.10			.05	-.03	.10	.05	
When purchasing products or services, my company does not give so much thought or care.	-.20	-.23	-.00	-.04	.11	.11	.00	
In my company, procurement is done very quickly, buying the first product or service or from the first supplier that seems good enough	-.17	.38	.56	.06	.03	.21	.19	.77
My company should plan its procurement more carefully than it usually does.	-.01	.28	.56	-.09	.05	.00	.24	
Decision Making Styles: My company buys impulsively.	-.05	.20	.72	.05	-.07	.03	.03	
My company makes thoughtless purchases that it later regrets.	-.16	.36	.61	-.04	.03	.18	-.11	

...

Source: Godbersen, H., Gully, V. (2022): B2B Decision-making Styles Scale. Zusammenstellung sozialwissenschaftlicher Items und Skalen (ZIS).



# Principal Component Analysis (& Cronbach's $\alpha$ ): Brief

Prof. Dr.  
Godbersen

- Steps of the analysis
  - 0) Preparation of a new dataset
  - 1) Test if a principal component analysis should be conducted through KMO (Kaiser-Meyer-Olkin test determines the proportion of shared variance among the variables)
  - 2) Determining the number of components through eigenvalues
    - Criterion: components with eigenvalues > 1
  - 3) Determining the loadings (correlation of an item with a component) & allocation of items to components:
    - Precondition: loading > .5
    - Decision: allocation of an item to a component according to the highest loading
  - 4) Interpretation (naming) of components
  - 5) Determining the reliability/one-dimensionality of each of the components through Cronbach's alpha (average inter-item correlation)

## KMO – accepted values:

- 0.00 to 0.49 - unacceptable
- 0.50 to 0.59 - miserable
- **0.60** to 0.69 - mediocre
- **0.70** to 0.79 - middling
- 0.80 to 0.89 - meritorious
- 0.90 to 1.00 - marvellous

Kaiser-Meyer-Olkin factor adequacy  
Call: KMO(r = pca\_radio\_data)  
Overall MSA = 0.85

Importance of components:

	Comp.1	Comp.2	Comp.3	Comp.4	Comp.5	Comp.6	Comp.7	Comp.8
Standard deviation	2.359856	1.2124941	1.16061942	1.00739666	0.98060991	0.84432456	0.78977668	0.73846855
Proportion of Variance	0.397780	0.1050101	0.09621696	0.07248915	0.06868541	0.05092028	0.04455337	0.03895256
Cumulative Proportion	0.397780	0.5027901	0.59900708	0.67149623	0.74018164	0.79110192	0.83565530	0.87460785

	Comp.9	Comp.10	Comp.11	Comp.12
Standard deviation	0.63632144	0.63167541	0.51744946	0.50533062
Proportion of Variance	0.02892178	0.02850099	0.01912528	0.01823993
Cumulative Proportion	0.90352964	0.93203063	0.95115591	0.96939584

4 components

Principal Components Analysis  
Call: principal(r = pca\_radio\_data, nfactors = 4, rotate = "varimax")  
Standardized loadings (pattern matrix) based upon correlation matrix

	RC1	RC4	RC3	RC2	h2	u2	com
service	0.16	0.78	0.18	-0.10	0.67	0.33	1.2
informative_news	0.12	0.90	0.03	0.02	0.83	0.17	1.0
Entertaining_news	0.48	0.47	0.04	0.35	0.58	0.42	2.8
current_music	0.19	-0.02	-0.02	0.77	0.62	0.38	1.1
old_music	0.29	0.16	0.65	-0.12	0.55	0.45	0.8

## Allocation of items

## Cronbach's $\alpha$ – accepted values:

- > 0.6 ("hard" criterion)
- > 0.7 ("soft" criterion)

```
> cronbach(comp_entertainment)
$sample.size
[1] 275

$number.of.items
[1] 5

$alpha
[1] 0.8511712
```

# Principal Component Analysis (& Cronbach's $\alpha$ ): Exercise & R Commands

## Exercise:

- Which components can be formed from the programme attributes? ("How can the programme attributes be combined to form new variables?")
- Evaluate the reliability (one-dimensionality) of the new components?

### Principal Component analysis (package: psych)

```
dataset_new <- cbind(dataset$variable1, dataset$variable2...)
colnames(dataset_new) <- c("Item_1", "Item_2")
KMO(dataset_new)
pcaX <- princomp(dataset_new, scores = TRUE, cor = TRUE)
summary(pcaX)                plot(pcaX)
principal(dataset_new, nfactors = x, rotate = "varimax")
```

### Cronbach's Alpha (package: psy)

```
factorX <- cbind(dataset$variable1, dataset$variable2...)
cronbach(factorX)
```

## Variables

- station
- station\_cluster
- service
- informative\_news
- entertaining\_news
- current\_music
- old\_music
- activating\_music
- relaxing\_music
- informative\_presentation
- entertaining\_presentation
- pleasant\_voice
- witty\_comedy
- funny\_comedy
- entertaining\_competitions
- attractive\_prices
- news
- music
- presentation
- comedy
- competitions
- programme
- gender
- age
- age\_groups
- current\_music\_A

Prof. Dr. Hendrik Godbersen