CARL VON OSSIETZKY universität OLDENBURG
FAKULTÄT VI
MEDIZIN UND GESUNDHEITSWISSENSCHAFTEN

Self-Test in Basic Statistics
Prerequisites for Research Master "Neurocognitive Psychology"

Psychological Methods and Statistics Lab
Contact: Prof. Dr. Andrea Hildebrandt
andrea.hildebrandt@uni-oldenburg.de

Content:

1. Descriptive statistics and graphical representation of uni- and bivariate data
2. Measures of central tendency and dispersion
3. Association of two variables
4. Basics of combinatorics
5. Elements of probability theory
6. Random variables and probability distributions
7. Statistical inference – Point and interval estimation
8. Hypothesis testing
9. Parametric tests for location parameters
10. Parametric tests for probabilities
11. Chi-square tests
12. The general linear model

Sample literature:

Heumann, C., Shalabh, & Schomaker, M. (2016). *Introduction to Statistics and Data Analysis.* With Exercises, Solutions and Applications in R. Springer.*

Navarro, D., Foxcroft, D., & Faulkenberry, T. J. (2020). Learning Statistics with JASP: A Tutorial for Psychology Students and Other Beginners. Open source: https://tomfaulkenberry.github.io/JASPbook/lsj.pdf

Note: In the following, you will find a few **exercises** (***taken from or inspired by ****) which delineate applications of basic statistics concepts. These are prerequisites for the Master's course module *Research Methods – Multivariate Statistics*.
Please try to solve the exercises and test your current knowledge. This self-test should help you to estimate the effort that you will need to invest in order to easily step into the multivariate statistics courses of our Master's programme.

- If you achieved **90-100%** correct responses, you will be **able to easily follow** the above mentioned Master's course.
- If you achieved **70-80%** correct responses, you will be able to follow the above mentioned Master's course, but **will have to catch up with some of the basics** to be able to easily follow the course.
- If you achieved **less than 70%** correct responses, we will advise you to attend our bridging module in basic statistics (no credits for the Master's degree) prior to or in parallel with multivariate statistics courses (mandatory modules for the Master's degree).

Exercise 1: Descriptive statistics and graphical representation of uni- and bivariate data

- Describe both the **population** and the **observations** for the following research questions:

(a) Evaluation of Oldenburg student's life satisfaction.

(b) Comparison of two drugs which deal with high blood pressure.

(c) Description of the marks of students from an assignment.

- Identify the **scale** of the following variables:

(a) Political party voted for in an election

(b) The difficulty of different levels in a computer game

(c) Production time of a car

(d) Age of turtles

(e) Calender year

(f) Price of a chocolate bar

(g) Identification number of a student

(h) Final ranking at a beauty contest

(i) Intelligence quotient

- Consider a **variable *X*** describing the time until the first goal was scored in the matches of the 2006 football World Cup competition. Only matches with at least one goal are considered, and goals during the *x*th minute of extra time are denoted as 90 + *x*:

6	24	90+1	8	4	25	3	83	89	34	25	24	18	6
23	10	28	4	63	6	60	5	40	2	22	26	23	26
44	49	34	2	33	9	16	55	23	13	23	4	8	26
70	4	6	60	23	90+5	28	49	6	57	33	56	7	

(a) What is the scale of *X*?

(b) Write down the frequency table of *X* based on the following categories: [0, 15), [15, 30), [30, 45), [45, 60), [60, 75), [75, 90), [90, 96).

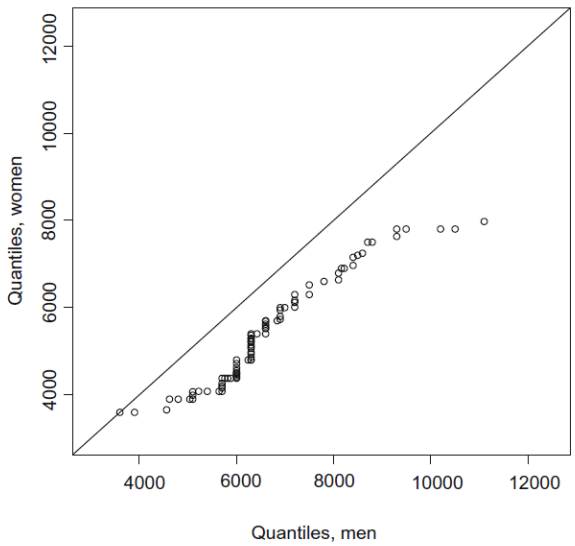
(c) Draw the histogram for *X* with intervals relating to the groups from the frequency table.

(d) Calculate the empirical cumulative distribution function for the grouped data.

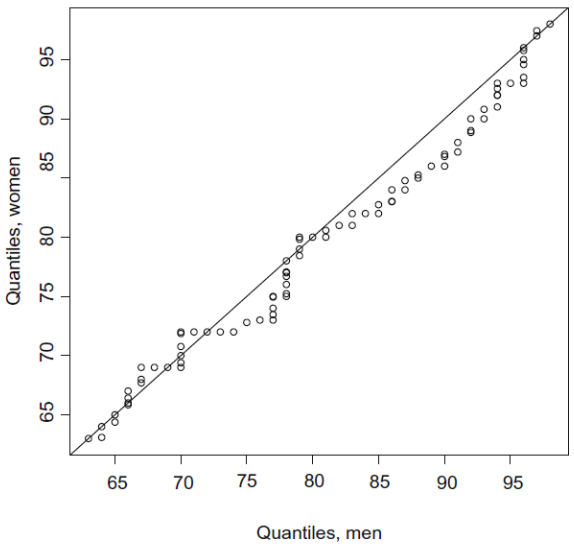
(e) Determine the time point at which in 80 % of the matches the first goal was scored at or before this time point.

Exercise 2: Measures of central tendency and dispersion

- Consider the monthly salaries Y of a well-reputed software company, as well as the length of service (in months, X), and gender (Z). The following figure shows the so called **QQ-plots** for both Y and X given Z. Interpret both graphs.



(a) for the salary



(b) for length of service

Exercise 3: Association of two variables

There has been a big debate about the usefulness of speed limits on public roads. Consider the following table which lists the speed limits for country roads (in miles/h) and traffic deaths (per 100 million km) for different countries in 1986 when the debate was particularly serious:

Country	Speed limit	Traffic deaths
Denmark	55	4.1
Japan	55	4.7
Canada	60	4.3
Netherlands	60	5.1
Italy	75	6.1

(a) Draw a scatter plot for the two variables.

(b) Calculate the Bravais–Pearson correlation coefficient.

(c) What are the effects on the correlation coefficients if the speed limit is given in km/h rather than miles/h?

(d) Consider one more observation: The speed limit for England was 70 miles/h and the death rate was 3.1. Add this observation to the scatter plot. Calculate the Bravais–Pearson correlation coefficient given this additional observation.

Exercise 4: Basics of combinatorics

- At a party with 10 guests, every guest shakes hands with each other guest. How many handshakes can be counted in total?

- A shop offers a special tray of beer: “Munich’s favourites”. Customers are allowed to fill the tray, which holds 20 bottles, with any combination of Munich’s 6 most popular beers (from 6 different breweries).

(a) What are the number of possible combinations to fill the tray?

(b) A customer insists of having at least one beer from each brewery in his tray. How many options does he have to fill the tray?

Exercise 5: Elements of probability theory

- A driving licence examination consists of two parts which are based on a theoretical and a practical examination. Suppose 25% of people fail the practical examination, 15% of people fail the theoretical examination, and 10% of people fail both the examinations. Now, if a person is randomly chosen, then what is the probability that this person

(a) fails at least one of the examinations?

(b) successfully passes both tests?

Exercise 6: Random variables and probability distributions

- Please fill in the gaps:

(a) If a random variable Y is normally distributed, Y^2 follows a _____ distribution with _____ degree of freedom.

(b) If $Y_1, Y_2, ..., Y_k$ are independent, normally distributed random variables, then the sum of their squares will follow a _____ distribution with _____ degrees of freedom.

(c) Like the standard normal distribution, _____ are symmetrical, unimodal distributions with an expected value of _____. Compared to the _____ distribution, _____ are more leptokurtic, which however decreases with increasing _____. With _____of degrees of freedom, the _____changes into _____ distribution.

Exercise 7: Statistical inference – Point and interval estimation

- A national park in Namibia determines the weight (in kg) of a sample of common eland antelopes:

450 730 700 600 620 660 850 520 490 670 700 820

910 770 760 620 550 520 590 490 620 660 940 790

Please calculate

(a) the point estimate of μ (expectation) and σ^2 (variance)

(b) the confidence interval for μ (α = 0.05).

Exercise 8: Hypothesis testing

- A producer of chocolate bars hypothesizes that his production does not adhere to the weight standard of 100 g. As a measure of quality control, he weighs 15 bars and obtains the following results in grams:

96.40, 97.64, 98.48, 97.67, 100.11, 95.29, 99.80, 98.80, 100.53, 99.41, 97.64, 101.11, 93.43, 96.99, 97.92

The producer further assumes that the production process is standardized in the sense that the variation is controlled to be σ = 2.

(a) What are the hypotheses regarding the expected weight μ for a two-sided test?

(b) Which test should be used to test these hypotheses?

(c) Conduct the test that you suggested to be used in (b). Please use α = 0.05.

(d) The producer wants to show that the expected weight is smaller than 100 g. What are the appropriate hypotheses to use?

(e) Conduct the test for the hypothesis in (d). Please use α = 0.05.

Exercise 9: Parametric tests for location parameters

Ten best friends try a new diet: the “Banting” diet. Each of them weighs him/herself before and after the diet. The data is as follows:

Person (i)	1	2	3	4	5	6	7	8	9	10
Before diet (x_i)	80	95	70	82	71	70	120	105	111	90
After diet (y_i)	78	94	69	83	65	69	118	103	112	88

(a) Please choose a test and a confidence interval to test whether there is a difference between the mean weight before and after the diet (α = 0.05).

Exercise 10: Parametric tests for probabilities

- A company producing clothing often finds deficient T-shirts among its production.

(a) The company’s controlling department decides that the production is no longer profitable when there are more than 10% deficient shirts. A sample of 230 shirts yields 30 shirts which contain deficiencies. Use the approximate binomial test to decide whether the T-shirt production is profitable or not (α = 0.05).

Exercise 11: Chi-square tests

- Please test the hypothesis that travel class on the Titanic and rescue status are not independent. The data can be summarized as follows:

	1. Class	2. Class	3. Class	Staff	Total
Rescued	202	125	180	211	718
Not rescued	135	160	541	674	1510

Exercise 12: The general linear model

- Consider the data visualized in the following table:

Obs	Test *x*	Test *y*	x^2	$x \cdot y$
1	31	15	961	465
2	128	95	16384	12160
3	67	35	4489	2354
4	46	40	2116	1840
5	180	80	32400	14400
Sum:	452	265	56350	31210

(a) Display the data in a scatter plot.

(b) Calculate the intercept and the slope for the following linear model: $y = \alpha + \beta x + \epsilon$

(c) Provide an interpretation for α and β.

Solutions

Solution of exercise 1: Descriptive statistics and graphical representation of uni- and bivariate data

- Describe both the **population** and the **observations** for the following research questions:

(a) Evaluation of Oldenburg student's life satisfaction.

The population consists of all students in Oldenburg. This may include students of different scientific disciplines. Each single student relates to an observation in the survey.

(b) Comparison of two drugs which deal with high blood pressure.

All individuals suffering high blood pressure in the study area (city, province, country, etc.), are the population of interest. Each individual is an observation.

(c) Description of the marks of students from an assignment.

The population comprises all students who take part in the examination. Each student represents an observation.

- Identify the **scale** of the following variables:

(a) Political party voted for in an election

Nominal

(b) The difficulty of different levels in a computer game

Ordinal

(c) Production time of a car

Continuous ratio scale

(d) Age of turtles

Continuous ratio scale

(e) Calender year

Continuous interval scale

(f) Price of a chocolate bar

Continuous ratio scale

(g) Identification number of a student

Nominal

(h) Final ranking at a beauty contest

Ordinal

(i) Intelligence quotient

Continuous interval scale

- Consider a **variable *X*** describing the time until the first goal was scored in the matches of the 2006 football World Cup competition. Only matches with at least one goal are considered, and goals during the *x*th minute of extra time are denoted as 90 + *x*:

6	24	90+1	8	4	25	3	83	89	34	25	24	18	6
23	10	28	4	63	6	60	5	40	2	22	26	23	26
44	49	34	2	33	9	16	55	23	13	23	4	8	26
70	4	6	60	23	90+5	28	49	6	57	33	56	7	

(a) What is the scale of *X*?

The scale of X is continuous.

(b) Write down the frequency table of *X* based on the following categories: [0, 15), [15, 30), [30, 45), [45, 60), [60, 75), [75, 90), [90, 96).

j	$[e_{j-1}, e_j)$	n_j	f_j	d_j	h_j	$F(x)$
1	[0, 15)	19	$\frac{19}{55}$	15	$\frac{19}{825}$	$\frac{19}{55}$
2	[15, 30)	17	$\frac{17}{55}$	15	$\frac{17}{825}$	$\frac{36}{55}$
3	[30, 45)	6	$\frac{6}{55}$	15	$\frac{6}{825}$	$\frac{42}{55}$
4	[45, 60)	5	$\frac{5}{55}$	15	$\frac{5}{825}$	$\frac{47}{55}$
5	[60, 75)	4	$\frac{4}{55}$	15	$\frac{4}{825}$	$\frac{51}{55}$
6	[75, 90)	2	$\frac{2}{55}$	15	$\frac{2}{825}$	$\frac{53}{55}$
7	[90, 96)	2	$\frac{2}{55}$	6	$\frac{2}{825}$	1
Total		56	1			

(c) Draw the histogram for *X* with intervals relating to the groups from the frequency table.

Obtain the heights for each category to obtain the histogram using $h_j = f_j / d_j$

(d) Calculate the empirical cumulative distribution function (ECDF) for the grouped data.

The ECDF values for F(x) are calculated using the relative frequencies f(x), see Table above.

(e) Determine the time point at which in 80 % of the matches the first goal was scored at or before this time point.

80% of the first goals happened up to 50.4 minutes.

Solution of exercise 2: Measures of central tendency and dispersion

Look at the figure at the right side first. The quantiles coincide approximately with the bisection line. This means that the distribution of "length of service" is similar for men and women. They have worked for about the same amount of time for the company. For example, the median service time should be approximately the same for men and

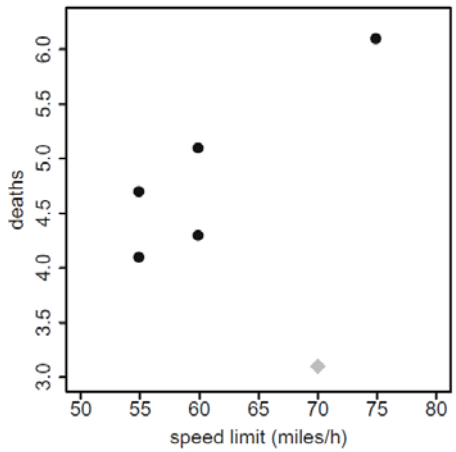
women, the first quartile should be very similar too, the third quartile should be similar too, and so on. However, the figure at the left side shows a somewhat different pattern: the quantiles for men are consistently higher than those for women. For example, the median salary will be higher for men. In conclusion, we see that men and women have a similar length of service in the company, but earn less.

Solution of exercise 3: Association of two variables

There has been a big debate about the usefulness of speed limits on public roads. Consider the following table which lists the speed limits for country roads (in miles/h) and traffic deaths (per 100 million km) for different countries in 1986 when the debate was particularly serious:

Country	Speed limit	Traffic deaths
Denmark	55	4.1
Japan	55	4.7
Canada	60	4.3
Netherlands	60	5.1
Italy	75	6.1

(a) Draw a scatter plot for the two variables.



(b) Calculate the Bravais–Pearson correlation coefficient.

There is a positive correlation of 0.891: the higher the speed limit, the higher the number of deaths.

(c) What are the effects on the correlation coefficients if the speed limit is given in km/h rather than miles/h?

The results stay the same. Pearson's correlation coefficient is invariant with respect to linear transformations.

(d) Consider one more observation: The speed limit for England was 70 miles/h and the death rate was 3.1. Add this observation to the scatter plot. Calculate the Bravais–Pearson correlation coefficient given this additional observation.

The overall pattern of the scatter plot changes with this new observation pair: the positive relationship between speed limit and number of traffic deaths is not so clear anymore: 0.351.

Solution of exercise 4: Basics of combinatorics

- At a party with 10 guests, every guest shakes hands with each other guest. How many handshakes can be counted in total?

There are n = 10 guests and m = 2 guests shake hands with each other. The order is not important: two guests shake hands no matter who is "drawn first". It is also not possible to shake hands with oneself which means that we have the "no replacement" case. Therefore, we obtain the solution as:

$$\binom{n}{m} = \binom{10}{2} = \frac{10 \cdot 9}{2 \cdot 1} = 45$$

- A shop offers a special tray of beer: "Munich's favourites". Customers are allowed to fill the tray, which holds 20 bottles, with any combination of Munich's 6 most popular beers (from 6 different breweries).

(a) What are the number of possible combinations to fill the tray?

The customer takes the beers "with replacement" because the customer can choose among any type of beer for each position in the tray. Therefore, we obtain the solution as:

$$\binom{n+m-1}{m} = \binom{6+20-1}{20} = \binom{25}{20} = 53,130$$

(b) A customer insists of having at least one beer from each brewery in his tray. How many options does he have to fill the tray?

If the customer insists on having at least one beer per brewery on his tray, then 6 out of the 20 positions of the tray are already occupied. Therefore, we obtain the solution as:

$$\binom{n+m-1}{m} = \binom{6+14-1}{14} = \binom{19}{14} = 11,628$$

Solution of exercise 5: Elements of probability theory

- A driving licence examination consists of two parts which are based on a theoretical and a practical examination. Suppose 25% of people fail the practical examination, 15% of people fail the theoretical examination, and 10% of people fail both the examinations. Now, if a person is randomly chosen, then what is the probability that this person

(a) fails at least one of the examinations?

If we ask for the probability of failing in at least one examination, we imply that either the theoretical examination, or the practical examination, or both are not passed. We can therefore calculate the union of events: 0.25 + 0.15 – 0.10 = 0.3

(b) successfully passes both tests?

1 – 0.30 = 0.7

Solution of exercise 6: Random variables and probability distributions

- Please fill in the gaps:

(a) If a random variable Y is normally distributed, Y^2 follows a *chi-square* distribution with *one* degree of freedom.

(b) If $Y_1, Y_2, ..., Y_k$ are independent, normally distributed random variables, then the sum of their squares will follow a *chi-square* distribution with *three* degrees of freedom.

(c) Like the standard normal distribution, *t-distributions* are symmetrical, unimodal distributions with an expected value of $\mu = 0$. Compared to the *standard normal* distribution, *t-distributions* are more leptokurtic, which however decreases with increasing *number of degrees of freedom*. With *increasing number* of degrees of freedom, the *t-distribution* changes into *a standard normal* distribution.

Solution of exercise 7: Statistical inference – Point and interval estimation

- A national park in Namibia determines the weight (in kg) of a sample of common eland antelopes:

450 730 700 600 620 660 850 520 490 670 700 820

910 770 760 620 550 520 590 490 620 660 940 790

Please calculate

(a) the point estimate of μ (expectation) and σ^2 (variance)

$\mu = 667.92$

$\sigma^2 \approx 18, 035$

(b) the confidence interval for μ (α = 0.05).

[611.17; 724.66]

Solution of exercise 8: Hypothesis testing

- A producer of chocolate bars hypothesizes that his production does not adhere to the weight standard of 100 g. As a measure of quality control, he weighs 15 bars and obtains the following results in grams:

96.40, 97.64, 98.48, 97.67, 100.11, 95.29, 99.80, 98.80, 100.53, 99.41, 97.64, 101.11, 93.43, 96.99, 97.92

The producer further assumes that the production process is standardized in the sense that the variation is controlled to be σ = 2.

(a) What are the hypotheses regarding the expected weight μ for a two-sided test?

H_0: μ = 100 versus H_1: μ ≠100.

(b) Which test should be used to test these hypotheses?

It is a one-sample problem for μ: thus, for known variance, the Gauss test can be used.

(c) Conduct the test that you suggested to be used in (b). Please use α = 0.05.

We need the arithmetic mean which is 98.08. t(x) will then be −3.72. We reject H_0 if |t(x)| > 1.96.

(d) The producer wants to show that the expected weight is smaller than 100 g. What are the appropriate hypotheses to use?

To show that μ < 100, we need to conduct a one-sided test. The research hypothesis of interest needs to be stated as the alternative hypothesis:

H_0: μ ≥ 100 versus H_1: μ < 100.

(e) Conduct the test for the hypothesis in (d). Please use α = 0.05.

The test statistic is the same as above: t(x) = −3.72. However, the critical region changes: We reject H_0 if t(x) < −1.64.

Solution of exercise 9: Parametric tests for location parameters

Ten best friends try a new diet: the "Banting" diet. Each of them weighs him/herself before and after the diet. The data is as follows:

Person (i)	1	2	3	4	5	6	7	8	9	10
Before diet (x_i)	80	95	70	82	71	70	120	105	111	90
After diet (y_i)	78	94	69	83	65	69	118	103	112	88

(a) Please choose a test and a confidence interval to test whether there is a difference between the mean weight before and after the diet (α = 0.05).

Since the data before and after the diet is dependent (weight measured on the same individuals), we need to use the paired t-test. The test statistic is 2.42. The null hypothesis is rejected because the test statistic is larger than 2.26 (the critical value).

Solution of exercise 10: Parametric tests for probabilities

- A company producing clothing often finds deficient T-shirts among its production.

(a) The company's controlling department decides that the production is no longer profitable when there are more than 10% deficient shirts. A sample of 230 shirts yields 30 shirts which contain deficiencies. Use the approximate binomial test to decide whether the T-shirt production is profitable or not (α = 0.05).

The production is no longer profitable if the probability of finding a deficient shirt is greater than 10%. This equates to the hypotheses:

H_0: $p \leq 0.1$ versus H_1: $p > 0.1$.

The null hypothesis gets rejected because $t(x) = 2.638 > z_{0.95} = 1.64$. It seems the production is no longer profitable.

Solution of exercise 11: Chi-square tests

- Please test the hypothesis that travel class on the Titanic and rescue status are not independent. The data can be summarized as follows:

	1. Class	2. Class	3. Class	Staff	Total
Rescued	202	125	180	211	718
Not rescued	135	160	541	674	1510

To answer this question, we need to conduct the $\chi 2$-independence test. The test statistic $t(x, y)$ is identical to Pearson's $\chi 2$ statistic ≈ 182. The null hypothesis of independence is rejected if $t(x, y) = \chi 2 > c_{(I-1)(J-1);1-\alpha}$. With $I = 2$ (number of rows), and $J = 4$ (number of columns) by using the chi-square table or a statistics software like R, we obtain $c_{3;0.95} = 7.81$. Since $182 > 7.81$, we reject the null hypothesis of independence.

Solution of exercise 12: The general linear model

- Consider the data visualized in the following table:

Obs	Test x	Test y	x^2	$x \cdot y$
1	31	15	961	465
2	128	95	16384	12160
3	67	35	4489	2354
4	46	40	2116	1840
5	180	80	32400	14400
Sum:	452	265	56350	31210

(a) Display the data in a scatter plot.

(b) Calculate the intercept and the slope for the following linear model: $y = \alpha + \beta x + \epsilon$

$y = 10.66 + 0.47\, x + \epsilon$

(c) Provide an interpretation for α and β.

α indicates the expected value of y, given x = 0.
β indicates the change of y for one unit change of x.