DRS Spring Kurs: Biostatistik

Lorenz Gygax & Vitaly Belik

Institute for Veterinary Epidemiology and Biostatistics, FU Berlin

Mar 11, 2020

Schedule

URL: http://belik.userpage.fu-berlin.de/springschool

	Wednesday, 11th March	Thursday, 12th March
8.30 – 10.00	R Intro (VB)	Proportions, Crosstabs, RR/IR and OR + Exercise (LG)
Break		
10.30 – 12.00	Basics of statistical hypothesis testing, general aspects in choice of a model + Exercise (LG)	Linear regression models, regression parameters and fit, (G)LM LinRreg (with ANOVA table) + Exercise (LG)
Lunch		
13.00 – 14.00	Descriptive statistics. Association concept, Graphical representation + Exercise (VB)	Model diagnostics, fit, outliers, residuals, assumptions (transformations, alternative distributions) + Exercise (LG)
Break		
14:15-15:15	ANOVA. Simple linear models + Exercise (VB)	Exercise: Model diagnostics, fit, outliers, residuals, assumptions (transformations, alternative distributions) (LG)
		Odds Ratios from logistic regression (VB)
Break		
15:30 -16:30	Exercise (VB)	Exercise: Odds Ratios from logistic regression (VB)
Break		
17:00-18:00	Extended exercises, self-study	Extended exercises, self-study

Contact

Prof. Dr. Vitaly Belik

Fachbereich Veterinärmedizin

Institut für Veterinär-Epidemiologie und Biometrie

Juniorprofessor

Leitung Arbeitsgruppe Systemmodellierung

Adresse	Königsweg 67 Raum 104 14163 Berlin
Telefon	+49 30 838 61129
Fax	+49 30 838 4 61129
E-Mail	vitaly.belik@fu-berlin.de
Homepage	Working Group Modelling

What is Statistics / Biostatistics?

- What are your expectations to this course?

What for could statistics be useful?

1. Task definition. After the question has been formulated precisely, a suitable choice of characteristics must be made, a measurement or observation method determined and an experimental plan can be drawn up.
2. Data acquisition. Obtaining the test material (taking the sample) and carrying out the measurements or observations on this material.
3. Data processing. The data obtained must be prepared graphically and mathematically, then conclusions must be drawn from the sample of the population; these are then checked and interpreted.

What is Statistics / Biostatistics? (1)

Statistics

is a scientific discipline, the object of which is the development and application of methods for data collection, description and analysis as well as the evaluation of the results. A distinction is made between:

Descriptive Statistics:

Methods for evaluating and clearly presenting and summarizing data.

Inference Statistics:

Methods for making sensible decisions in under uncertainty or risk.
"Getting a grip on chance". "Gain security over uncertainty".

Biostatistics

applied statistics for the description, modeling and assessment of biological-scientific phenomena.

Examples

- How certain is the result of a diagnostic test to detect a disease?
- How many trials have to be carried out to ensure improvement of a product (e.g. test)?

Descriptive statistics are sometimes referred to as *exploratory* and conclusive statistics as *confirmatory data analysis*.

Statistical Inference

- Estimating the population's unknown parameters. "Estimate a parameter from the data in the sample to be *as close as possible* to the unknown reality."
- Specification of confidence intervals (areas of trust). "Based on the data of the sample, provide an interval that overlaps with the true value (population parameter) with a certain probability."
- Decide by means of a statistical test whether, based on the data of the sample, a statement about a parameter of the population (e.g. difference in means between groups) is true or false.

Learning objectives of the course

The aim of the course is to provide you with the most important statistical methods for planning and evaluating the experiments and data from scientific studies.

You should understand the necessities, possibilities and limits of basic statistical analyzes and be able to carry out simple statistical calculations themselves.

If necessary, you should be able to communicate your concerns securely to statistical consultants.

Literature

1. R. Kabacoff. *R in Action*
2. A. Field *Discovering Statistics Using R*
3. M. Crawley. *The R Book*
4. I. Dohoo et al. *Veterinary Epidemiologic Research*
5. A. Field. *An Adventure in Statistics: The Reality Enigma*

Data as a Table

	survived	pclass	sex	age	sibsp	parch	fare	embarked	class	who	adult_male	deck	embark_town	alive	alone
0	0	3	male	22.0	1	0	7.2500	S	Third	man	True	NaN	Southampton	no	False
1	1	1	female	38.0	1	0	71.2833	C	First	woman	False	C	Cherbourg	yes	False
2	1	3	female	26.0	0	0	7.9250	S	Third	woman	False	NaN	Southampton	yes	True
3	1	1	female	35.0	1	0	53.1000	S	First	woman	False	C	Southampton	yes	False
4	0	3	male	35.0	0	0	8.0500	S	Third	man	True	NaN	Southampton	no	True
5	0	3	male	NaN	0	0	8.4583	Q	Third	man	True	NaN	Queenstown	no	True
6	0	1	male	54.0	0	0	51.8625	S	First	man	True	E	Southampton	no	True
7	0	3	male	2.0	3	1	21.0750	S	Third	child	False	NaN	Southampton	no	False
8	1	3	female	27.0	0	2	11.1333	S	Third	woman	False	NaN	Southampton	yes	False
9	1	2	female	14.0	1	0	30.0708	C	Second	child	False	NaN	Cherbourg	yes	False
10	1	3	female	4.0	1	1	16.7000	S	Third	child	False	G	Southampton	yes	False
11	1	1	female	58.0	0	0	26.5500	S	First	woman	False	C	Southampton	yes	True
12	0	3	male	20.0	0	0	8.0500	S	Third	man	True	NaN	Southampton	no	True
13	0	3	male	39.0	1	5	31.2750	S	Third	man	True	NaN	Southampton	no	False
14	0	3	female	14.0	0	0	7.8542	S	Third	child	False	NaN	Southampton	no	True
15	1	2	female	55.0	0	0	16.0000	S	Second	woman	False	NaN	Southampton	yes	True

Features (characteristics)

Individuals or objects of investigation that are the basis of a survey / investigation, i.e. on / from which data is collected, is referred to as a *statistical unit*, feature carrier or examination units.

The properties that are examined with regard to the objective of the statistical unit are called *features* or *characteristics* .

Example: students' data

- gender, height, year of birth
- grew up close to the city or in the country
- the wish to work in a specific company after graduation

Selection of suitable features

Objectivity

The form of the characteristic must be clearly determined regardless of the person of the evaluator.

Reliability

The feature allows reproducible measurement (or observation) results, so when repeated, the results are the same.

Validity

The characteristic in its manifestations reflects the essential properties for the question. Also called “validity” or “meaningfulness”.

Characterization of features

quantitative features:

Examination units differ in absolute (numerical) value. - e.g. Age, weight, temperature, number of germs, company size, pollutant content, ...

qualitative features:

Examination units differ in their characteristics (type) - e.g. Gender, name, medical finding, race, therapy, type of husbandry, region, ...

Scales of features

nominal (qualitative, categorical) Scale:

the values are not ranked and are not comparable. - e.g. Name, gender, race, attitude, form of therapy, pathological classification

ordinal (qualitative oder semi-quantitative) Scale:

the values are ranked, but the distances between the values on the scale cannot be interpreted. - e.g. Evaluation (ratings, grades), state of health, degree of contamination with germs (-, +, ++, +++)

metric (quantitative) Scale:

the values are ranked and the distances between the values on the scale can be interpreted. - z.B. Gewicht, Betriebsgröße, Keimzahlen, ...

Scales of features (1)

A distinction is also made between

interval scale

The distances between characteristic values can be compared. The Scale is continuous.

- e.g. Temperature in degrees Celsius

Ratio scale

Not only the difference, but also the ratio of two measured values may be used.

- e.g. Temperature in Kelvin, length in centimeters

Scales features (2)

Options for statistical evaluation depend on the scale type, because more information can be recorded and evaluated at a higher scale type than with lower scale type.

Besides we should consider the effort for the additional information gain.

Not every number is a number.

(categorical) data is often encrypted to facilitate subsequent data processing

- School grades: 1, 2, 3, 4, 5, 6 (ordinal)
- Test result: 1, 0 (nominal)
- District codes: 3253, 3351 (nominal)

Descriptive Statistics

Deskriptive Statistics (1)

Tables

Charts

Charactersistic numbers

Tables

Table 1: Titanic dataset

Index	survived	pclass	sex	age	deck	fare	alone
0	0	3	male	22		7.2500	False
1	1	1	female	38	C	71.2833	False
2	1	3	female	26		7.9250	True
3	1	1	female	35	C	53.1000	False
4	0	3	male	35		8.0500	True
5	0	3	male	NA		8.4583	True
6	0	1	male	54	E	51.8625	True
7	0	3	male	2		21.0750	False
8	1	3	female	27		11.1333	False
9	1	2	female	14		30.0708	False

Frequency table

Table 2: Frequency table

age	freq
0.42	1
0.67	1
0.75	2
0.83	2
0.92	1
1	7
2	10
3	6
4	10
5	4
6	3
7	3
8	4
9	8
10	2
11	4
12	1
13	2
14	6
14.5	1
15	5
16	17
17	13
18	26
19	25
20	15
20.5	1
21	24
22	27
23	15

Charts

Bar chart

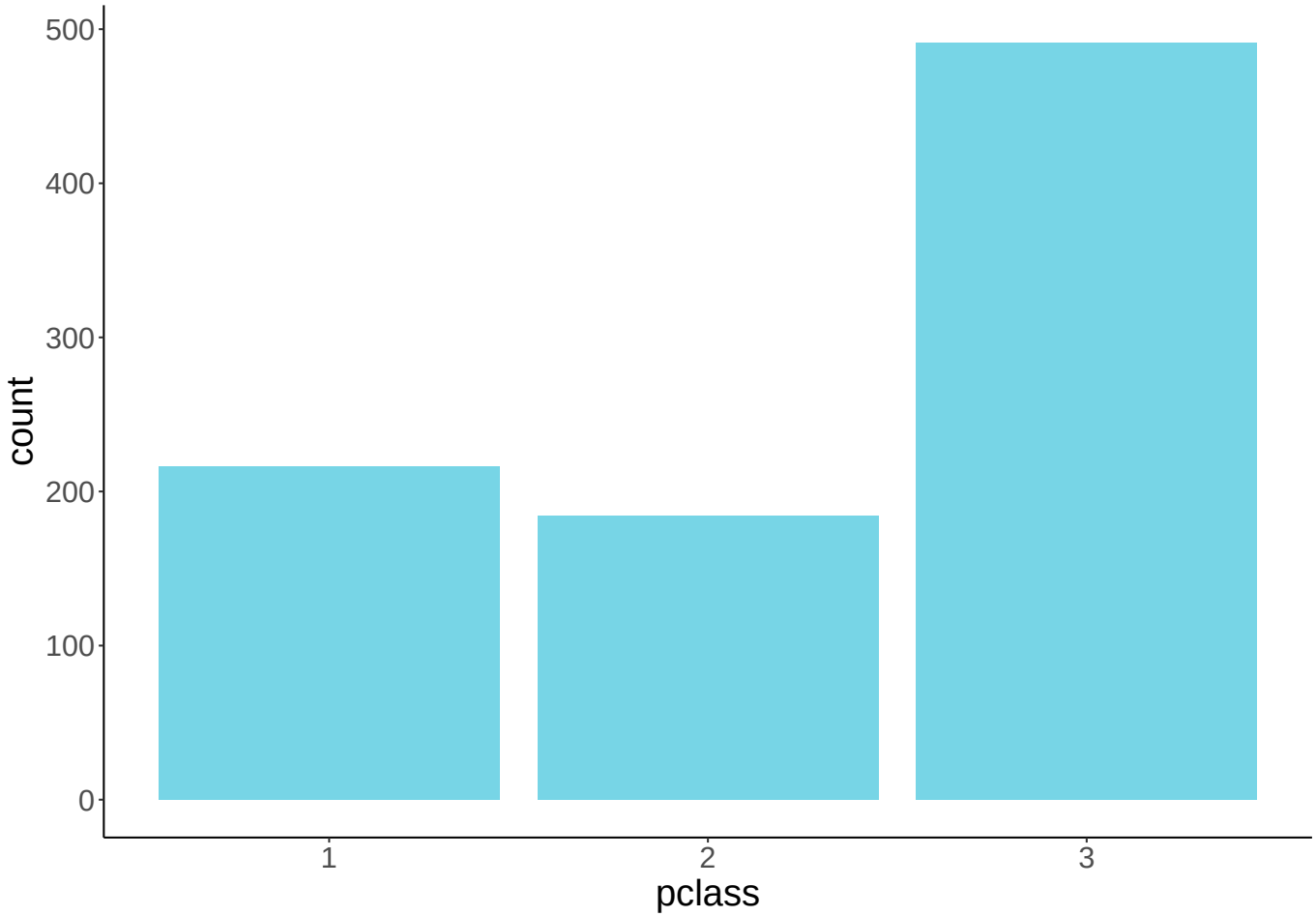


Figure 1: Bar chart

Bar chart (relative frequency)

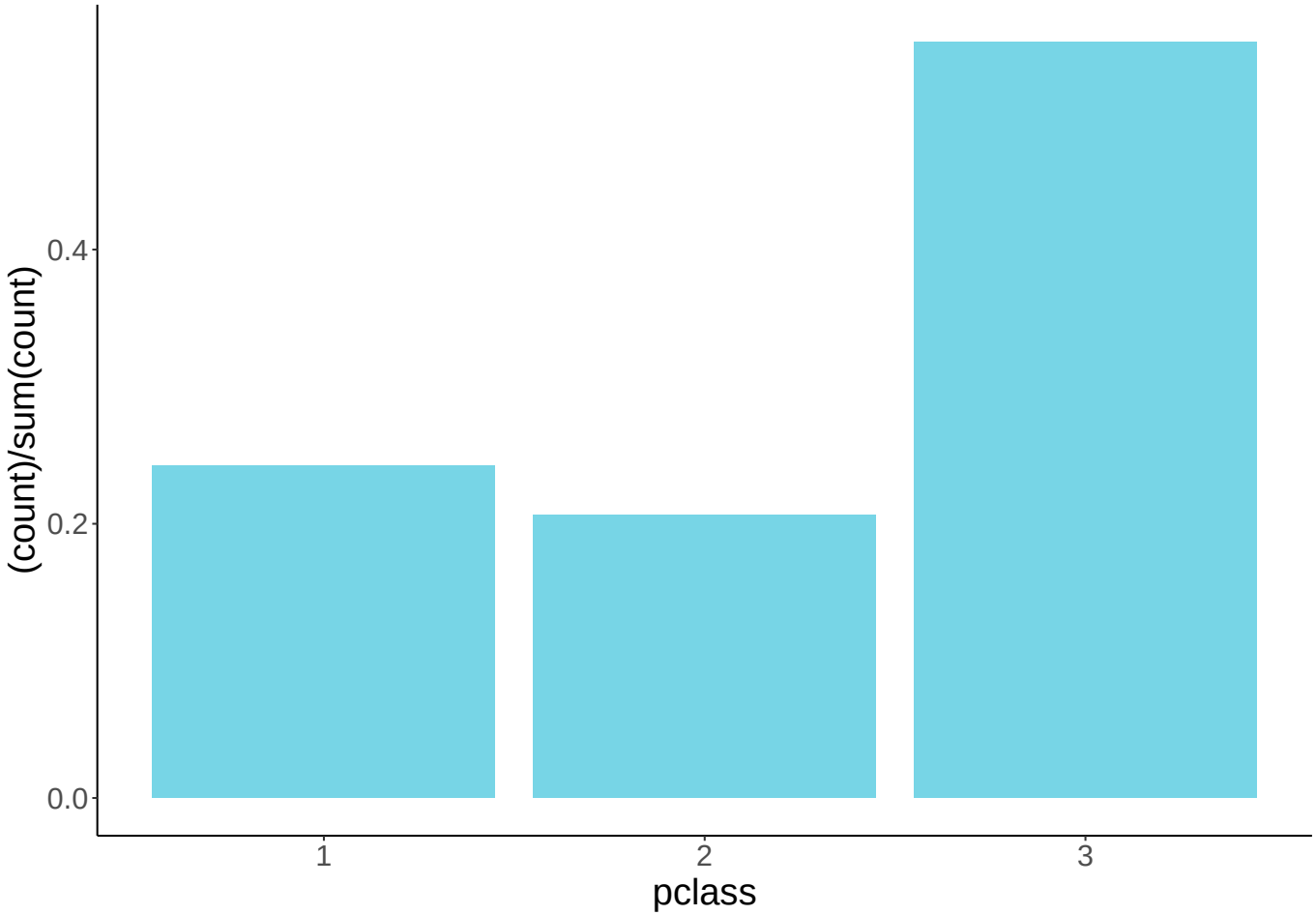


Figure 2: Bar chart (relative frequency)

Component bar chart

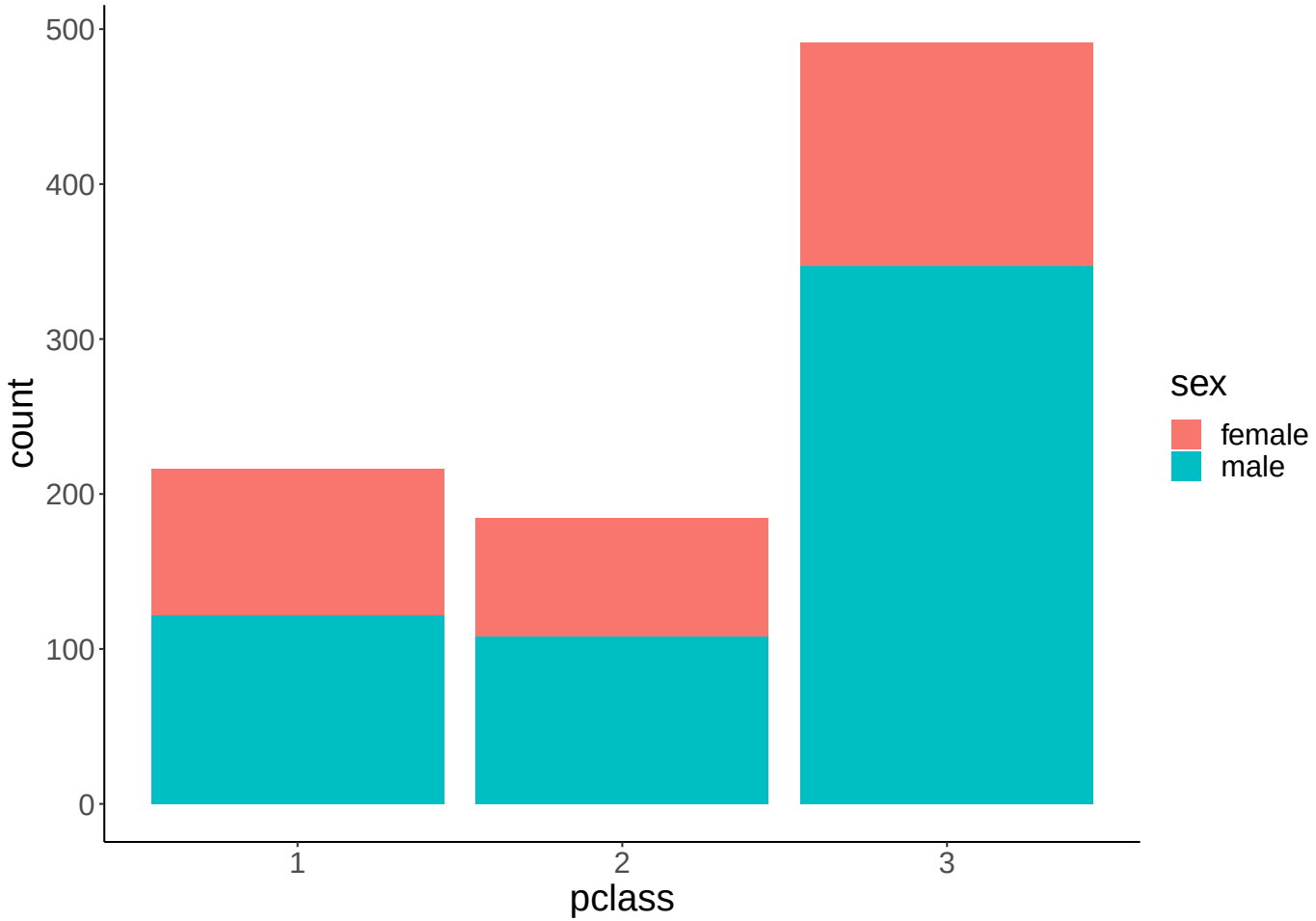


Figure 3: Component bar chart

Component bar chart (relative frequency)

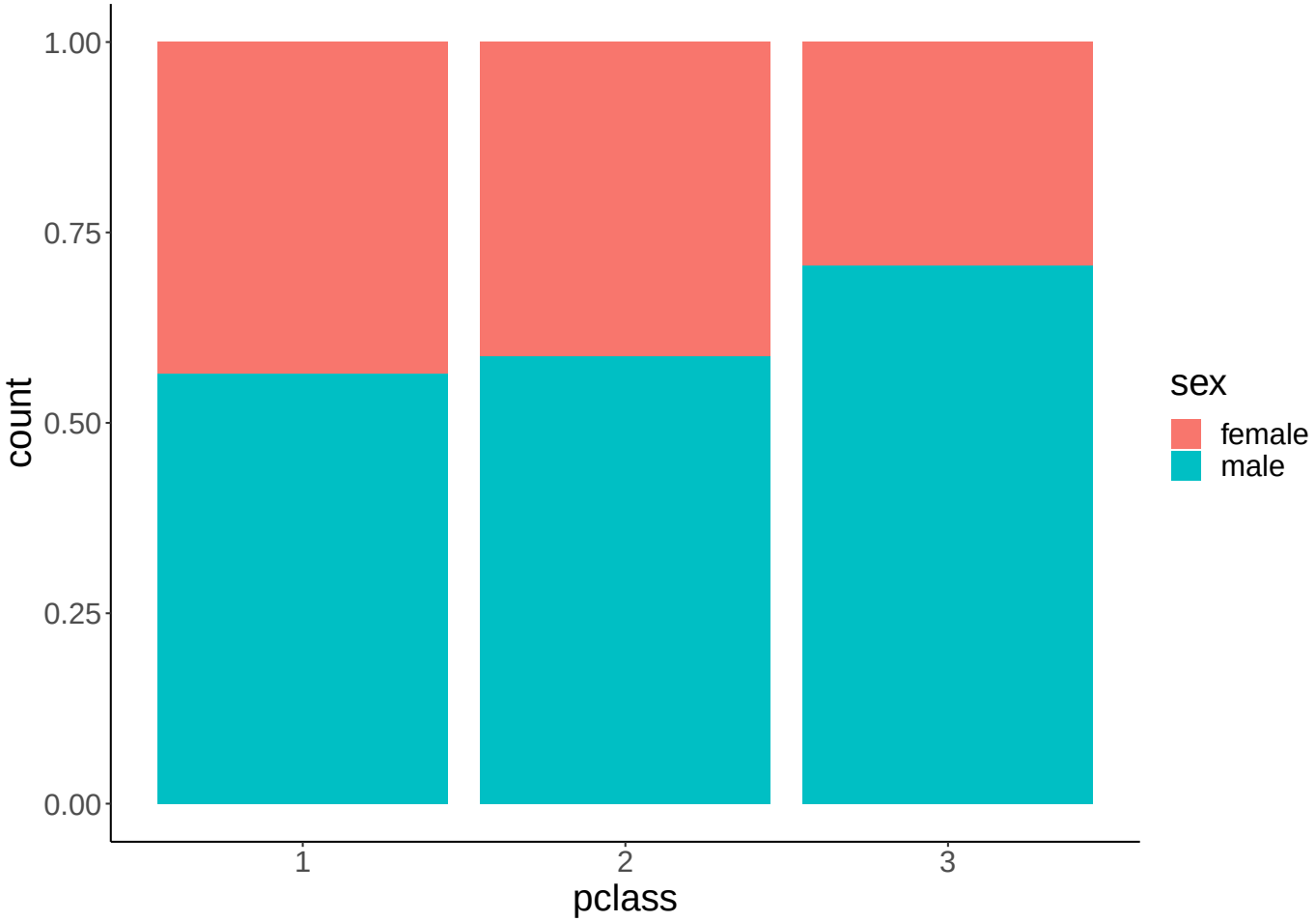


Figure 4: Component bar chart (relative frequency)

Component bar chart (relative frequency) (1)

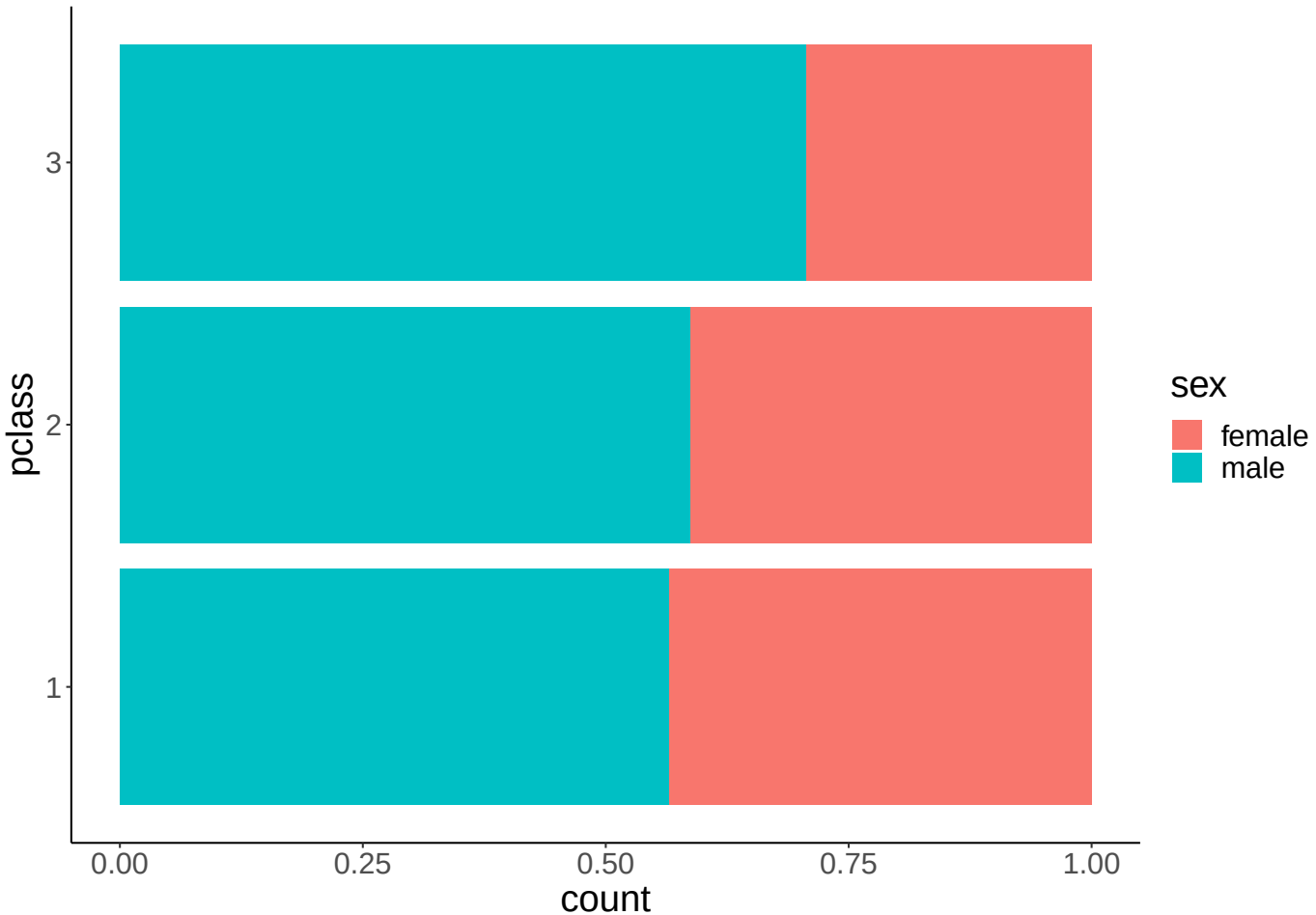


Figure 5: Component bar chart (relative frequency). Flipped.

Histogram

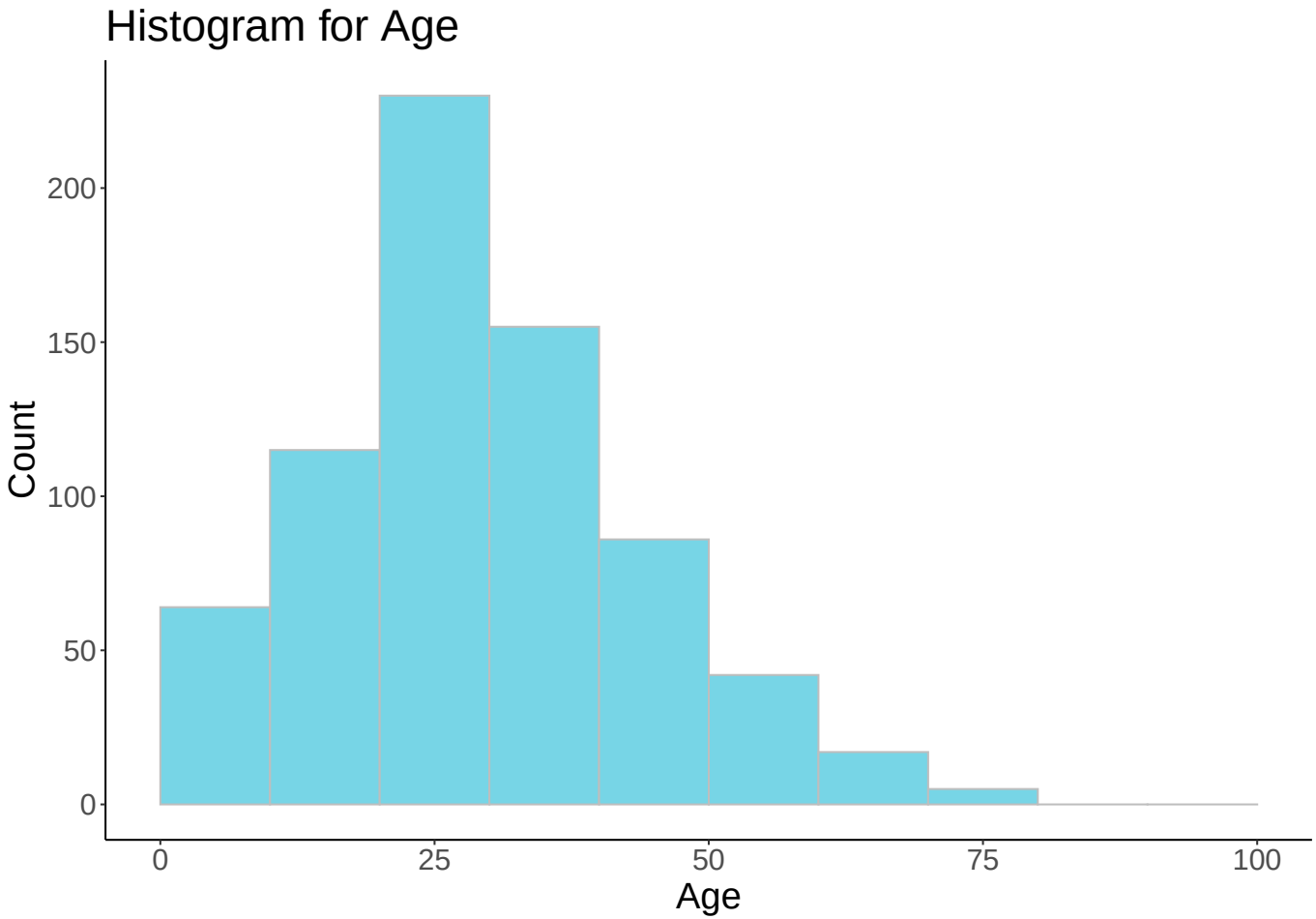


Figure 6: Histogram

Histogram (relative frequency)

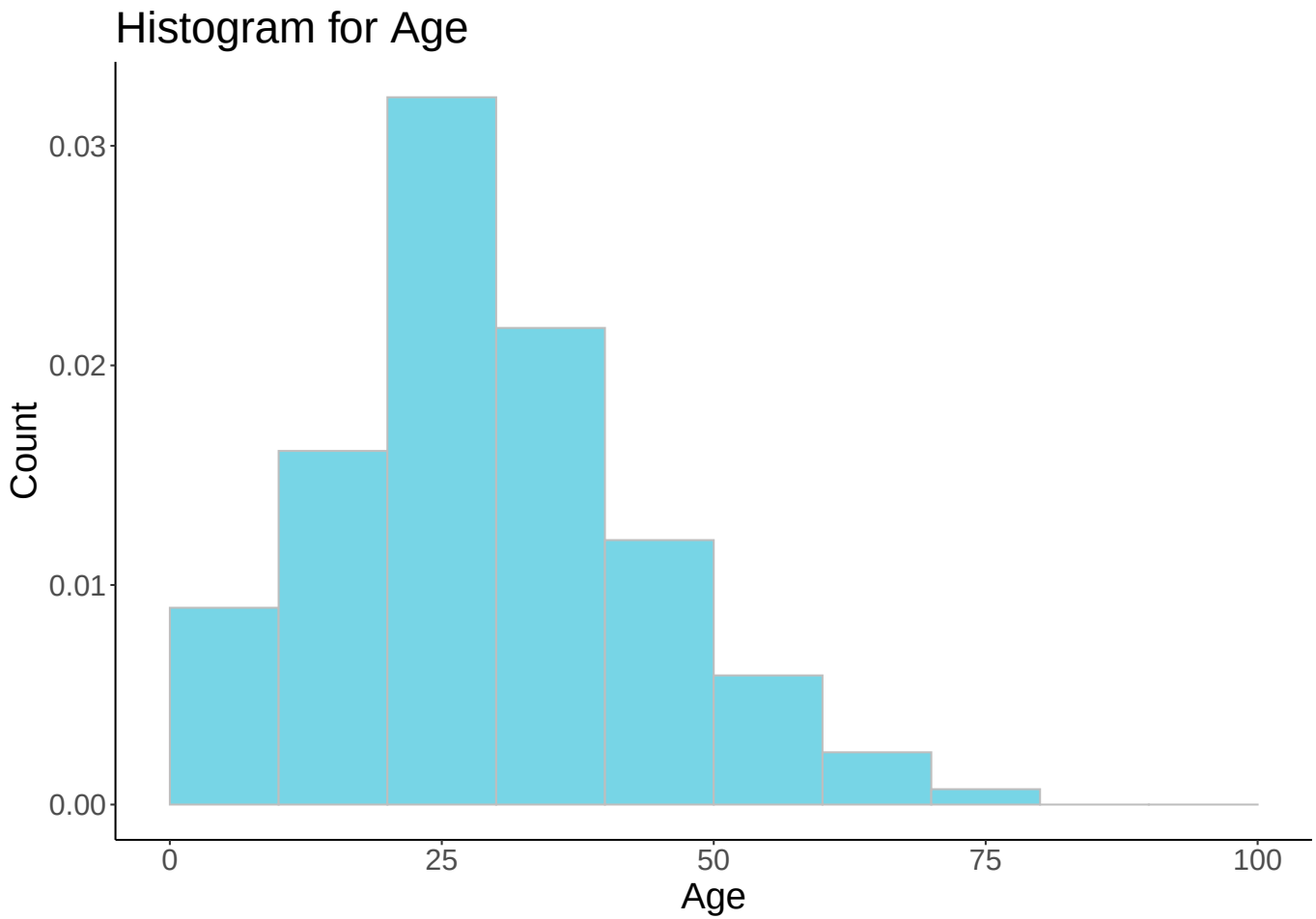


Figure 7: Histogram (relative frequency)

Histogram (relative frequency) (2)

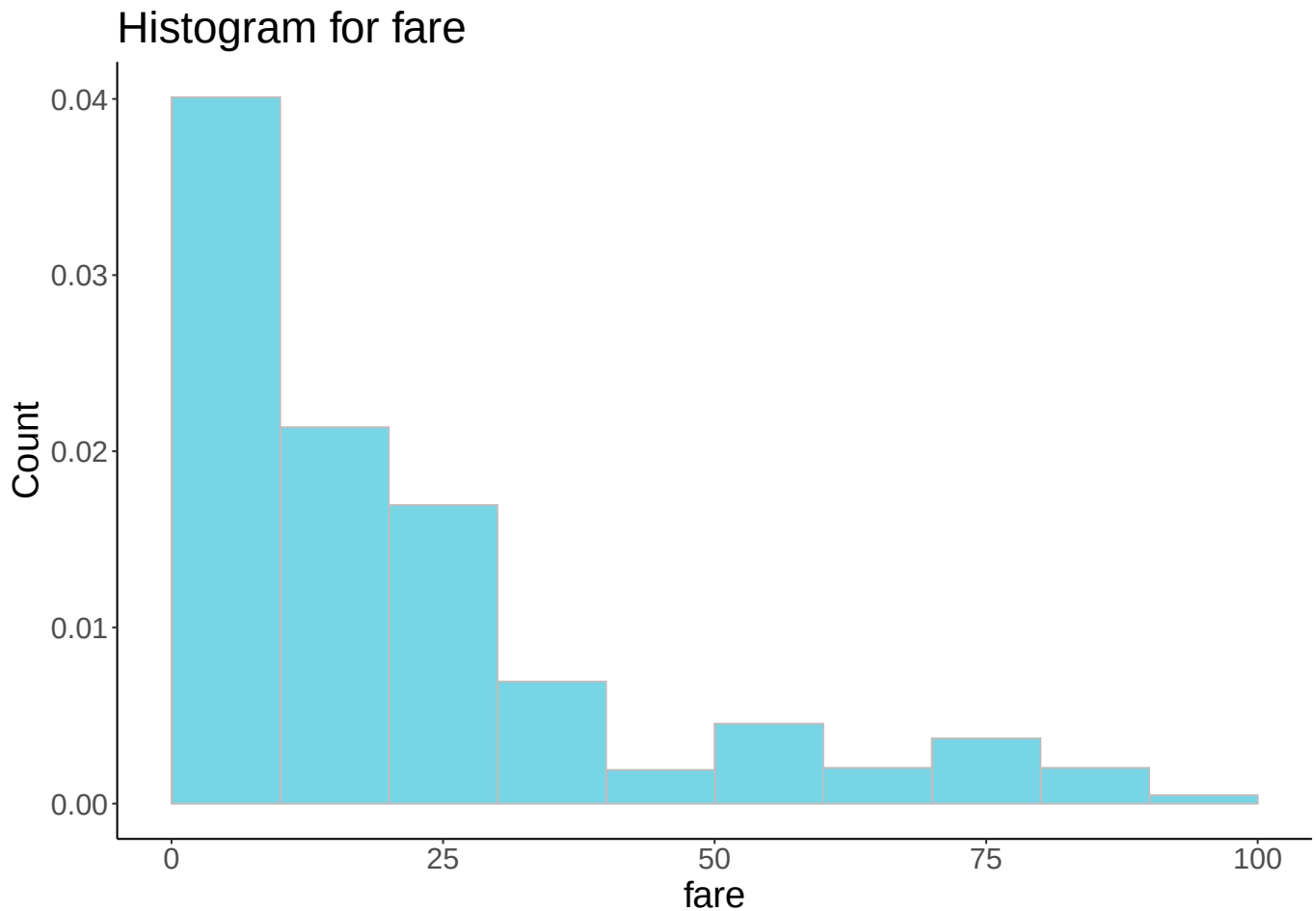


Figure 8: Histogram fare

Bin size calculation

Bin size b after *von Sturges*

$$b = \frac{V}{1 + 3.32 \cdot \lg n} \approx \frac{V}{5 \lg n}$$

n - the sample size (number of measurements)

V - the range of variation (span)

$\lg n$ - Decimal logarithm of n

Histogram (relative frequency): different number of bins

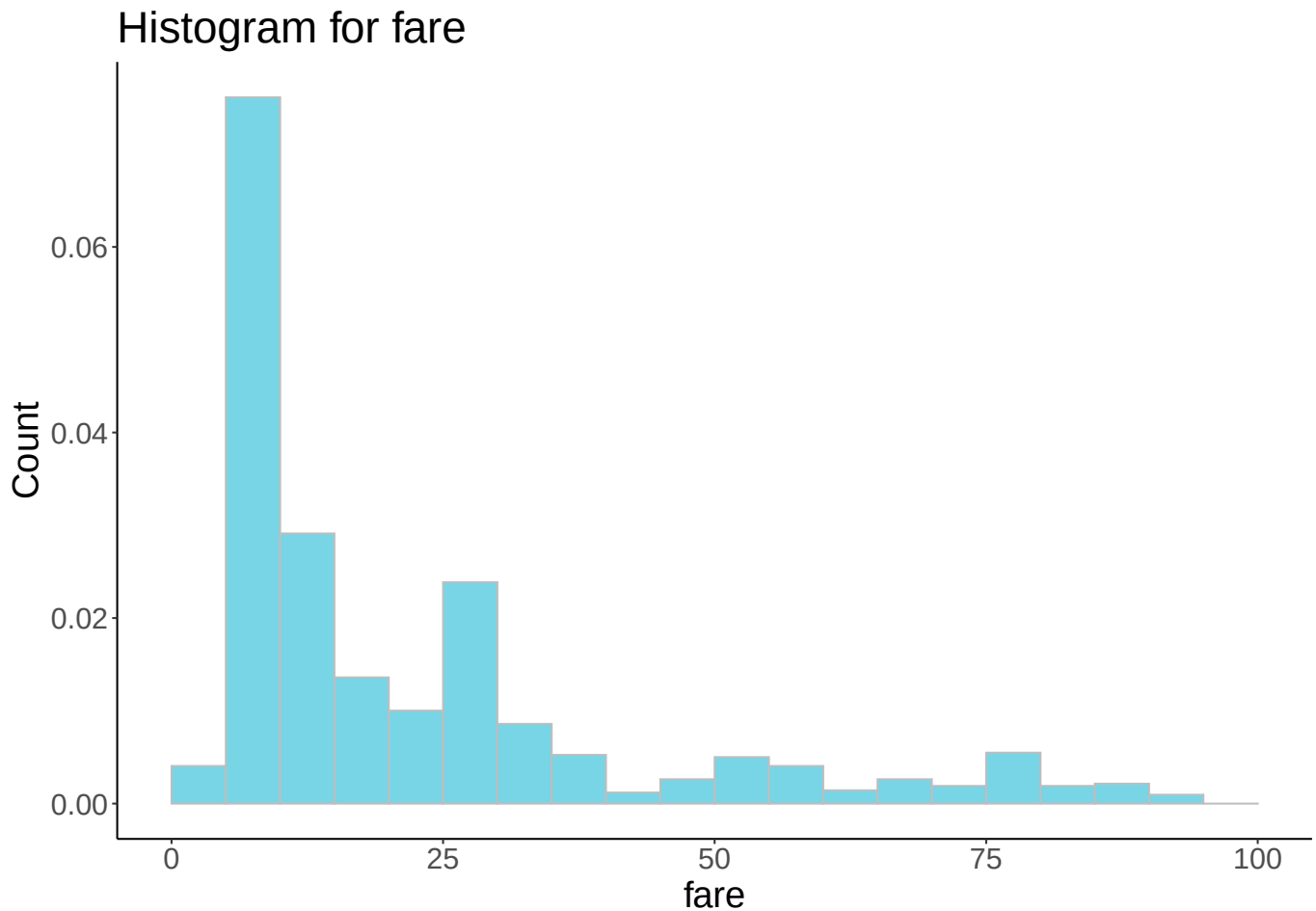


Figure 9: Histogram fare

Characteristic numbers (measures)

Characteristic numbers

Measures of central tendency

Dispersion parameters

Measures of central tendency

Variation range (span)

$$V_x = \max(x) - \min(x)$$

Arithmetic mean (average)

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

The average is very sensitive to extreme values

Median

Central value

Odd number of observations

$\left(\frac{n+1}{2}\right)$th observed (ordered) value

Even number of observations

Average of $\left(\frac{n}{2}\right)$th and $\left(\frac{n}{2}+1\right)$th observed (ordered) values.

The median is mainly determined by the values in the middle of the sample and is less dependent on the extreme values

Modus oder Modal score

The most frequent observed value. If all values occur the same number of times, the mode does not exist.

Geometrical Mean

$$G = \sqrt[n]{x_1 \cdot \ldots \cdot x_n} = e^{\frac{1}{n}\sum_{i=1}^{n} \ln x_i}$$

e.g. used to determine the MIC (minimal inhibitory concentration) $(2^k c,\ k = 1, 2, \ldots)$

Harmonic Mean

$$H = \frac{1}{n}\left(\frac{1}{x_1} + \ldots + \frac{1}{x_n}\right)^{-1}$$

Comparison of measures of central tendency

Histogram (relative frequency) (3)

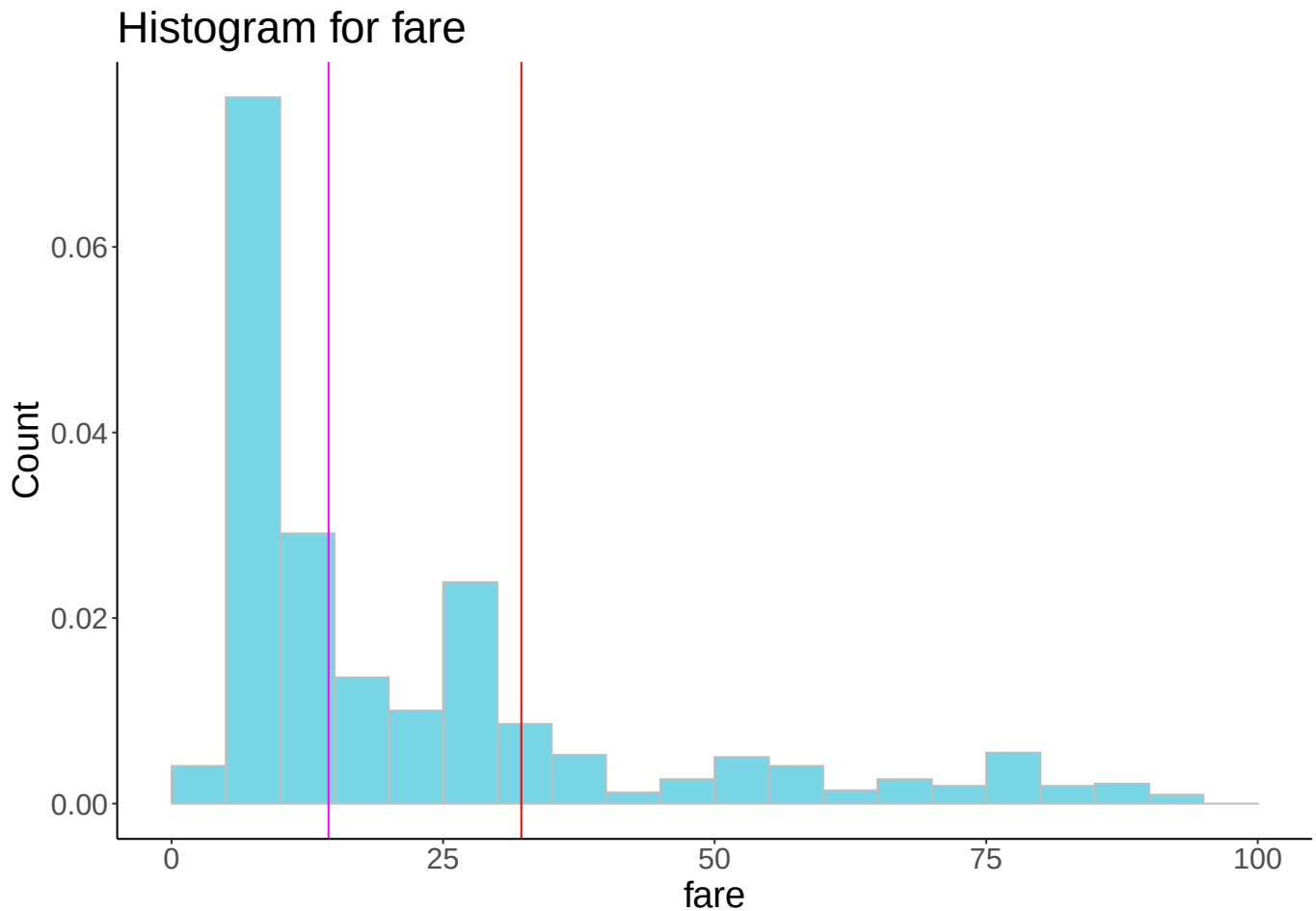


Figure 10: Histogram fare

Dispersion measures

Variance und Standard deviation

Variance

$$s_x^2 = \frac{1}{n-1} \sum_1^n (x_i - \bar{x})^2$$

$n - 1$ - *degrees of freedom*

Standard deviation

$$s_x = \sqrt{\frac{1}{n-1} \sum_1^n (x_i - \bar{x})^2}$$

Standard error of the mean (SEM)

$\bar{x}$ is scattered around the true mean μ of the population

$$s_{\bar{x}} = \frac{s}{\sqrt{n}}$$

Coefficient of variation

The ratio of the standard deviation to the mean

$$\text{CV} = \frac{s}{|\bar{x}|}$$

Allows you to compare the spread of data regardless of the mean.

Quantiles

- $(k+1)$th data point if $\frac{np}{100}$ non integer $(k < \frac{np}{100}, k \in \mathbb{N})$
- Average of kth and $(k+1)$th data points if $\frac{np}{100}$ is integer

1. Quantile (Q_1) - The data point where 25% data points are below and 75% above
2. Quantile (Q_3) - The data point where 75% data points are below and 25% above
3. Quantile (Q_2) - Median

Interquartile range and MAD

Interquartile range

$$IQR = Q_{0.75} - Q_{0.25}$$

Mean Abslolute Deviation (MAD)

$$|x_i - Q_{0.5}|$$

Cumulative distribution

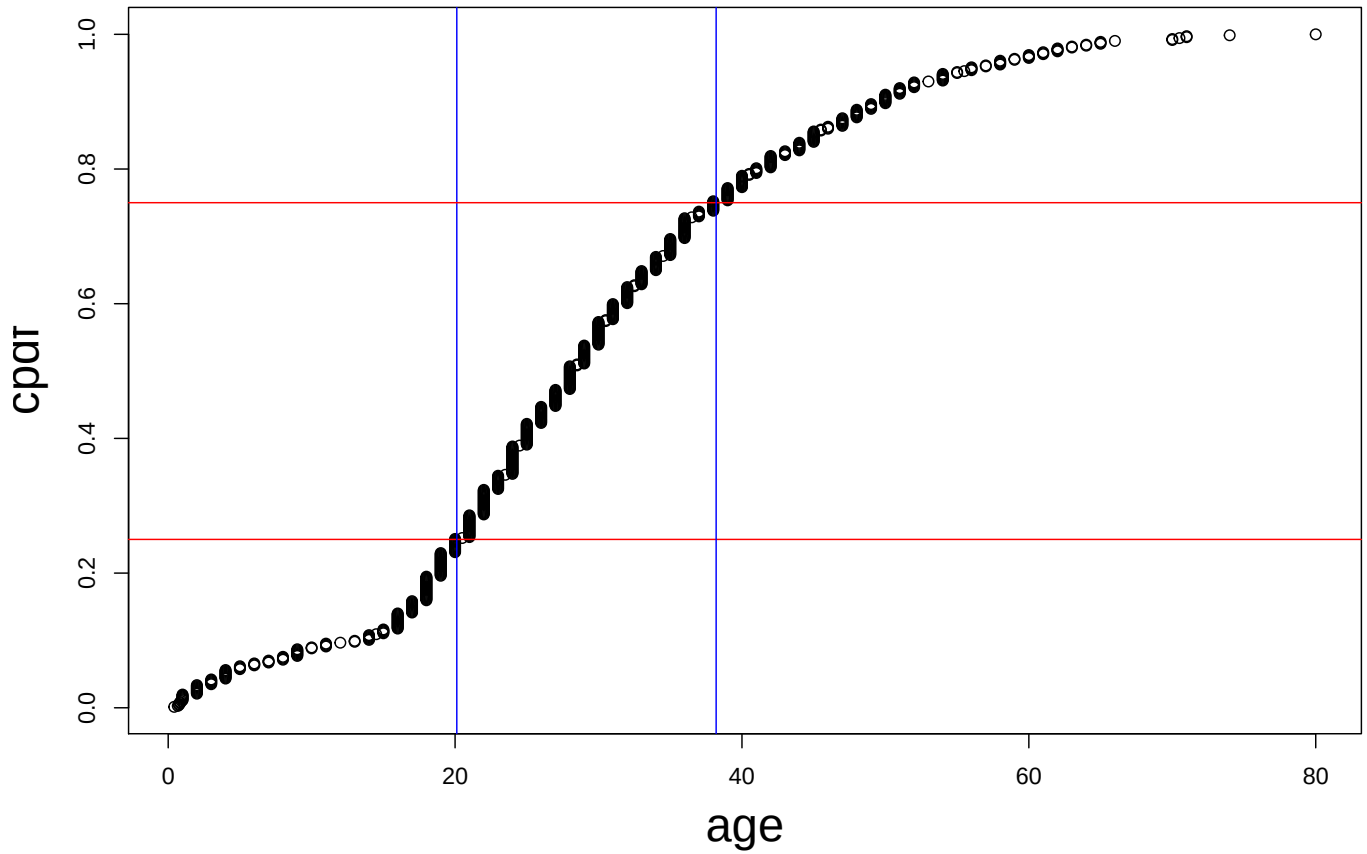
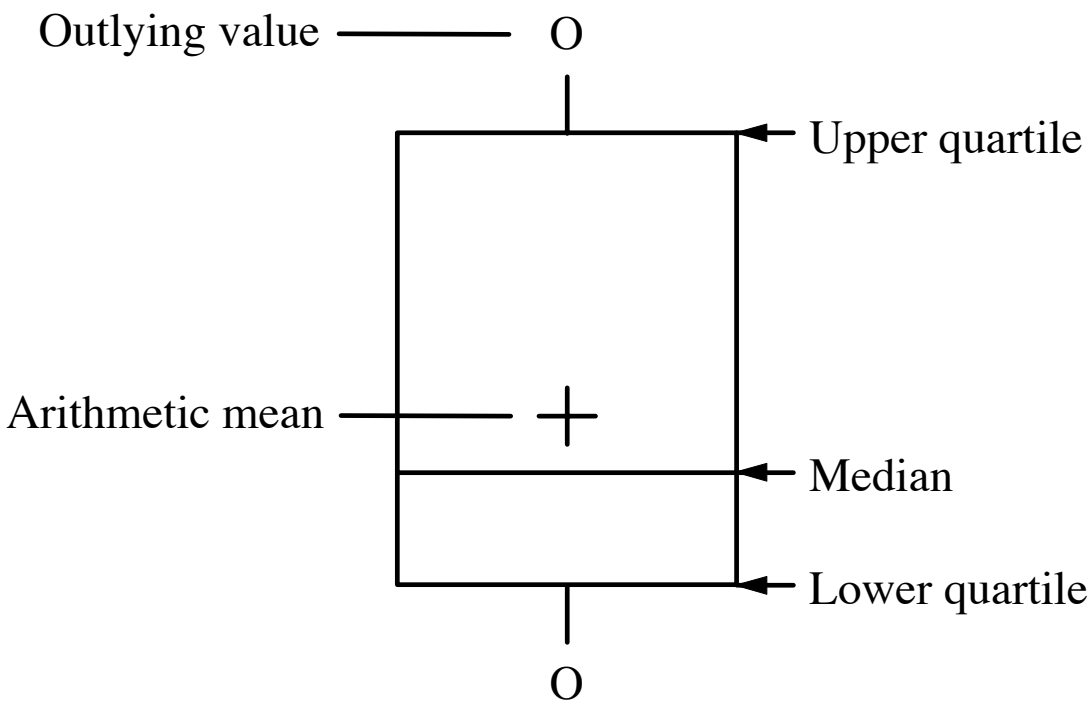


Figure 11: Cumulative distribution

Boxplot


Outlying value
O
Upper quartile
Arithmetic mean
Median
Lower quartile
O

Boxplot

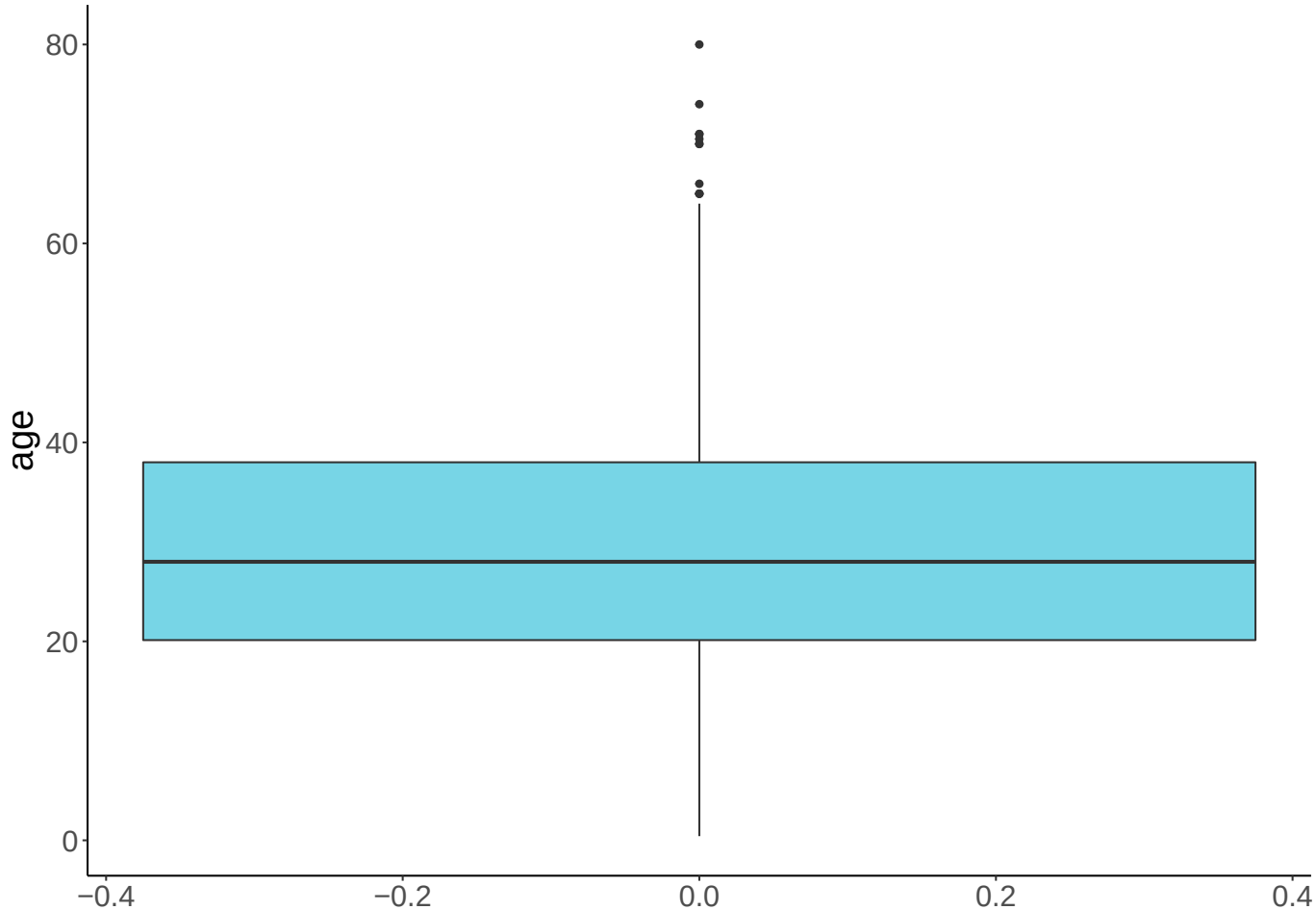


Figure 12: Boxplot

Boxplot

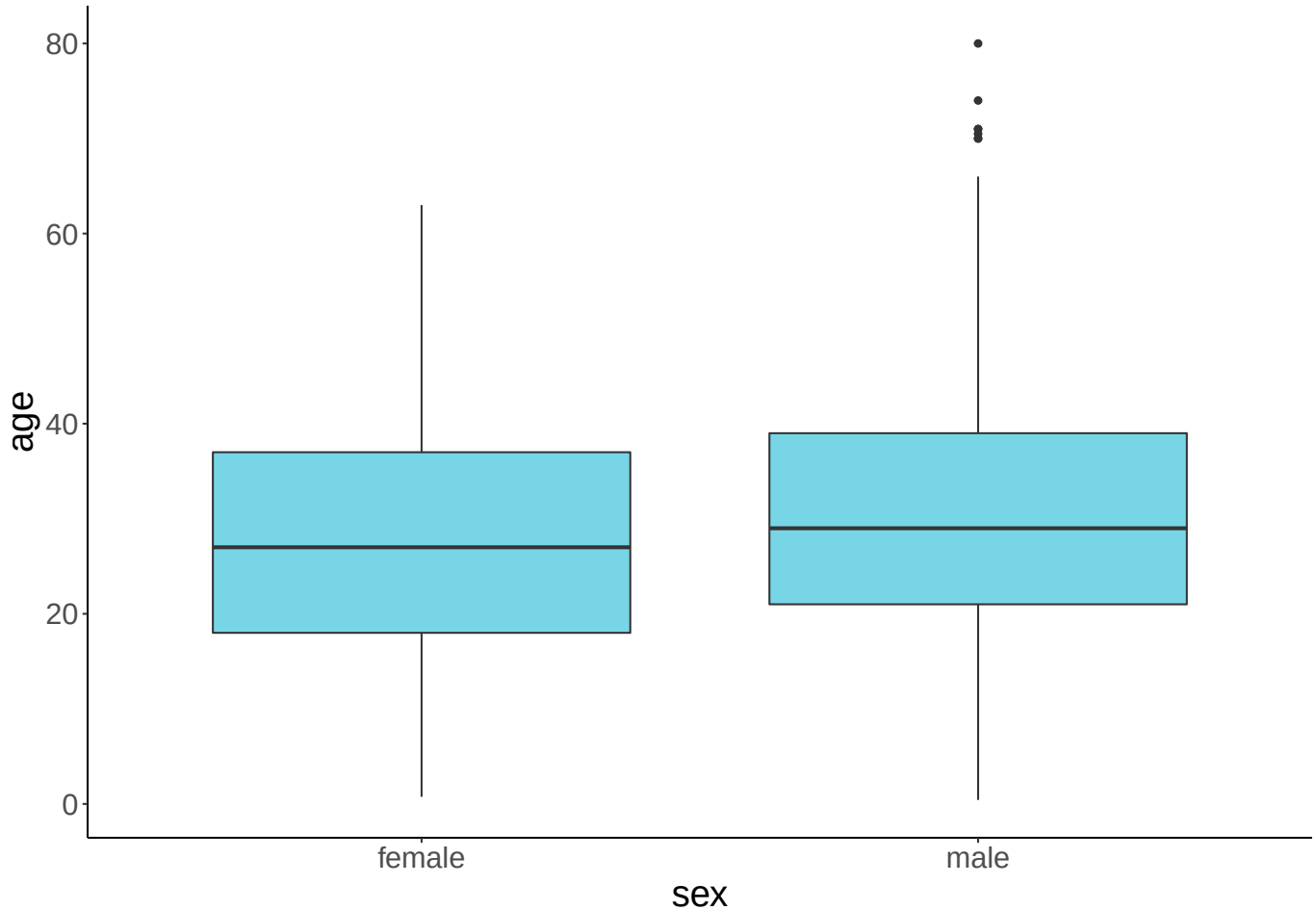


Figure 13: Boxplot

Diversity index (Shannon Entropy)

$$H = \sum_i p_i \log p_i$$

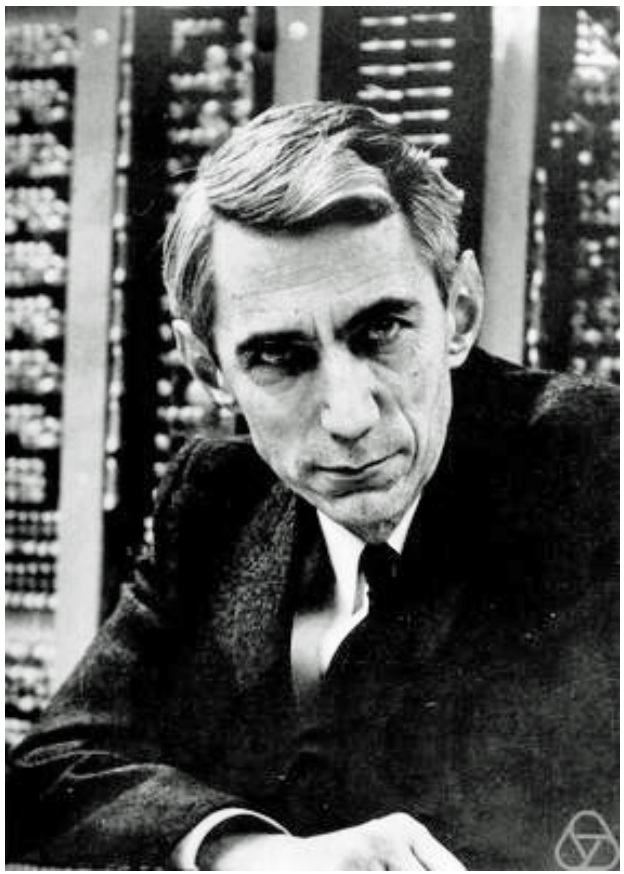
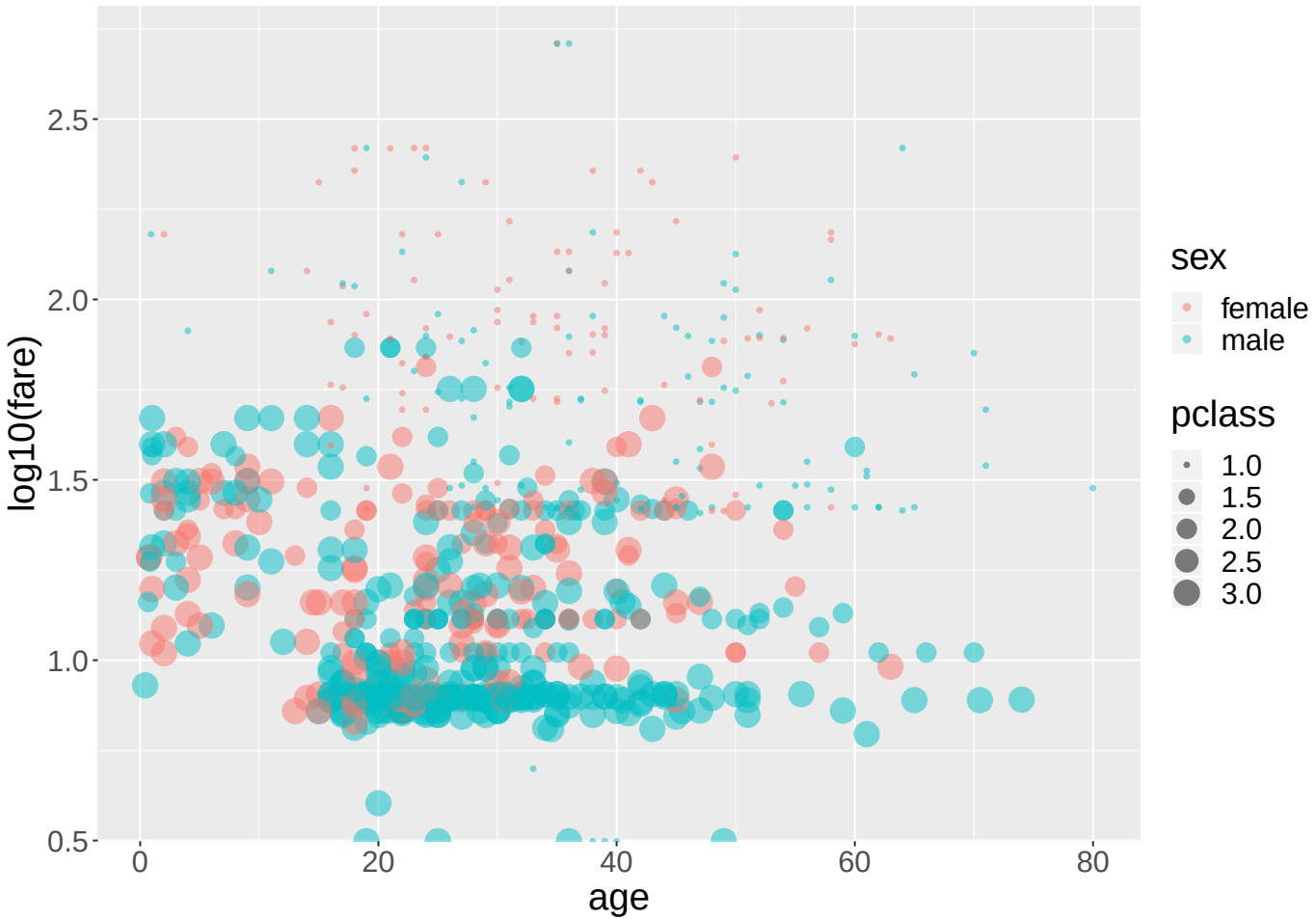


Figure 14: Claude Shannon (1910 — 2001)

Correlations in Data

Bivariate Data



Correlations in continuous data

- Pearson correlation coefficient
- Spearman rank correlation coefficient

Pearson correlation coefficient

Variance

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

Covariance

$$\mathrm{cov}(x, y) = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{n-1}$$

Pearson correlation coefficient

$$r = \frac{\operatorname{cov}(x, y)}{s_x s_y} = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{(n-1) s_x s_y}$$

Corrected z-score

$$z_r = \frac{1}{2} \ln \left(\frac{1+r}{1-r} \right)$$

Standard error of z-score

$$SE_{z_r} = \frac{1}{\sqrt{n-3}}$$

Confidence intervals of z-score and Pearson correlation coefficient

$$z_r \pm (1.96 \times SE_{z_r}), \text{ Transformation: } r = \frac{e^{2z_r} - 1}{e^{2z_r} + 1}$$

Pearson correlation coefficient (example)

```
#fig.show = 'hide'
#results='hide'
library(ggplot2)
df <- read.table('/Users/vitaly/Lehre/WS1920/Beuth/4_Statistische_Tests_and_Regression/Data/Data\ files/Album\ Sales\ 1.dat', s
cor(df)
```

```
##            adverts     sales
## adverts 1.0000000 0.5784877
## sales   0.5784877 1.0000000
```

```
cor.test(df$adverts,df$sales)
```

```
##
##  Pearson's product-moment correlation
##
## data:  df$adverts and df$sales
## t = 9.9793, df = 198, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.4781207 0.6639409
## sample estimates:
##       cor
## 0.5784877
```

Spearman (rank) correlation coefficient (example)

```r
cor(df, method = "spearman")
```

```
##           adverts     sales
## adverts 1.0000000 0.5541557
## sales   0.5541557 1.0000000
```

```r
cor.test(df$adverts, df$sales, alternative = "greater", method = "spearman")
```

```
##
##  Spearman's rank correlation rho
##
## data:  df$adverts and df$sales
## S = 594444, p-value < 2.2e-16
## alternative hypothesis: true rho is greater than 0
## sample estimates:
##       rho
## 0.5541557
```

Kendall-tau (example)

```r
cor(df, method = "kendall")
```

```
##           adverts     sales
## adverts 1.0000000 0.3985301
## sales   0.3985301 1.0000000
```

```r
cor.test(df$adverts, df$sales, alternative = "greater", method = "kendall")
```

```
##
##  Kendall's rank correlation tau
##
## data:  df$adverts and df$sales
## z = 8.2362, p-value < 2.2e-16
## alternative hypothesis: true tau is greater than 0
## sample estimates:
##       tau
## 0.3985301
```

Regression

Regression

$$Y = f(X_1, ..., X_n)$$

Regression is a broad term for a number of methods used to predict a *response* variable Y (or a *dependent*, *predicted*, *resulting*) from one or more *predictor* variables X_i (*independent*, *explanatory*).

- In the case of *one* predictor variable, one speaks of *simple* regression
- In the case of *several* predictor variables, one speaks of *multiple*regression

Regression goals

- *Determination* of the explanatory variables that refer to the response variable
- *Description* of the form of the relationship
- *To Provide* an equation for *predicting* the response variables from the explanatory variables No *causality* could be derived directly from the regression!

Regression, example (1)

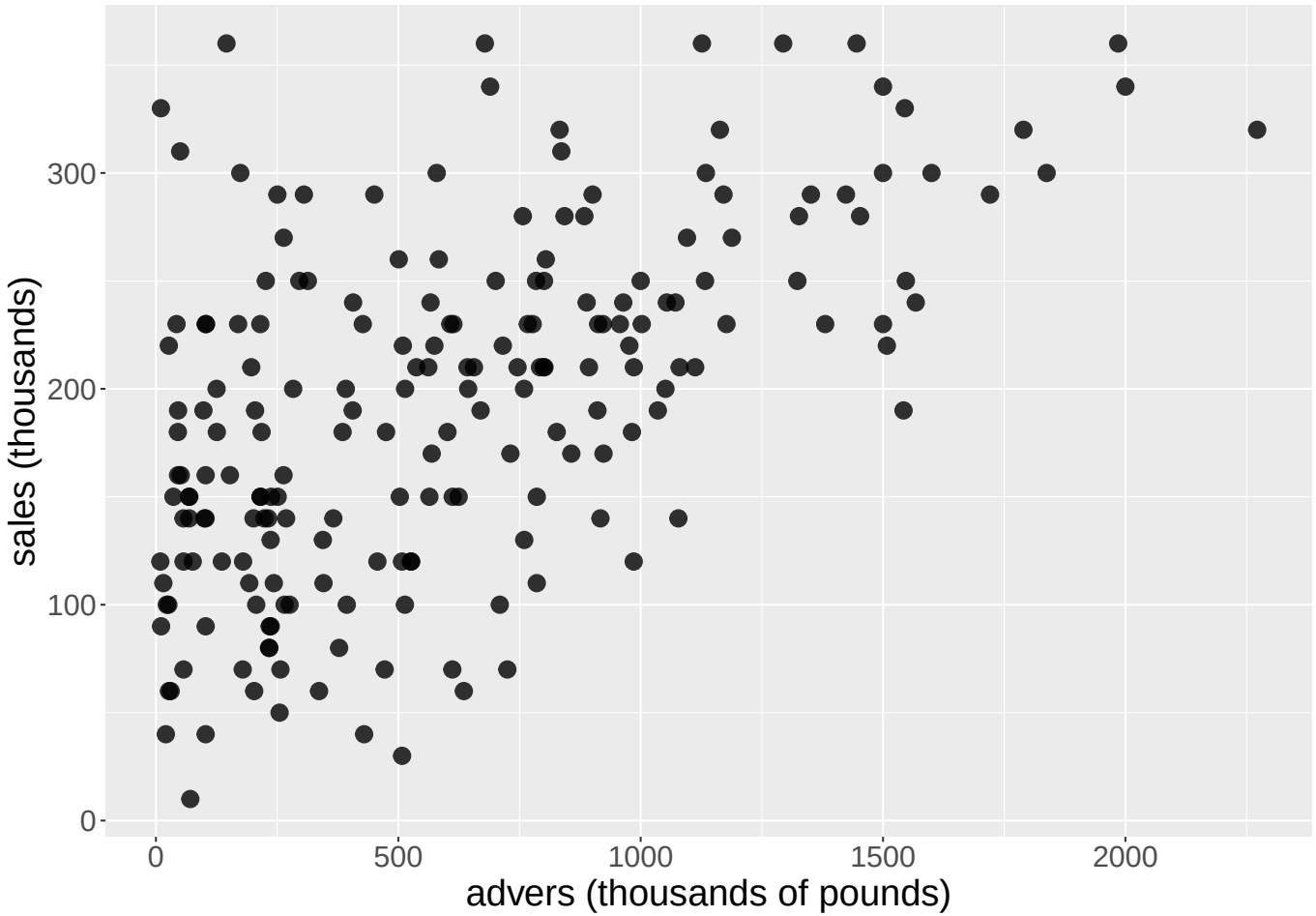


Figure 15: Impact of the amount of money spent on advertising in album sales. [A. Field et al.]

Regression, example (2)

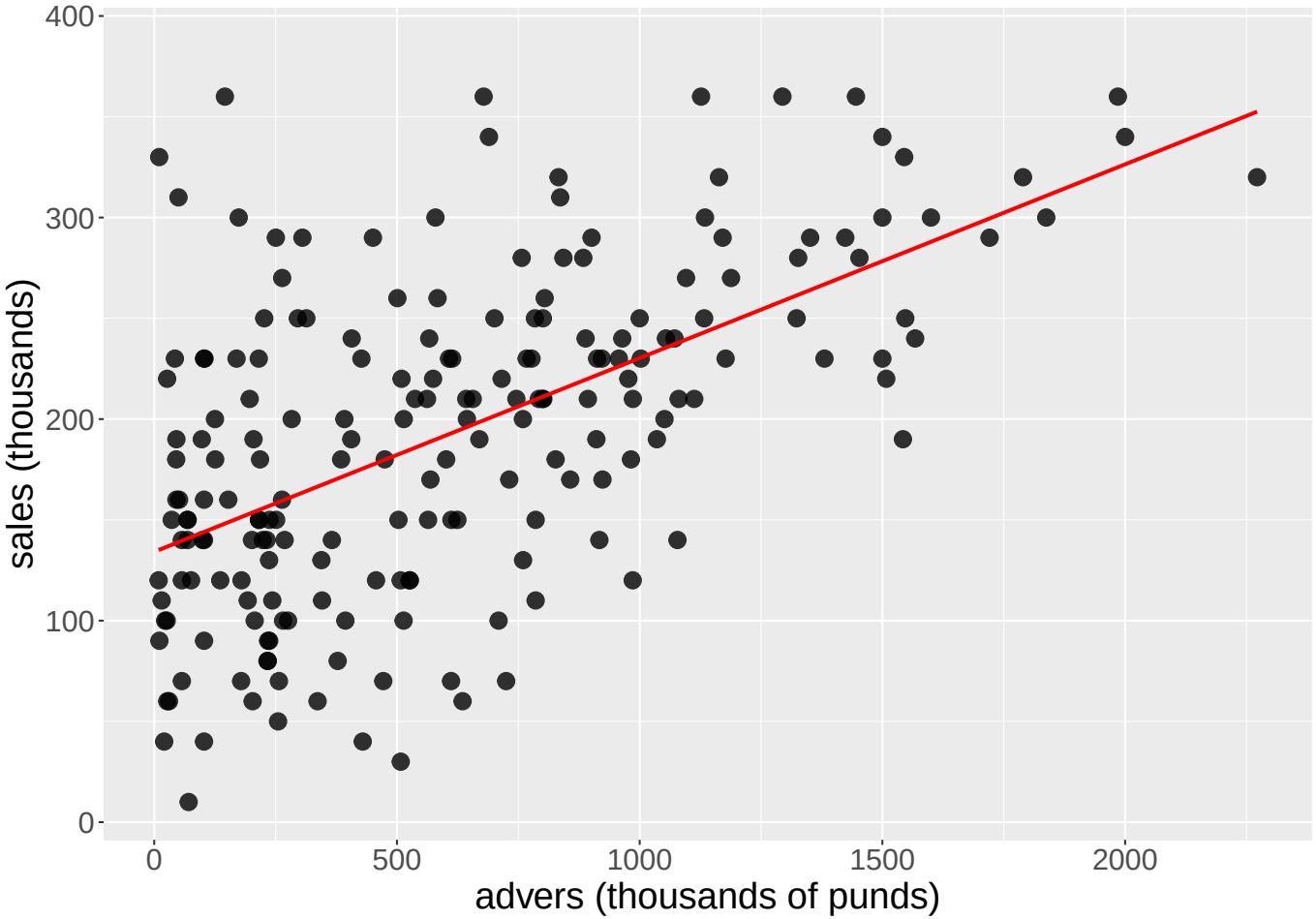


Figure 16: Impact of the amount of money spent on advertising in album sales. [A. Field et al.]

Regression, example (3)

```
library(ggplot2)
df <- read.table('/Users/vitaly/Lehre/WS1920/Beuth/4_Statistische_Tests_and_Regression/Data/Data\ files/Album\ Sales\ 1.dat', s
ggplot(df, aes(x = adverts, y = sales)) +
theme(text = element_text(size = 20)) +
geom_point(alpha = 0.8, size = 4) +
labs(x = "advers (thousands of punds)", y = "sales (thousands)") +
geom_smooth(fill=NA,color="red",size=1,method="lm")
```

Regression, example: model output

$$Y = \beta_0 + \beta_1 X + \varepsilon$$

```
model <- lm(sales~adverts, data = df)
summary(model)
```

```
##
## Call:
## lm(formula = sales ~ adverts, data = df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -152.949  -43.796   -0.393   37.040  211.866
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 1.341e+02  7.537e+00  17.799   <2e-16 ***
## adverts     9.612e-02  9.632e-03   9.979   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 65.99 on 198 degrees of freedom
## Multiple R-squared:  0.3346, Adjusted R-squared:  0.3313
## F-statistic: 99.59 on 1 and 198 DF,  p-value: < 2.2e-16
```

Regression, example: ANOVA-Table

```
anova(model)
```

```
## Analysis of Variance Table
##
## Response: sales
##            Df Sum Sq Mean Sq F value    Pr(>F)
## adverts     1 433688  433688  99.587 < 2.2e-16 ***
## Residuals 198 862264    4355
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

ANOVA-Table

$$\hat{Y} = \beta_0 + \beta_1 X$$

To assess how much information the variable X contains about the variable Y, we consider how much of the *sum of squares* (SS) of the variable Y could be explained by the knowledge of the variable X.

Table 3: Decomposition of sums of squares in a regression model with k predictor variables. [Dohoo, p. 327]

Source of variation	Sum of squares	Degrees of freedom	Mean square	F-test
Model	$SSM = \sum_{i=1}^{n} \left(\hat{Y}_i - \bar{Y}\right)^2$	$dfM = k$	$MSM = \frac{SSM}{dfM}$	$\frac{MSM}{MSE}$
Error	$SSE = \sum_{i=1}^{n} \left(Y_i - \hat{Y}_i\right)^2$	$dE = n - (k+1)$	$MSE = \frac{SSE}{dfE}$	
Total	$SST = \sum_{i=1}^{n} \left(Y_i - \bar{Y}\right)^2$	$dfT = n - 1$	$MST = \frac{SST}{dfT}$	

$MSE = \sigma^2$ and σ is called *standard error of prediction*.

F-Test for Model quality assessement

H_0: $\beta_1 = \beta_2 ... \beta_k = 0$, $\beta_0 \neq 0$

H_1: at least some of coefficients $\beta_i \neq 0$, $i \neq 0$

- Variables are selected to maximize F statistics
- F - test has a simple meaning when manipulating independent variables in a controlled experiment
- Be careful with *observation variables* (influenced by the number of variables, their correlations, and the sample size)

t-test (with $n - (k + 1)$ degrees of freedom) (1)*

H_0: $\beta_i = \beta^* \; e.g. \beta^* = 0$

H_1: $\beta_i \neq \beta^* \; e.g. \beta^* = 0$

$$t = \frac{\beta_i - \beta^*}{SE(\beta_i)}$$

$SE(\beta_i)$ is standard error of the esimated coefficient

In case of a single predictor

$$SE(\beta_1) = \sqrt{\frac{MSE}{SSX(\beta_1)}},$$

wobei $SSX_1 = \sum_{i=1}^{n} \left(X_{1,i} - \bar{X_1}\right)^2$ is sum of squares of variable X_1.

t-test (with $n - (k + 1)$ degrees of freedom) (2)

- Variables are selected so that the t statistics are maximized and their significance is guaranteed
- t -test has a simple meaning only if independent variables are manipulated in a controlled experiment
- Caution should be exercised when using observation variables (influenced by the number of variables, their correlations and the sample size).

Prediction intervals (1)

Deviations from forecast estimates:

- due to the estimation of the regression parameters (SE)
- due to the uncertainty associated with a new observation x^* (variation over the regression equation for the mean)

Error of the mean

For simple regression (single predictor) *error of the mean* of a large number of new observations

$$SE_{\text{mean}}(Y|x^*) = \sigma\sqrt{\frac{1}{n} + \frac{\left(x^* - \bar{X}\right)^2}{SSX}}$$

Prediction intervals (1)

Standard error of a new observation

For simple regression

$$SE_{\text{obs}}(Y|x^*) = \sigma\sqrt{1 + \frac{1}{n} + \frac{\left(x^* - \bar{X}\right)^2}{SSX}}$$

95% Confidence interval

$$95\%CI = Y \pm t_{0.05}(SE)$$

Coefficient of determination, R^2

- R^2 describes the amount of variance in the result variable that is explained by the predictor variables.
- R^2 is a squared *correlation coefficient* between the predicted and observed Y values

$$R^2 = \frac{SSM}{SST} = 1 - \frac{SSE}{SST}$$

Adjusted R^2

Because R^2 always increases with the number of variables, *adjusted* R^2 is used

$$R^2_{\text{adjusted}} = 1 - \frac{MSE}{MST}$$

Akaike Information Criterion (AIC)

When estimating a model's loss of information, AIC takes into account the compromise between the goodness of fit of the model and the simplicity of the model. In other words, AIC deals with both the risk of overfitting and the risk of underfitting.

$$AIC = n \ln\left(\frac{SSE}{n}\right) + 2k$$

n - number of observations

k - number of predictor variables

One should prefer a model with a smaller AIC!

```
model1 <- lm(milk120~cc, data = daisy2)
AIC(model1)
```

```
## [1] 112227.5
```

```
model2 <- lm(milk120~parity, data = daisy2)
AIC(model2)
```

```
## [1] 104745.6
```

Simple Regression, example

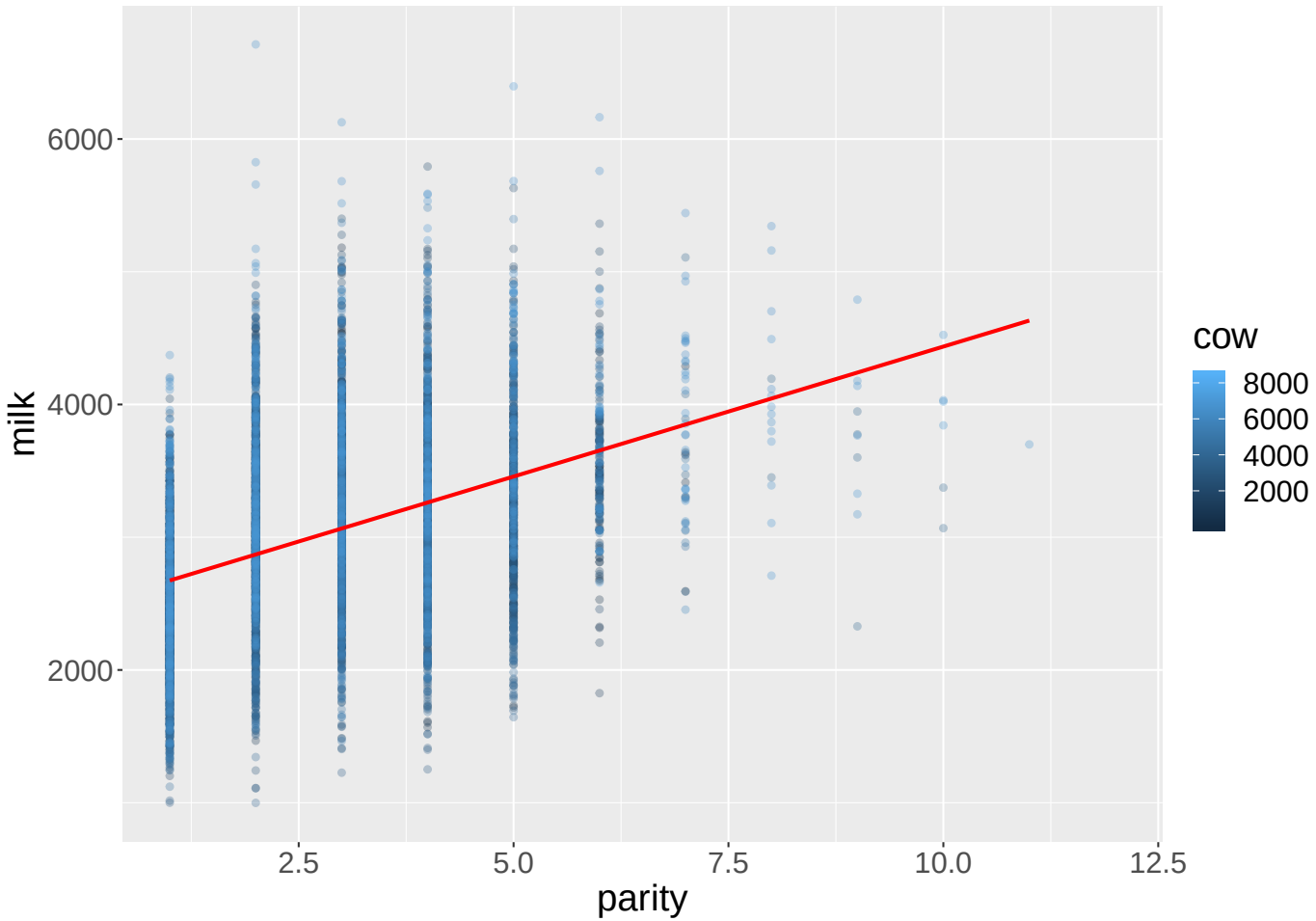


Figure 17: Impact of parity on Milk volume. 9383 lactation records in 42 year-round calving herds. http://projects.upei.ca/ver/data-and-samples/.

Simple Regression, example: model output

$$Y = \beta_0 + \beta_1 X + \varepsilon$$

```
model <- lm(milk120~parity, data = daisy2)
summary(model)
```

```
##
## Call:
## lm(formula = milk120 ~ parity, data = daisy2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2010.3  -454.6   -36.3   413.5  3842.9
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 2478.286     17.176  144.28   <2e-16 ***
## parity       195.807      5.385   36.36   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 667.8 on 6608 degrees of freedom
##   (2773 observations deleted due to missingness)
## Multiple R-squared:  0.1667, Adjusted R-squared:  0.1666
## F-statistic:  1322 on 1 and 6608 DF,  p-value: < 2.2e-16
```

Simple Regression, example: ANOVA-Table

```
anova(model)
```

```
## Analysis of Variance Table
##
## Response: milk120
##             Df     Sum Sq   Mean Sq F value    Pr(>F)
## parity       1  589631568 589631568  1322.1 < 2.2e-16 ***
## Residuals 6608 2947090043    445988
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Simple Regression, example: Confidence Intervals (1)

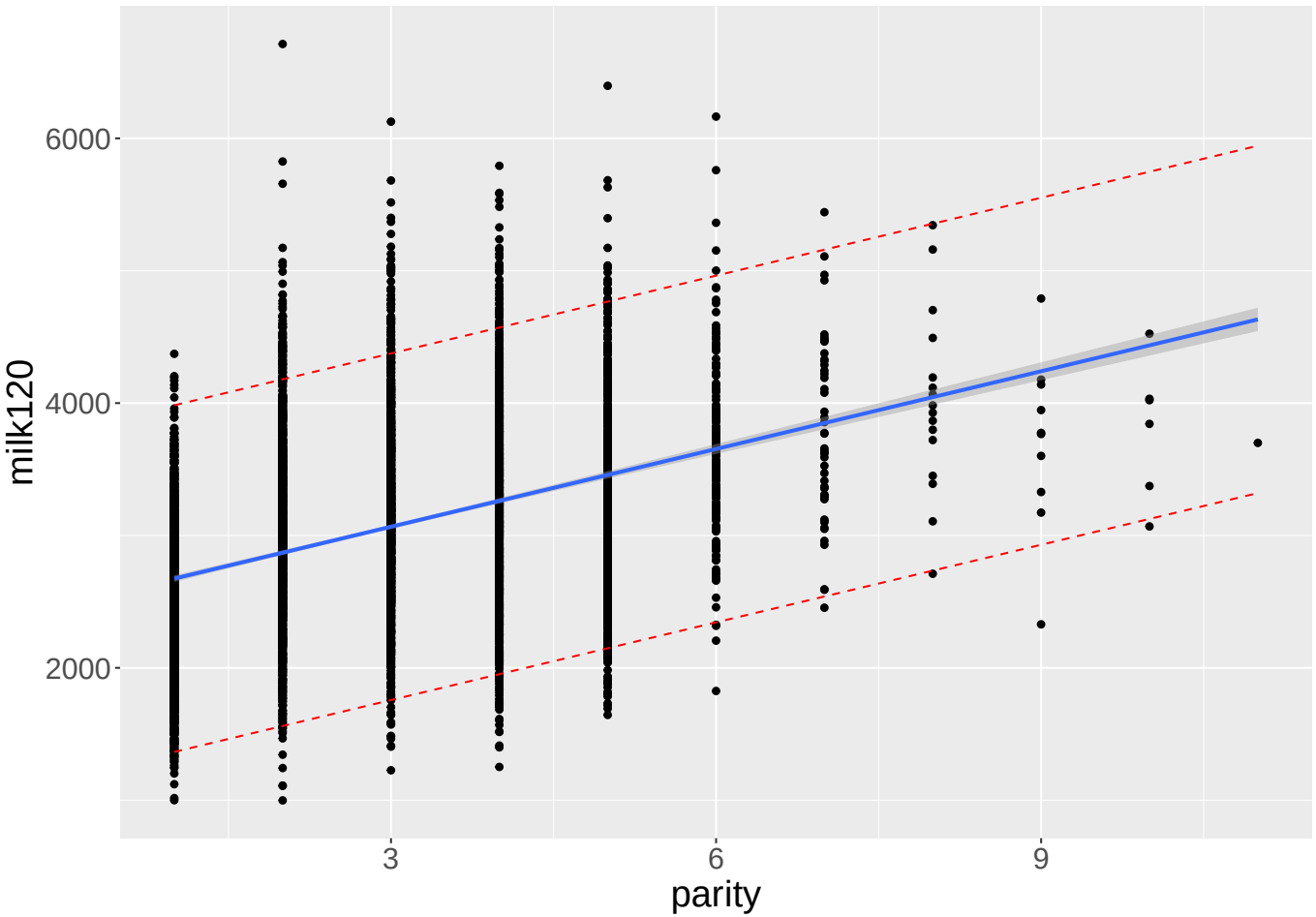


Figure 18: Prediction interval for the regression milk volume on parity.

Simple Regression, example: Confidence Intervals (2)

```
temp_var <- predict(model, interval="prediction")
new_df0 <- na.omit(data.frame("milk120" = daisy2$milk120, "parity" = daisy2$parity))
new_df <- cbind(new_df0, temp_var)
ggplot(new_df, aes(x=parity, y=milk120))+
    geom_point() +
    geom_line(aes(y=lwr), color = "red", linetype = "dashed")+
    geom_line(aes(y=upr), color = "red", linetype = "dashed")+
    geom_smooth(method=lm, se=TRUE)
```

ANOVA

ANOVA

If the response variable (*response*) is *continuous* and the independent variables categorical (*factors*), it makes sense to consider the differences between the different groups of factor levels or factor gradations.

The statistical techniques for comparing the mean values of several groups are called **ANOVA** (Analysis of Variance).

ANOVA can and should be viewed as a *GLM* (generalized *regression*).

Terminology

- Simple ANOVA: an explanatory (*categorical*) variable (factor). Comparisons between levels by this factor is performed.
- Two way (factorial) ANOVA: two explanatory (*categorical*) variables (factor). Comparisons between levels of these factors is performed.
- Multiple ANOAVA – **MANOVA**: several independent variables.
- ANCOVA (Analysis of covariances): in addition to the ANOVA situation, there is also a covariate (variable) for the dependent variable.
- Repeated measures: the measurements are carried out *several times* for the *same objects*.

Einfache ANOVA

ANOVA example

```
df <- read.table('/Users/vitaly/Lehre/WS1920/Beuth/Data files/Viagra.dat', sep = '', header = TRUE)
knitr::kable(df, caption = "Viagra-Datensatz [A. Field]")
```

Table 4: Viagra-Datensatz [A. Field]

person	dose	libido
1	1	3
2	1	2
3	1	1
4	1	1
5	1	4
6	2	5
7	2	2
8	2	4
9	2	2
10	2	3
11	3	7
12	3	4
13	3	5
14	3	3
15	3	6

ANOVA example (1)

```
library(gplots)
  plotmeans(df$libido ~ df$dose)
```

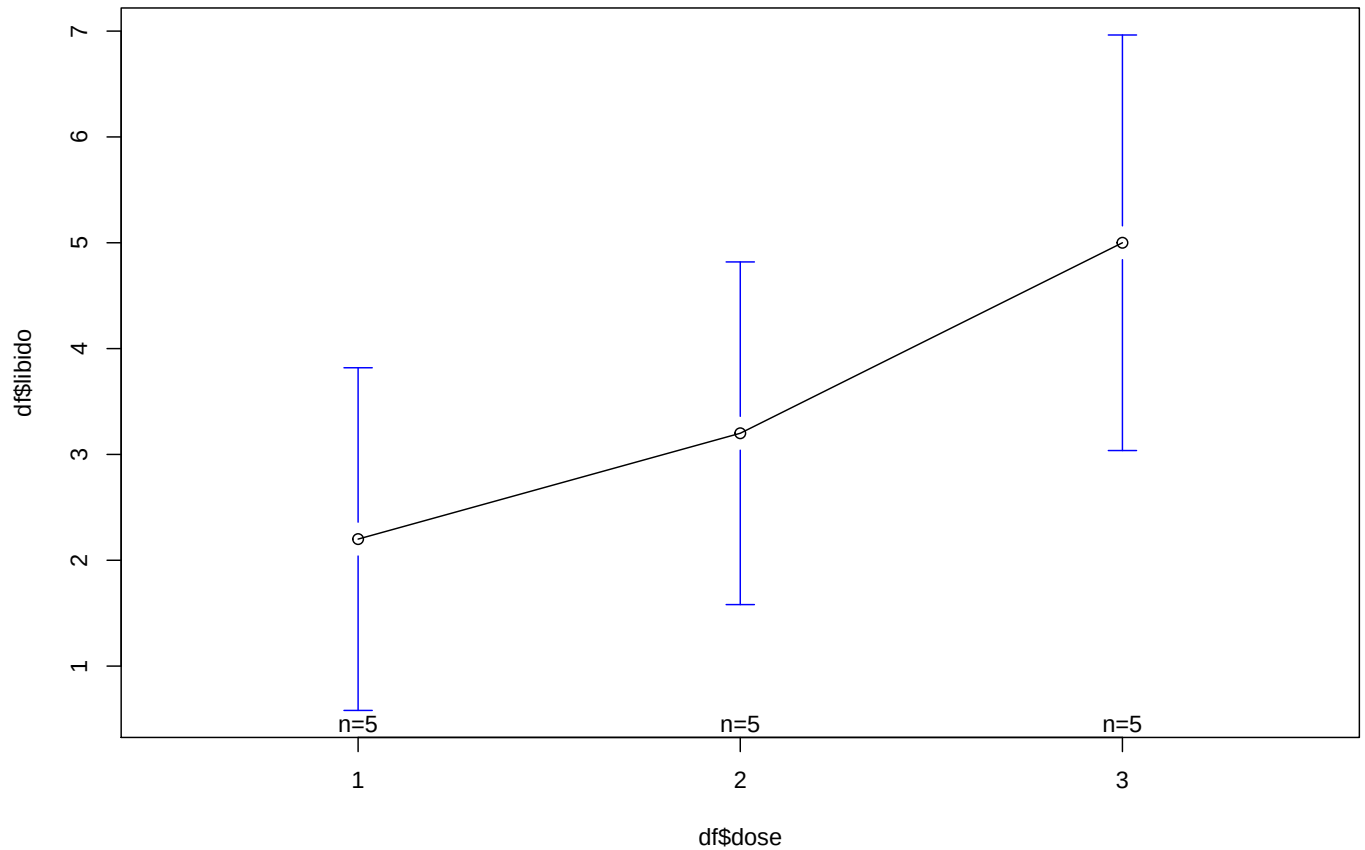


Figure 19: Viagra dataset with confidence intervals

ANOVA example (2)

```
library(pastecs)
by(df$libido, df$dose, stat.desc)
```

```
## df$dose: 1
##      nbr.val     nbr.null       nbr.na          min          max        range
##    5.0000000    0.0000000    0.0000000    1.0000000    4.0000000    3.0000000
##          sum       median         mean      SE.mean CI.mean.0.95          var
##   11.0000000    2.0000000    2.2000000    0.5830952    1.6189318    1.7000000
##      std.dev     coef.var
##    1.3038405    0.5926548
## ------------------------------------------------------------
## df$dose: 2
##      nbr.val     nbr.null       nbr.na          min          max        range
##    5.0000000    0.0000000    0.0000000    2.0000000    5.0000000    3.0000000
##          sum       median         mean      SE.mean CI.mean.0.95          var
##   16.0000000    3.0000000    3.2000000    0.5830952    1.6189318    1.7000000
##      std.dev     coef.var
##    1.3038405    0.4074502
## ------------------------------------------------------------
## df$dose: 3
##      nbr.val     nbr.null       nbr.na          min          max        range
##    5.0000000    0.0000000    0.0000000    3.0000000    7.0000000    4.0000000
##          sum       median         mean      SE.mean CI.mean.0.95          var
##   25.0000000    5.0000000    5.0000000    0.7071068    1.9632432    2.5000000
##      std.dev     coef.var
##    1.5811388    0.3162278
```

ANOVA example (Levene's test)

```
library(pastecs)
library(car)
leveneTest(df$libido, df$dose, center = median)
```

```
## Levene's Test for Homogeneity of Variance (center = median)
##       Df F value Pr(>F)
## group  2  0.1176   0.89
##       12
```

If the *Levene's test* is non-significant, the *data variances* are very homogenous, i.e. similar.

If not significant, you can use *Welch's F test* or *robust* ANOVA.

ANOVA example (2)

$$\text{libido}_i = \text{dose}_i + \text{error}$$

ANOVA example (*lm*)

```r
df <- read.table('/Users/vitaly/Lehre/WS1920/Beuth/Data files/Viagra.dat', sep = '', header = TRUE)
df$dose <- as.factor(df$dose)
library(ggplot2)
viagraModel <- lm(libido ~ dose, data = df)
summary(viagraModel)
```

```
##
## Call:
## lm(formula = libido ~ dose, data = df)
##
## Residuals:
##    Min     1Q Median     3Q    Max
##   -2.0   -1.2   -0.2    0.9    2.0
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   2.2000     0.6272   3.508  0.00432 **
## dose2         1.0000     0.8869   1.127  0.28158
## dose3         2.8000     0.8869   3.157  0.00827 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.402 on 12 degrees of freedom
## Multiple R-squared:  0.4604, Adjusted R-squared:  0.3704
## F-statistic: 5.119 on 2 and 12 DF,  p-value: 0.02469
```

ANOVA example (*aov*)

```
ViagraModel <- aov(libido ~ dose, data = df)
summary(ViagraModel)
```

```
##             Df Sum Sq Mean Sq F value Pr(>F)
## dose         2  20.13  10.067   5.119 0.0247 *
## Residuals   12  23.60   1.967
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

F-Test checks whether there are differences between the groups.

F-Test does not tell us between which groups the differences exist.

ANOVA example (*aov* 1)

```
library(car)
qqPlot(lm(libido ~ dose, data = df), simulate = TRUE, main = 'Q-Q Plot', labels = FALSE)
```

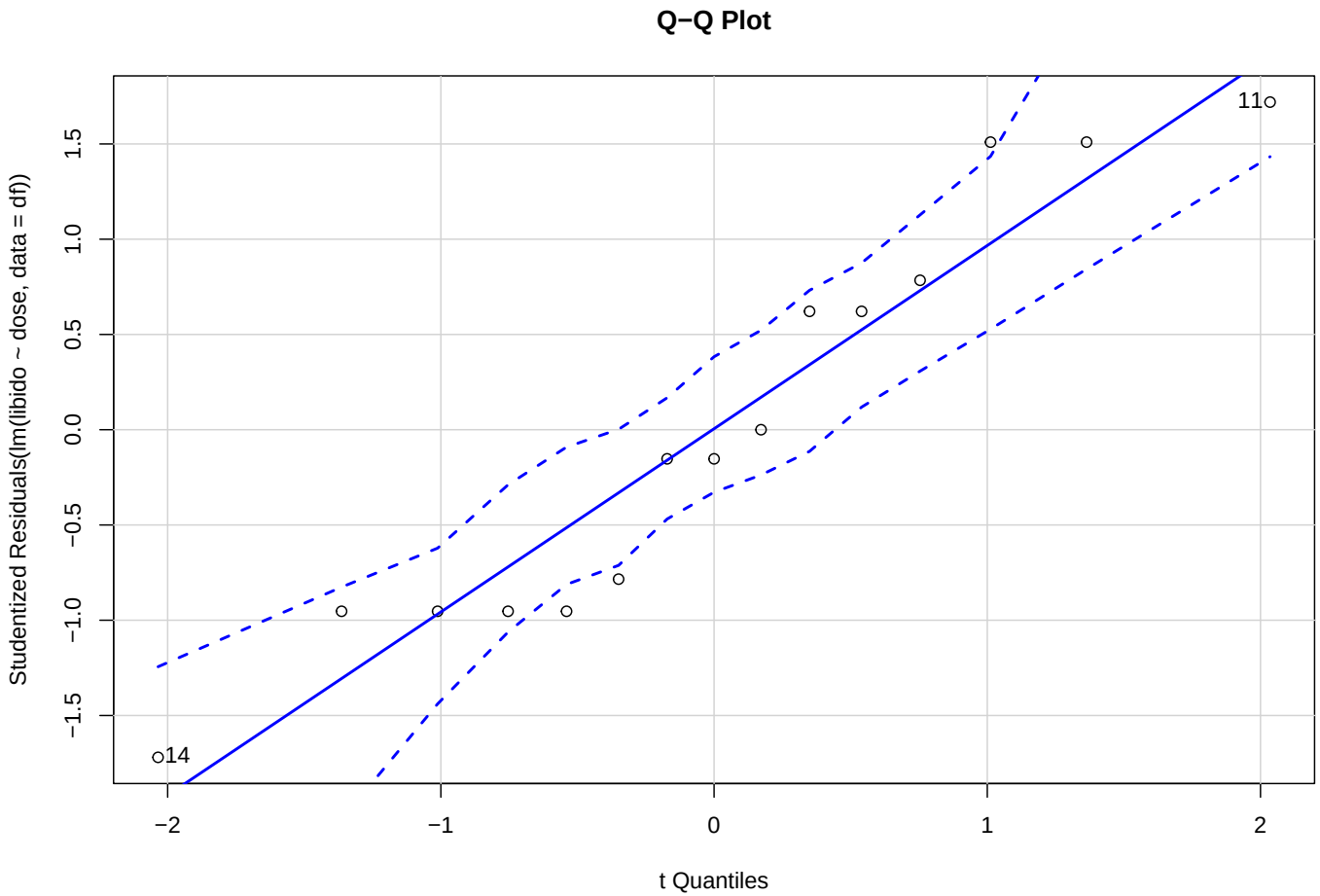


Figure 20: Q-Q plot.

```
## [1] 11 14
```

ANOVA example: Bartlett-Test

```
bartlett.test(libido ~ dose, data = df)
```

```
##
##  Bartlett test of homogeneity of variances
##
## data:  libido by dose
## Bartlett's K-squared = 0.1853, df = 2, p-value = 0.9115
```

ANOVA example: outliers

```
outlierTest(ViagraModel)
```

```
## No Studentized residuals with Bonferroni p < 0.05
## Largest |rstudent|:
##    rstudent unadjusted p-value Bonferroni p
## 11  1.71959            0.11348           NA
```

ANOVA example (*aov* 3)

```
par(mfrow=c(2,2))
plot(ViagraModel)
```

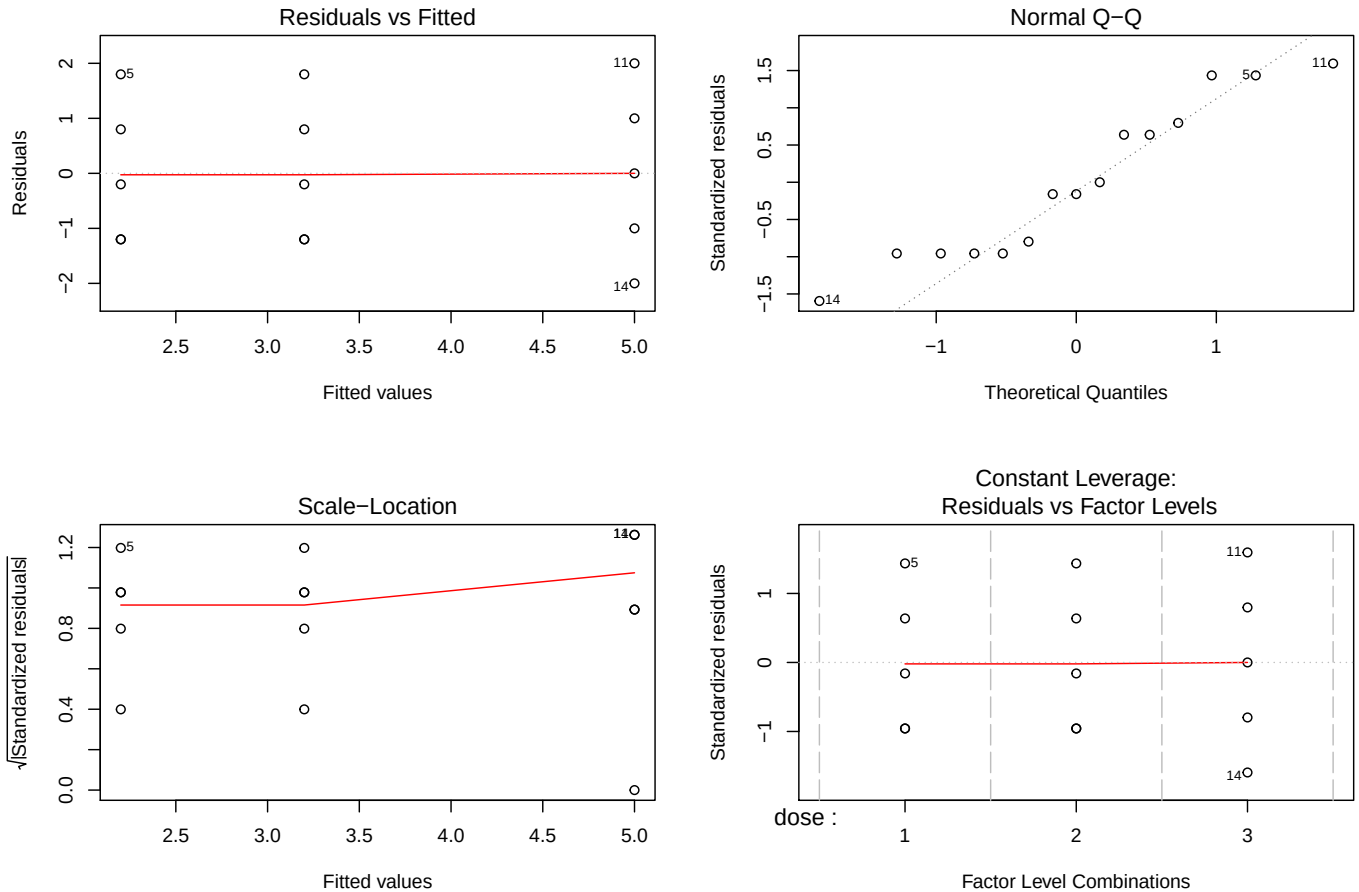


Figure 21: Output of aov-Function

Welch's F test

Welch's F-Test is similar to *t*-test for group mean differences takes into account the existing differences in group variances.

```
oneway.test(libido ~ dose, data = df)
```

```
##
##  One-way analysis of means (not assuming equal variances)
##
## data:  libido and dose
## F = 4.3205, num df = 2.0000, denom df = 7.9434, p-value = 0.05374
```

In our example there are no differences in the group variances!

Robust ANOVA

```
viagraWide <- unstack(df, libido ~ dose)
viagraWide1 <- data.frame(data = viagraWide)
colnames(viagraWide1) = c('Placebo','Low.Dose','High.Dose')
knitr::kable(viagraWide1, caption = "Viagra data set in wide format [A. Field]")
```

Table 5: Viagra data set in wide format [A. Field]

Placebo	Low.Dose	High.Dose
3	5	7
2	2	4
1	4	5
1	2	3
4	3	6

Robust ANOVA (1)

```
library(WRS2)
t1way(libido ~ dose, data = df, tr = 0.1)
```

```
## Call:
## t1way(formula = libido ~ dose, data = df, tr = 0.1)
##
## Test statistic: F = 4.3205
## Degrees of freedom 1: 2
## Degrees of freedom 2: 7.94
## p-value: 0.05374
##
## Explanatory measure of effect size: 0.71
```

```
med1way(libido ~ dose, data = df) # median
```

```
## Call:
## med1way(formula = libido ~ dose, data = df)
##
## Test statistic F: 4.7829
## Critical value: 5.473
## p-value: 0.07
```

```
t1waybt(libido ~ dose, data = df)
```

```
## Call:
## t1waybt(formula = libido ~ dose, data = df)
##
## Effective number of bootstrap samples was 384.
##
## Test statistic: 3
## p-value: 0.09115
## Variance explained 0.623
## Effect size 0.789
```

All tests are not significant and therefore the dose has no effect on libido.

Post hoc-Tests

Post hoc-Tests

- *F*-test only tells us that *there are* differences, but not between which groups.
- To make the pairwise comparisons, we need *post hoc* tests.

Bonferroni and Benjamini-Hochberg *post hoc*-tests

```r
pairwise.t.test(df$libido, df$dose, p.adjust.method = "bonferroni")
```

```
##
##  Pairwise comparisons using t tests with pooled SD
##
## data:  df$libido and df$dose
##
##   1     2
## 2 0.845 -
## 3 0.025 0.196
##
## P value adjustment method: bonferroni
```

```r
pairwise.t.test(df$libido, df$dose, p.adjust.method = "BH")
```

```
##
##  Pairwise comparisons using t tests with pooled SD
##
## data:  df$libido and df$dose
##
##   1     2
## 2 0.282 -
## 3 0.025 0.098
##
## P value adjustment method: BH
```

Tukey HSD pairwise comparisons

```
ViagraModel3 <- aov(libido ~ dose, data = df)
TukeyHSD(ViagraModel3)
```

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = libido ~ dose, data = df)
##
## $dose
##     diff         lwr      upr     p adj
## 2-1  1.0 -1.3662412 3.366241 0.5162761
## 3-1  2.8  0.4337588 5.166241 0.0209244
## 3-2  1.8 -0.5662412 4.166241 0.1474576
```

Tukey HSD pairwise comparisons (1)

```
par(las=2)
par(mar=c(5,8,4,2))
ViagraModel3 <- aov(libido ~ dose, data = df)
plot(TukeyHSD(ViagraModel3), cex = 2, cex.main=2, cex.lab=2, cex.axis=2)
```

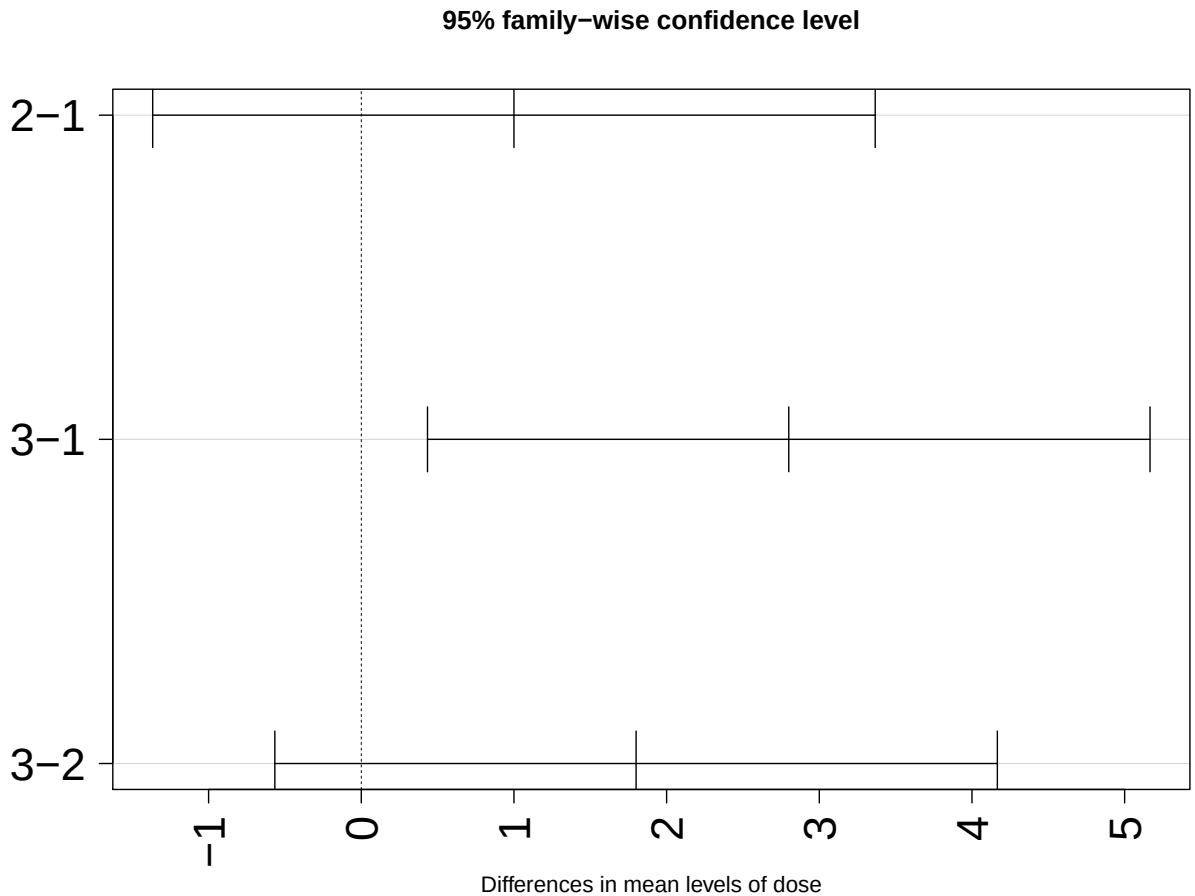


Figure 22: Tukey HSD pairwise comparisons.

Tukey HSD pairwise comparisons (2)

```
library(multcomp)
par(mar=c(5,4,6,2))

tuk <- glht(ViagraModel3, linfct = mcp(dose = "Tukey"))
summary(tuk)
```

```
##
##   Simultaneous Tests for General Linear Hypotheses
##
## Multiple Comparisons of Means: Tukey Contrasts
##
##
## Fit: aov(formula = libido ~ dose, data = df)
##
## Linear Hypotheses:
##            Estimate Std. Error t value Pr(>|t|)
## 2 - 1 == 0   1.0000     0.8869   1.127   0.5163
## 3 - 1 == 0   2.8000     0.8869   3.157   0.0209 *
## 3 - 2 == 0   1.8000     0.8869   2.029   0.1475
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## (Adjusted p values reported -- single-step method)
```

Tukey HSD pairwise comparisons (3)

```
confint(tuk)
```

```
##
##   Simultaneous Confidence Intervals
##
## Multiple Comparisons of Means: Tukey Contrasts
##
##
## Fit: aov(formula = libido ~ dose, data = df)
##
## Quantile = 2.6658
## 95% family-wise confidence level
##
##
## Linear Hypotheses:
##            Estimate lwr     upr
## 2 - 1 == 0  1.0000  -1.3644  3.3644
## 3 - 1 == 0  2.8000   0.4356  5.1644
## 3 - 2 == 0  1.8000  -0.5644  4.1644
```

For significant differences, the confidence intervals must not cross the zero!

Tukey HSD pairwise comparisons (4)

```
par(mar=c(5,8,4,2))
plot(cld(tuk, level = .05, col = "lightblue"), cex = 2, cex.main=2, cex.lab=2, cex.axis=2)
```

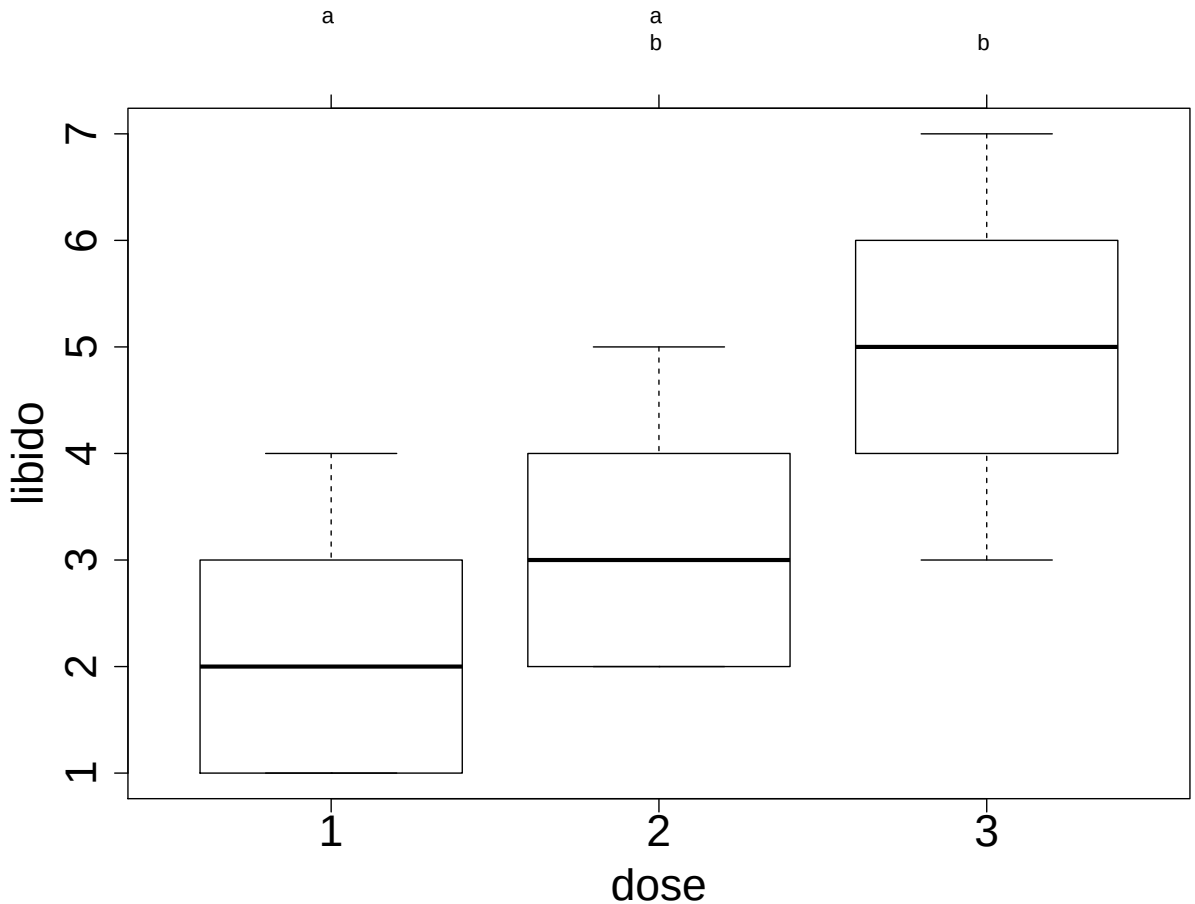


Figure 23: Tukey HSD pairwise comparisons. Groups that have the same letters at the top are not significantly different.

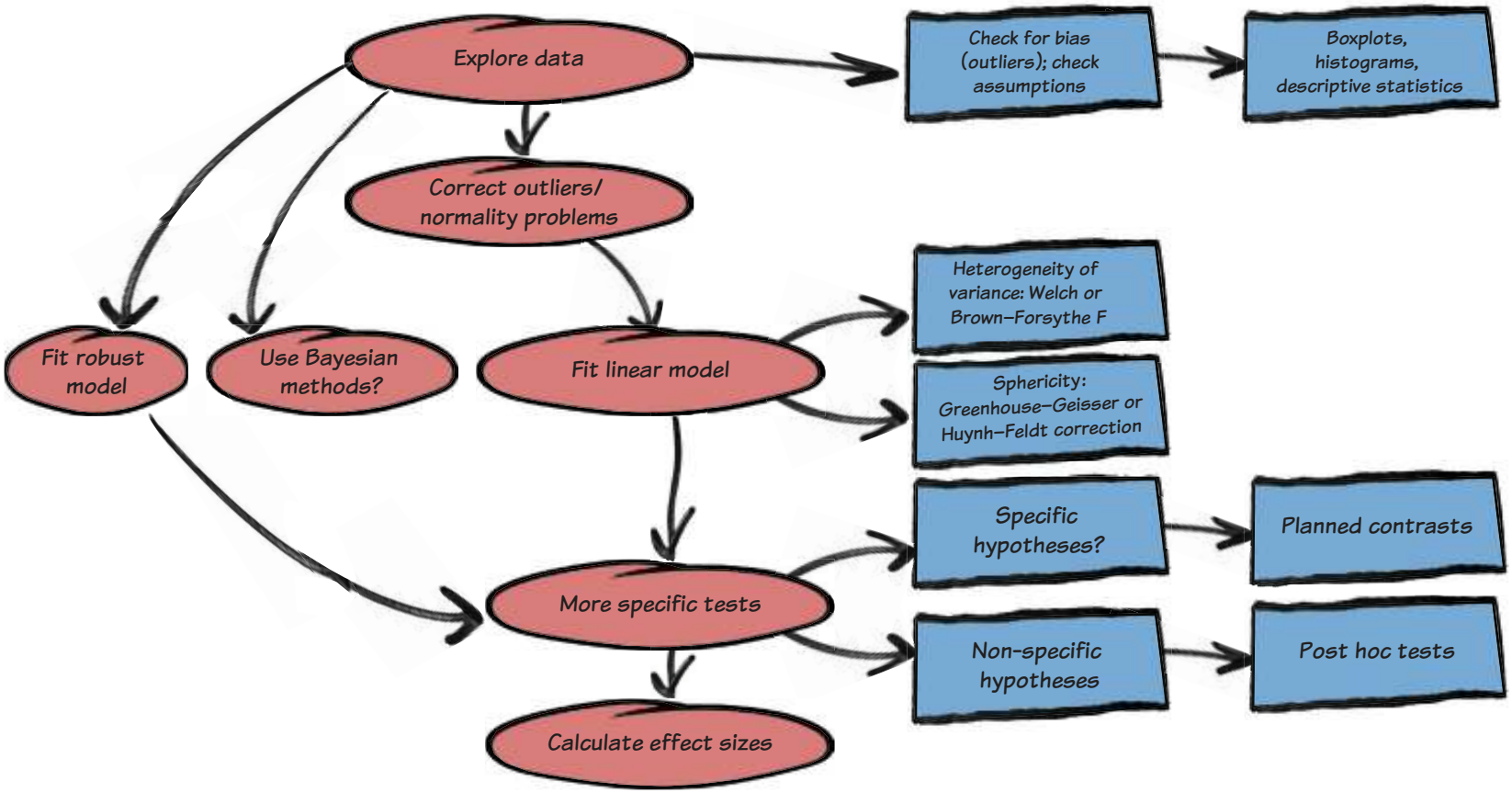


Figure 24: Comparison of several means [Quelle: A. Field]

Effect size

$$R^2 = \frac{SSM}{SST}$$

Effect size is given by R (0.68)

One denotes r^2 η^2 in ANOVA.

$$\omega^2 = \frac{SSM - (df_M)MSE}{SST + MSE}$$

$\omega = 0.60$

- $\omega^2 = 0.01$ (small effects)
- $\omega^2 = 0.06$ (medium effects)
- $\omega^2 = 0.14$ (large effects)

Contrasts (Additional material for ANOVA)

Contrasts

```
summary.lm(viagraModel)
```

```
##
## Call:
## lm(formula = libido ~ dose, data = df)
##
## Residuals:
##    Min     1Q Median     3Q    Max
##   -2.0   -1.2   -0.2    0.9    2.0
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   2.2000     0.6272   3.508  0.00432 **
## dose2         1.0000     0.8869   1.127  0.28158
## dose3         2.8000     0.8869   3.157  0.00827 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.402 on 12 degrees of freedom
## Multiple R-squared:  0.4604, Adjusted R-squared:  0.3704
## F-statistic: 5.119 on 2 and 12 DF,  p-value: 0.02469
```

Contrasts

```
contrasts(df$dose)
```

```
##   2 3
## 1 0 0
## 2 1 0
## 3 0 1
```

Planned contrasts

Given the variables or their statistics $\theta_1, \theta_2, ..., \theta_k$ and constants $a_1, a_2, ..., a_k$.

Linear combination $\sum_i a_i\theta_i$ is called *contrast* if $\sum_i a_i = 0$.

Two contrasts $\sum_i a_i\theta_i$ and $\sum_i b_i\theta_i$ are called *orthogonal* if $\sum_i a_i b_i = 0$.

Geplante Kontraste (1)

Contrast	Description
`contr.helmert`	Contrasts the second level with the first, the third level with the average of the first two, the fourth level with the average of the first three, and so on.
`contr.poly`	Contrasts are used for trend analysis (linear, quadratic, cubic, and so on) based on orthogonal polynomials. Use for ordered factors with equally spaced levels.
`contr.sum`	Contrasts are constrained to sum to zero. Also called *deviation contrasts*, they compare the mean of each level to the overall mean across levels.
`contr.treatment`	Contrasts each level with the baseline level (first level by default). Also called *dummy coding*.
`contr.SAS`	Similar to `contr.treatment`, but the baseline level is the last level. This produces coefficients similar to contrasts used in most SAS procedures.

Figure 25: Different built-in contrasts [Quelle: A. Kabacoff]

```
contrasts(df$dose) <- contr.helmert(3)
df$dose
```

```
##  [1] 1 1 1 1 1 2 2 2 2 2 3 3 3 3 3
## attr(,"contrasts")
##   [,1] [,2]
## 1   -1   -1
## 2    1   -1
## 3    0    2
## Levels: 1 2 3
```

Planned contrasts (3)

```
contrasts(df$dose) <- cbind(c(-2,1,1),c(0,-1,1))
df$dose
```

```
##  [1] 1 1 1 1 1 2 2 2 2 2 3 3 3 3 3
## attr(,"contrasts")
##   [,1] [,2]
## 1   -2    0
## 2    1   -1
## 3    1    1
## Levels: 1 2 3
```

- Planned contrasts allow to compare the groups with the positive sign against the groups with the negative sign.

- In contrast 1 we compare the placebo group and two test groups
- In contrast 2 we compare the low dose group and the high dose group

Planned contrasts (4)

```r
ViagraModel3 <- aov(libido ~ dose, data = df)
summary.lm(ViagraModel3)
```

```
##
## Call:
## aov(formula = libido ~ dose, data = df)
##
## Residuals:
##    Min     1Q Median     3Q    Max
##   -2.0   -1.2   -0.2    0.9    2.0
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   3.4667     0.3621   9.574 5.72e-07 ***
## dose1         0.6333     0.2560   2.474   0.0293 *
## dose2         0.9000     0.4435   2.029   0.0652 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.402 on 12 degrees of freedom
## Multiple R-squared:  0.4604, Adjusted R-squared:  0.3704
## F-statistic: 5.119 on 2 and 12 DF,  p-value: 0.02469
```

Planned contrasts (5)

- Because we assume that libido increases with the administration of the drug, we can replace the two-sided test with the one-sided test and halve the corresponding p values.
- Without the hypothesis that the effect increases with the dose, our result would not be significant
- This shows that it is important to have the hypothesis before we collect the data.

Logistic regression

Generalized Linear Models

Generalized Linear Models

In statistics, the generalized linear model (GLM) is a flexible generalization of ordinary linear regression that allows for response variables that have error distribution models other than a normal distribution.

$$y = \beta_0 + \sum_{i=1}^{p} \beta_i X_i,\ y \sim \mathcal{N}(\mu, \sigma)$$

The GLM generalizes linear regression by allowing the linear model to be related to the response variable via a *link function* and by allowing the magnitude of the variance of each measurement to be a function of its predicted value. - Linearität nur in den Koeffizienten

$$g(y) = \beta_0 + \sum_{i=1}^{p} \beta_i X_i$$

Generalized Linear Models

- binomial (dichotomic) logistic regression
- multinominal logistic regression

Table 6: Link functions

Datentyp	Transformation	Verteilung
kontinuierlich	$\log(x)$	Log-normal
Anzahl	$\sqrt{x}$ oder $\log(x + 0.5)$	Poisson, Negative binomiale, ...
Verhältnis	$\arcsin\sqrt{x}$ oder logit $= \log\frac{x}{1-x}$	Bernoulli / binomiale, Beta binomiale, ...

Logistic Regression

There are situations when the *response variable* is not normally distributed. E.g. it can be categorical and *binomial* or *multinomial*.

$$\log\left(\frac{\pi}{1-\pi}\right) = \beta_0 + \sum_{i=1}^{p} \beta_i X_i.$$

Here $\pi = \mu_Y$ is a conditional mean (i.e. the probability that $Y = 1$ provided the existing X values). $\frac{\pi}{1-\pi}$ is Odds-Ratio, that $Y = 1$.

$\log\left(\frac{\pi}{1-\pi}\right)$ ist *log odds* oder *logit*.

Logistic Regression: example

What personal, demographic, and relationship variables can predict infidelity?

Table 7: Extract from the infidelity dataset [aftre Kabacoff / Green&Fair]

	affairs	gender	age	yearsmarried	children	religiousness	education	occupation	rating
4	0	male	37	10.00	no	3	18	7	4
5	0	female	27	4.00	no	4	14	6	4
11	0	female	32	15.00	yes	1	12	1	4
16	0	male	57	15.00	yes	5	18	6	5
23	0	male	22	0.75	no	2	17	6	3
29	0	female	32	1.50	no	2	17	5	5
44	0	female	22	0.75	no	2	12	1	3
45	0	male	57	15.00	yes	2	14	4	4
47	0	female	32	15.00	yes	4	16	1	2
49	0	male	22	1.50	no	4	14	4	5
50	0	male	37	15.00	yes	2	20	7	2
55	0	male	27	4.00	yes	4	18	6	4
64	0	male	47	15.00	yes	5	17	6	4
80	0	female	22	1.50	no	2	17	5	4
86	0	female	27	4.00	no	4	14	5	4

Logistic Regression: example (1)

```
summary(Affairs)
```

```
##     affairs          gender         age         yearsmarried    children
##  Min.   : 0.000   female:315   Min.   :17.50   Min.   : 0.125   no :171
##  1st Qu.: 0.000   male  :286   1st Qu.:27.00   1st Qu.: 4.000   yes:430
##  Median : 0.000                Median :32.00   Median : 7.000
##  Mean   : 1.456                Mean   :32.49   Mean   : 8.178
##  3rd Qu.: 0.000                3rd Qu.:37.00   3rd Qu.:15.000
##  Max.   :12.000                Max.   :57.00   Max.   :15.000
##  religiousness     education       occupation       rating
##  Min.   :1.000   Min.   : 9.00   Min.   :1.000   Min.   :1.000
##  1st Qu.:2.000   1st Qu.:14.00   1st Qu.:3.000   1st Qu.:3.000
##  Median :3.000   Median :16.00   Median :5.000   Median :4.000
##  Mean   :3.116   Mean   :16.17   Mean   :4.195   Mean   :3.932
##  3rd Qu.:4.000   3rd Qu.:18.00   3rd Qu.:6.000   3rd Qu.:5.000
##  Max.   :5.000   Max.   :20.00   Max.   :7.000   Max.   :5.000
```

Logistic Regression: example (2)

```
knitr::kable(table(Affairs$affairs))
```

Var1	Freq
0	451
1	34
2	17
3	19
7	42
12	38

Logistic Regression: example (3)

- Transformation into binary (0 and 1 values) variables

```
Affairs$ynaffair[Affairs$affair > 0] <- 1
Affairs$ynaffair[Affairs$affair == 0] <- 0
Affairs$ynaffair <- factor(Affairs$ynaffair,
                           levels = c(0,1),
                           labels = c("No","Yes"))
knitr::kable(table(Affairs$ynaffair))
```

Var1	Freq
No	451
Yes	150

Logistische Regression: Beispiel(3)

```r
fit.full <- glm(ynaffair ~ gender + age + yearsmarried + children + religiousness + education + occupation + rating,
                data = Affairs, family = binomial() )
summary(fit.full)
```

```
##
## Call:
## glm(formula = ynaffair ~ gender + age + yearsmarried + children +
##     religiousness + education + occupation + rating, family = binomial(),
##     data = Affairs)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.5713  -0.7499  -0.5690  -0.2539   2.5191
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept)    1.37726    0.88776   1.551 0.120807
## gendermale     0.28029    0.23909   1.172 0.241083
## age           -0.04426    0.01825  -2.425 0.015301 *
## yearsmarried   0.09477    0.03221   2.942 0.003262 **
## childrenyes    0.39767    0.29151   1.364 0.172508
## religiousness -0.32472    0.08975  -3.618 0.000297 ***
## education      0.02105    0.05051   0.417 0.676851
## occupation     0.03092    0.07178   0.431 0.666630
## rating        -0.46845    0.09091  -5.153 2.56e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 675.38  on 600  degrees of freedom
## Residual deviance: 609.51  on 592  degrees of freedom
## AIC: 627.51
##
## Number of Fisher Scoring iterations: 4
```

Beispiel: reduziertes Modell

```r
fit.reduced <- glm(ynaffair ~ age + yearsmarried + religiousness + rating,
                   data = Affairs, family = binomial() )
summary(fit.reduced)
```

```
##
## Call:
## glm(formula = ynaffair ~ age + yearsmarried + religiousness +
##     rating, family = binomial(), data = Affairs)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.6278  -0.7550  -0.5701  -0.2624   2.3998
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept)    1.93083    0.61032   3.164 0.001558 **
## age           -0.03527    0.01736  -2.032 0.042127 *
## yearsmarried   0.10062    0.02921   3.445 0.000571 ***
## religiousness -0.32902    0.08945  -3.678 0.000235 ***
## rating        -0.46136    0.08884  -5.193 2.06e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 675.38  on 600  degrees of freedom
## Residual deviance: 615.36  on 596  degrees of freedom
## AIC: 625.36
##
## Number of Fisher Scoring iterations: 4
```

Example: Model comparison (χ^2)

```r
anova(fit.reduced, fit.full, test = "Chisq")
```

```
## Analysis of Deviance Table
##
## Model 1: ynaffair ~ age + yearsmarried + religiousness + rating
## Model 2: ynaffair ~ gender + age + yearsmarried + children + religiousness +
##     education + occupation + rating
##   Resid. Df Resid. Dev Df Deviance Pr(>Chi)
## 1       596     615.36
## 2       592     609.51  4   5.8474   0.2108
```

Example: Interpretation of the coefficient

Regression coefficients give the change (in *log(odds)*) in the response variable if all other variables remain constant.

```
coef(fit.reduced)
```

```
##   (Intercept)           age  yearsmarried religiousness        rating
##    1.93083017   -0.03527112    0.10062274   -0.32902386   -0.46136144
```

```
exp(coef(fit.reduced))
```

```
##   (Intercept)           age  yearsmarried religiousness        rating
##     6.8952321     0.9653437     1.1058594     0.7196258     0.6304248
```

```
exp(confint(fit.reduced))
```

```
##                    2.5 %     97.5 %
## (Intercept)   2.1255764 23.3506030
## age           0.9323342  0.9981470
## yearsmarried  1.0448584  1.1718250
## religiousness 0.6026782  0.8562807
## rating        0.5286586  0.7493370
```

Overdispersion

$$\sigma^2 = n\pi(1-\pi).$$

```
fit <-  fit.reduced
fit.od <- glm(ynaffair ~ age + yearsmarried + religiousness + rating,
                data = Affairs, family = quasibinomial() )
pchisq(summary(fit.od)$dispersion*fit$df.residual, fit$df.residual, lower = F)
```

```
## [1] 0.340122
```

Probability distributions (Additional material)

Normal distribution or Gaussian distribution

$$p(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right] = \mathcal{N}(\mu, \sigma)$$

μ — Mean

σ — Standard deviation

Normal distribution or Gaussian distribution (1)

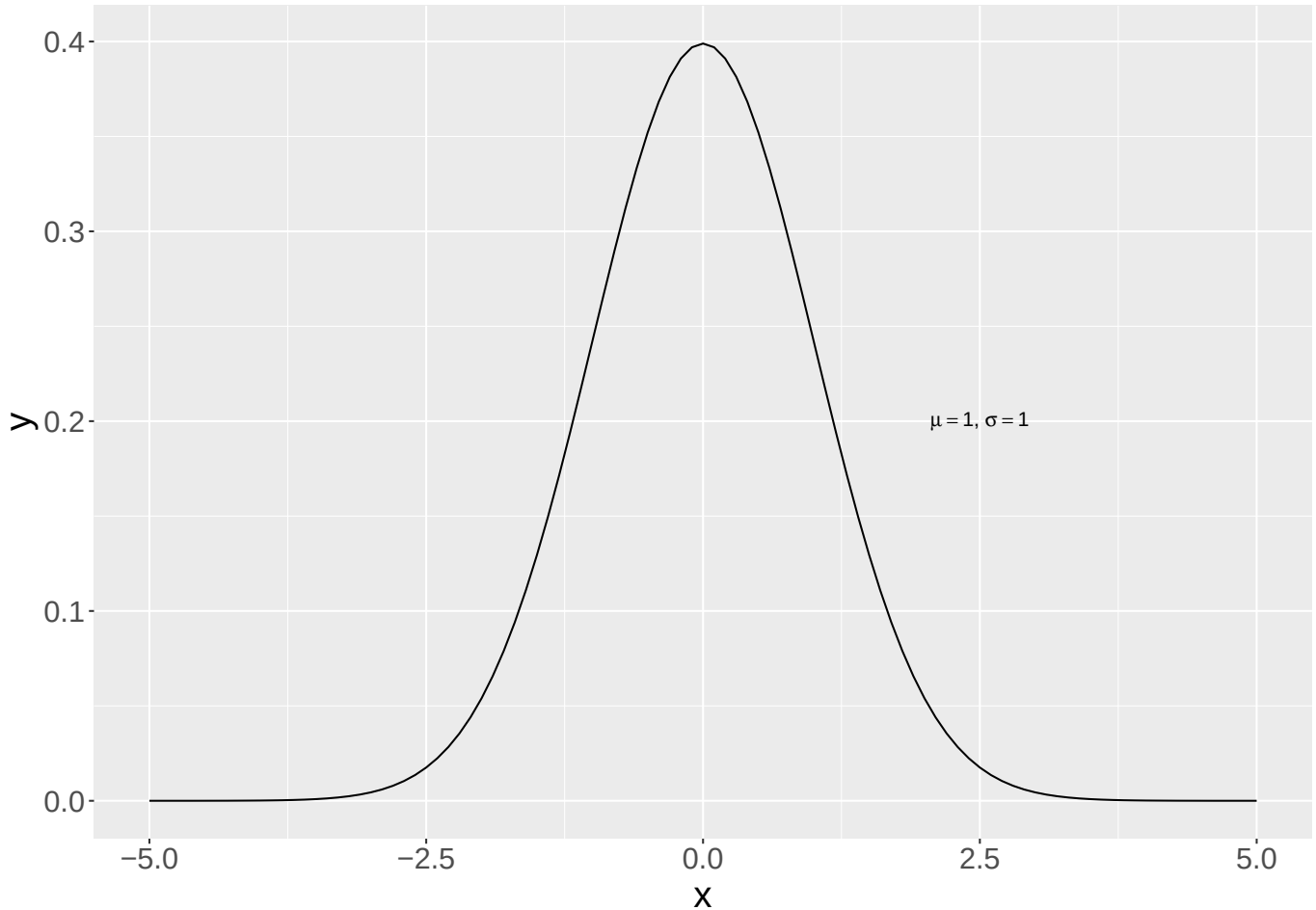


Figure 26: Normal distribution

Binomial distribution

$$p(k|n,p) = \binom{n}{k} p^k (1-p)^n - k$$

n - Number of attempts (coin flippings)

p - Success or hit probability

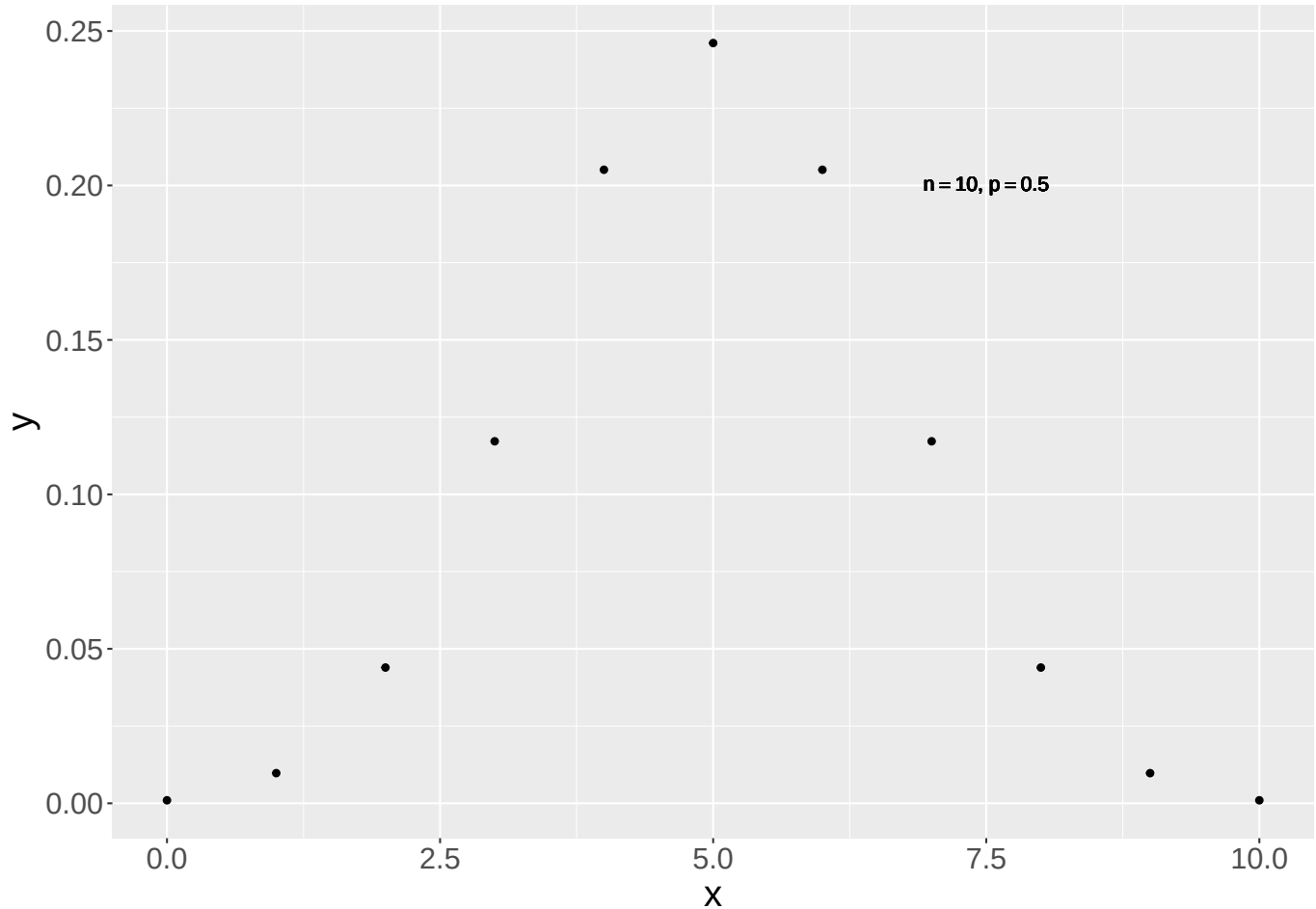


Figure 27: Binomial distribution

Poisson distribution

Poissson distribution is used e.g. for count (involving integer number) processes.

$$P_\lambda(k) = \frac{\lambda^k}{k!}\,\mathrm{e}^{-\lambda}$$

λ Expected value and the variance of the Poisson distribution are equal

Poisson distribution (1)

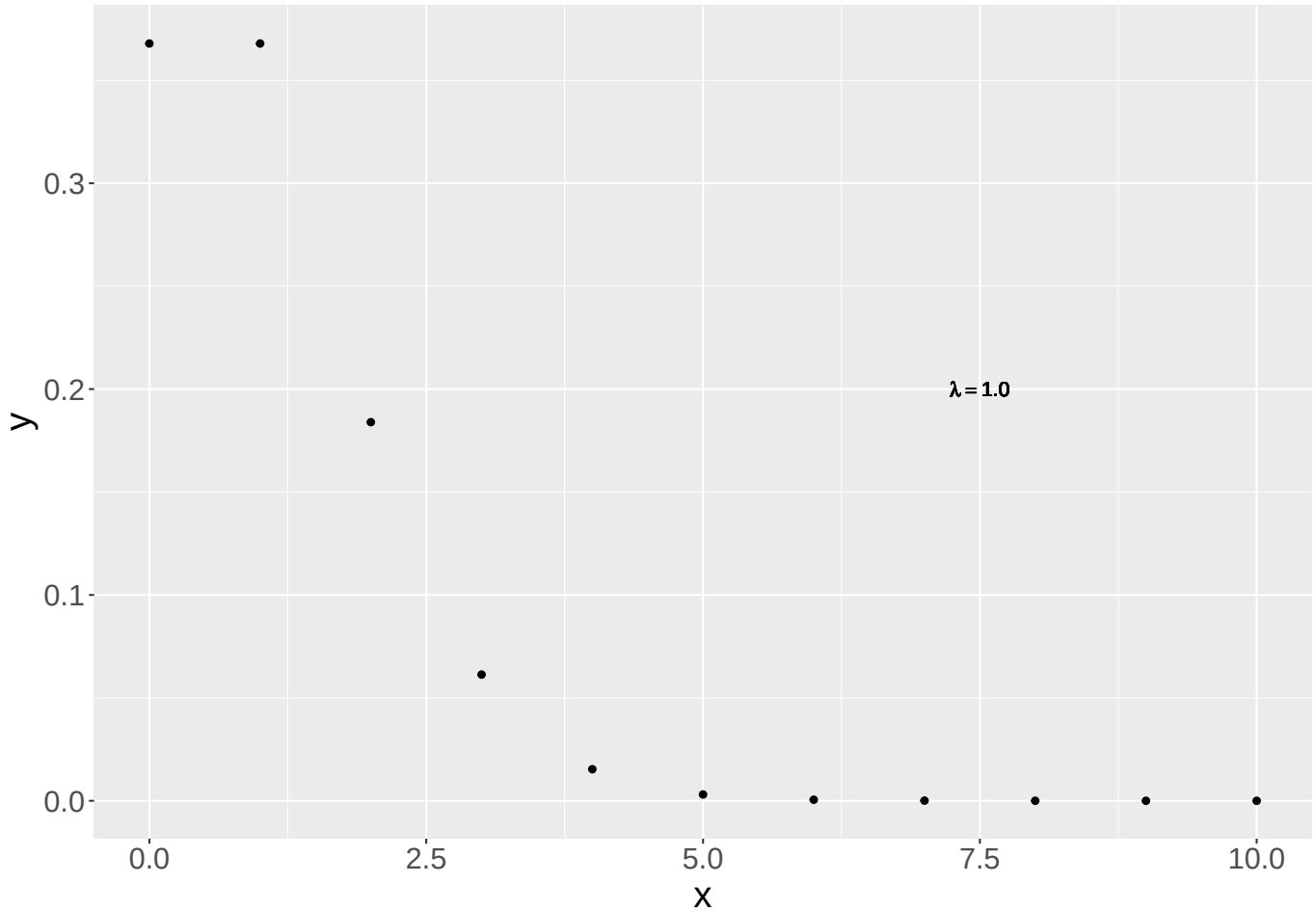


Figure 28: Poisson distribution