Biostatistical calculations

Last update: November 17, 2016

University of Szeged
Department of Medical Physics and Informatics

1 Distribution of discrete and continuous variables

1.1 Questionnaire (sample)

1. Identification number

☐☐☐

2. Gender

1. male
2. female

3. Age (in years)

☐☐

4. Education

1. no
2. primary
3. secondary
4. university

5. Body mass (kg)

☐☐☐

6. Height (cm)

☐☐☐

7. Eye colour

1. blue
2. green
3. grey
4. brown
5. black
6. other

8. Hobby

- ☐ sport
- ☐ music listening
- ☐ collecting stamps
- ☐ dancing
- ☐ fine arts
- ☐ other

1.1.1 Create variables using the questionnaire!

variable	type
ID	
GENDER	categorical, binary
AGE	
EDUCATION	categorical, ordinal
WEIGHT	
HEIGHT	continuous, quantitative

variable	type
E_COLOUR	categorical, nominal
SPORT	
MUSIC	
STAMP	
DANCE	
FINEART	categorical, binary

1.1.2 Dataset (sample) based on the questionnaire

ID	GENDER	AGE	EDUCATION	WEIGHT	HEIGHT	E_COLOUR	SPORT	MUSIC
1	1	20	3	65	185	3	1	1
2	2	17	3	60	170	4	1	2
3	1	22	3	62	177	2	2	1
4	2	28	4	62	176	4	2	1
5	1	9	1	32	148	4	2	2
6	1	5	1	19	125	3	2	2
7	2	26	3	70	166	4	2	2
8	1	60	4	75	180	1	1	1
9	2	35	3	49	155	4	2	1
10	2	51	4	61	162	4	2	1
11	1	17	2	61	178	4	2	1
12	2	50	2	65	164	4	2	2
13	1	9	1	30	130	2	1	2
14	2	10	1	40	135	1	2	1
15	1	19	3	86	187	3	1	1
16	1	22	3	67	179	4	2	2
17	1	25	3	103	186	4	1	1
18	1	29	4	74	176	1	1	1
19	2	27	4	67	164	4	1	1
20	1	19	3	70	180	4	1	1

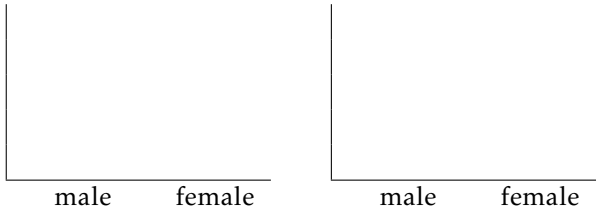
1.2 Discrete variables: distribution, absolute and relative frequency, bar chart

1.2.1 Characterize the GENDER variable.

Frequency table

	frequency	relative frequency (%)
male female		
Total		

Bar charts (absolute and relative frequencies)

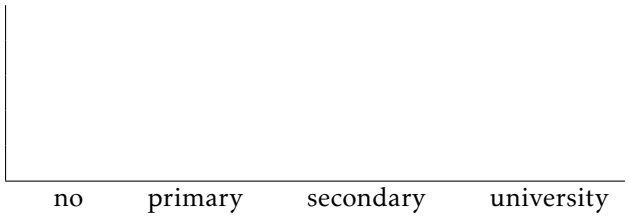


1.2.2 Characterize the EDUCATION variable!

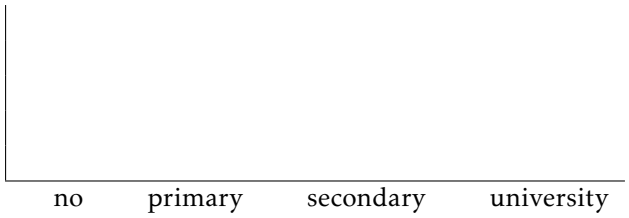
Frequency table

	frequency	relative frequency (%)
No Primary school Secondary school University		
Total		

Bar chart – (absolute) frequency



Bar chart – relative frequency



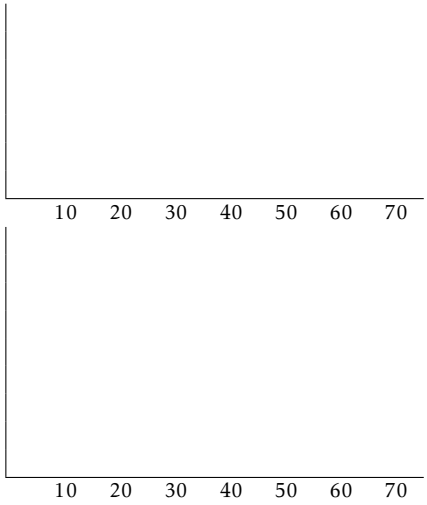
1.3 Continuous variables: absolute and relative frequency, histogram

1.3.1 Characterize the AGE variable! Create a histogram, make scale on y-axis. Interpret the results!

Frequency table

AGE	frequency	relative frequency (%)
[0,10)		
[10,20)		
[20,30)		
[30,40)		
[40,50)		
[50,60)		
[60,70)		
total		

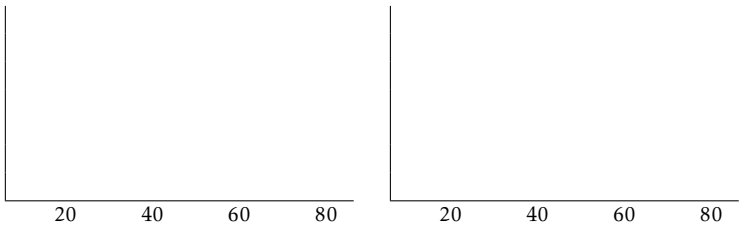
Frequency and relative frequency histogram



1. Is the shape of the distribution symmetrical (or skewed)?
2. Which interval contains the most/fewest elements?

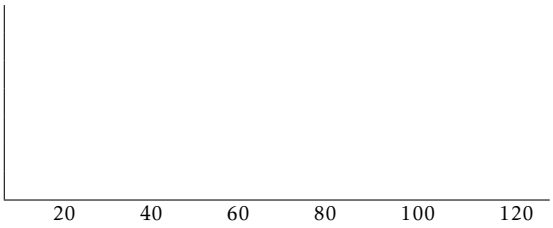
Double the interval width. Create histograms using data of variable AGE. Make scale on y-axis.

AGE	frequency	relative frequency (%)
[0, 20)		
[20, 40)		
[40, 60)		
[60, 80)		
total		



1.3.2 Characterize the WEIGHT variable as done it for variable AGE!

WEIGHT	frequency	relative frequency (%)
[0, 20)		
[20, 40)		
[40, 60)		
[60, 80)		
[80, 100)		
[100, 120)		
total		



1.4 Continuous variables: measures of central tendency and variability

1.4.1 Calculate mean, median, mode, range, standard deviation of these random samples.

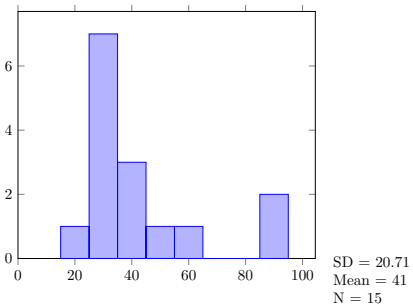
	n	Random sample	Mean	Median	Mode	Range	Standard deviation
1.	$n = 4$	1 2 4 1					
2.	$n = 4$	10 20 40 10					
3.	$n = 4$	2 4 8 2					
4.	$n = 4$	2 3 5 2					
5.	$n = 6$	1 3 2 4 0 2					

1.4.2 The numbers of some special operations made by 15 man operators are the following

20 25 25 27 28 31 33 34 36 37 44 50 59 85 86

Based on the shape of the histogram, which statement is true? Why?

1. the mean is smaller than the median
2. the mean is approximately equal to the median
3. the mean is greater than the median



Find the quartiles of the above sample.

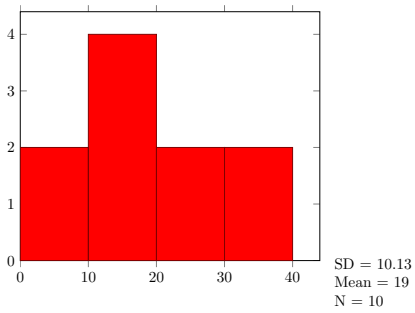
1. first (lower) quartile (25% percentile):
2. second quartile (50% percentile, median):
3. third (upper) quartile (75% percentile):

Draw a box plot. Compare this to the histogram.
Draw a mean-SD diagram. Can you conclude to the symmetry of the distribution based on this diagram?

1.4.3 The numbers of some special operations made by 10 woman operators are the following:

5 7 10 14 18 19 25 29 31 33

Based on the shape of the histogram, which statement is true? Why?



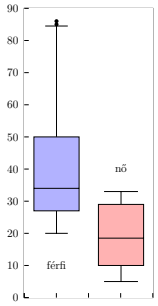
1. the mean is smaller than the median
2. the mean is approximately equal to the median
3. the mean is greater than the median

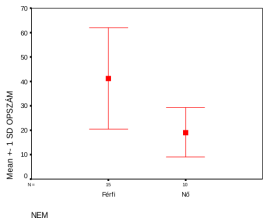
Find the quartiles of the above sample.

1. first (lower) quartile (25% percentile):
2. second quartile (50% percentile, median):
3. third (upper) quartile (75% percentile):

Draw a box plot. Compare this to the histogram.
Draw a mean-SD diagram. Can you conclude to the symmetry of the distribution based on this diagram?

1.4.4 Putting the diagrams side by side, the two distributions can be compared. Which diagram gives more information and why?





1.5 Calculations with R

1.5.1 Type the samples from the Exercise 1.4.1 in R and calculate the descriptive statistics.

1.5.2 Open the `smallquest_labels.csv` data file! Repeat the characterization of both GENDER, EDUCATION, AGE and WEIGHT variables using the R commands

1.5.3 Open the `bank.csv` data file. Characterize the continuous variables. Calculate the descriptive statistics of all the continuous variables, e.g. salnow (current salary). Interpret the results.

1.6 Homework

1.6.1 Here are the blood types of 20 patients. Characterize the distribution of this data.

A 0 0 B AB 0 A B B AB 0 A 0 AB B 0 AB 0 B AB

1.6.2 Given the following temperature values. Calculate summary statistics for the sample, construct a frequency histogram and a box diagram. Conclude to the symmetry of the distribution.

35.1 36.1 35.2 36.2 36.5 36.5 37 36.2 36.8 36.7 36.5

1.6.3 Consider the following ordered set of data. Find the mean, first quartile, median, third quartile, range, mode, standard deviation of this data.

44 49 50 51 53 57 58 62 66 66 68 71 75 77 80 85

1.6.4 Interpret the results of the table below (Varga Z et al: Individualized positioning for maximum heart protection breast irradion. Acta Oncologica 2013)

Table I. Baseline characteristics of the patients included in the study (n = 138).

	Training set (n = 83)			Validation set (n = 55)		
Variable	Median	Mean ± SE	Range	Median	Mean ± SE	Range
Age (years)	60.2	59.4 ± 1.1	31.1–79.1	59.1	58.4 ± 1.4	26.0–76.5
Weight (kg)	73.0	73.4 ± 1.4	46.0–112.0	75.0	77.1 ± 1.9	49.0–120.0
Height (cm)	162.0	162.4 ± 0.7	149.0–178.0	162.0	161.1 ± 0.8	140.0–179.0
BMI (kg/m^2)	27.5	27.8 ± 0.5	17.1–38.9	28.4	29.7 ± 0.7	20.3–44.1
Breast volume (cm^3)	897.0	983.1 ± 46.8	197.0–2448.0	1061.0	1050.6 ± 65.1	257.0–2838.0
Heart volume (cm^3)	515.0	522.0 ± 11.6	307.0–965.0	540.0	553.5 ± 14.7	360.0–862.0
d_{median} (cm)	1.3	1.33 ± 0.1	0.4–2.2	1.4	1.39 ± 0.1	0.3–3.3
A_{heart} (mm^2)	549.0	599.9 ± 43.9	0–1820.0	455.0	476.9 ± 50.0	0–1627.0

2 Probability, conditional probability, diagnostic tests, discrete and normal distributions

2.1 Probability calculus

2.1.1 If we roll a dice, there are 6 possible outcomes.

If X represents the value of the outcome, find the following probabilities:

a) $P(X = 1) =$ _____

b) $P(X > 1) =$ _____

c) $P(1 < X < 4) =$ _____

2.1.2 If we roll two dices, there are 36 possible outcomes.

If X represents the sum, Y the product of the rolled numbers, find the following probabilities:

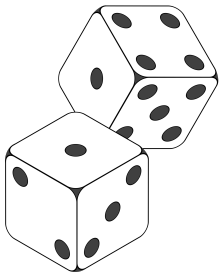
a) $P(X = 2) =$ _____

b) $P(X > 2) =$ _____

c) $P(X = 10) =$ _____

d) $P(Y = 2) =$ _____

e) $P(Y = 12) =$ _____



2.1.3 A fair coin is tossed twice.

a) List the possible outcomes: _____

b) Find the probability of getting two heads: _____

head tail

2.1.4 A penny is tossed once and a dice is rolled once

a) List the possible outcomes: _____

Find the probabilities of the following outcomes:

b) tossing a head and rolling an even number: _____

c) tossing a head or rolling an even number: _____

d) tossing a head and rolling a 5: _____

e) tossing a head or rolling a 5: _____

f) rolling either a 4 or a 6: _____

2.1.5 The blood groups of 200 people are distributed as follows

50 have blood type A, 65 have blood type B, 70 have blood type 0 and 15 have blood type AB. If a person from this group is selected at random, what is the probability that this person has blood type 0?

2.2 Discrete distributions

2.2.1 A penny is tossed 3 times

Elementary events: _____

Define the variable X to be the number of heads. Prepare the distribution of the variable X.

X	0	1	2	3
P(X)				

2.2.2 Which of the following distributions are probability distributions?

a)

X	0	5	10	15	20
P(X)	1/5	1/5	1/5	1/5	1/5

b)

X	0	2	4	6
P(X)	1/4	1/8	1/16	9/16

c)

X	0	2	4	6
P(X)	-1	1.5	0.30	0.2

d)

X	0	2	4	6
P(X)	1/4	1/8	1/16	2/16

2.3 Conditional probability

2.3.1 We have large number of families with two children. After selecting randomly a girl from this set of families estimate the probability of the event that there is a boy also in that family.

a) List of the elementary events: _____

b) Event A: *there is at least one girl in the family* GB, BG, GG P(A) = _____

c) Event B: *there is at least one boy in the family* _____ P(B) = _____

d) Event A and B ($A \cdot B$): _____

Possible cases from the list of elementary events: _____

P(AB) = _____

e) Conditional probability is $P(B|A) = \frac{P(AB)}{P(A)}$ = _____

2.3.2 Calculate the probability of a 2 being rolled by a dice if it is already known that the result is even

_____

2.3.3 A dice is rolled twice, we got different numbers.

What is the probability that at least one of them is a 6?

_____

2.4 Diagnostic tests

2.4.1 Relation between results of liver scan and correct diagnosis are summarised in the following table.

Calculate sensitivity, specificity, positive (PPV) and negative (NPV) predictive values.

Source: D G Altman, J M Bland BMJ 1994; 308:1552

	Pathology		
Liver scan	abnormal (+)	normal (-)	Total
abnormal (+)	231	32	263
normal (-)	27	54	81
Total	258	86	344

a) Sensitivity: _____

b) Specificity: _____

c) PPV: _____

d) NPV: _____

e) Proportion of all correct diagnosis: _____

f) What does sensitivity mean? _____

g) What is the probability of abnormal pathology test result given abnormal liver scan result? _____

2.4.2 In the following table, the observed frequencies of two diagnostic tests are summarised. Calculate sensitivity, specificity, positive and negative predictive values.

	standard test		
new test	+	-	Total
+	60	35	
-	40	65	
Total			

a) Sensitivity: _____

b) Specificity: _____

c) PPV: _____

d) NPV: _____

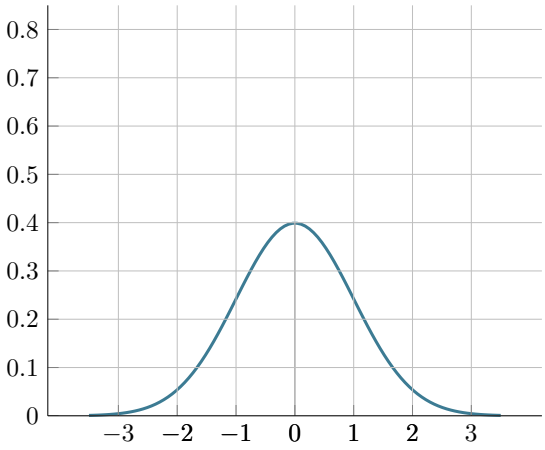
e) Proportion of all correct diagnosis: _____

f) What does positive predictive value (PPV) mean? _____

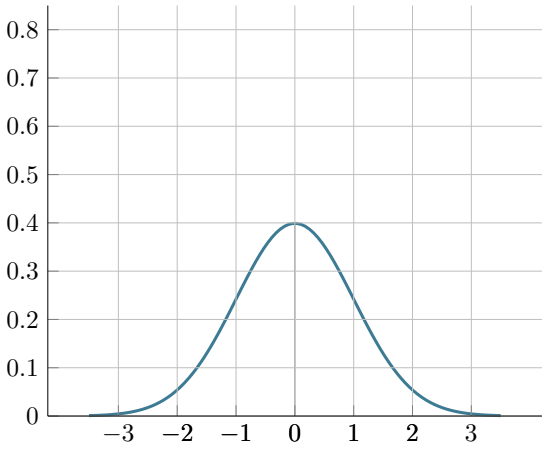
g) What is the probability of negative new test result in case of negative standard test result? _____

2.5 Normal distributions

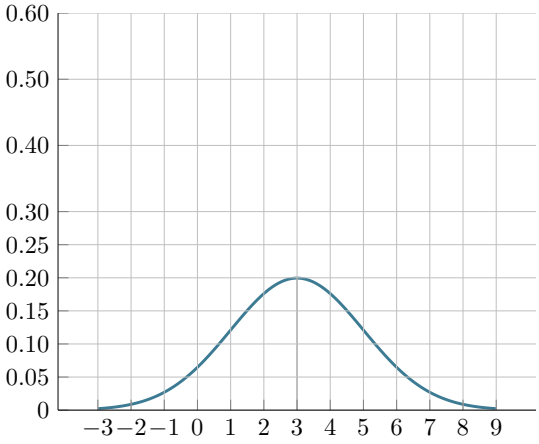
2.5.1 Draw the $\mathcal{N}(1,1), \mathcal{N}(2,1), \mathcal{N}(-1,1)$ normal distributions, if the curve of the $\mathcal{N}(0,1)$ distribution is given:



2.5.2 Draw the $\mathcal{N}(0,2^2), \mathcal{N}(0,0.5^2)$ normal distributions, if the curve of the $\mathcal{N}(0,1)$ distribution is given!



2.5.3 Draw the $\mathcal{N}(2,2^2), \mathcal{N}(1,2^2)$ normal distributions, if the curve $\mathcal{N}(3,2^2)$ is given!



2.5.4 For a standard normal distribution, find the following probabilities:

a) $P(Z < 0) =$ _____

b) $P(Z > 0) =$ _____

c) $P(Z < -1) =$ _____

d) $P(Z > 1) =$ _____

e) $P(Z < -1,96) =$ _____

f) $P(-1 < Z < 1) =$ _____

g) $P(-1,96 < Z < 1,96) =$ _____

h) $P(-2 < Z < 2) =$ _____

i) Find x value such that the area to the left of x is 0.025: _____

j) Find x value such that the area to the left of x is 0.5: _____

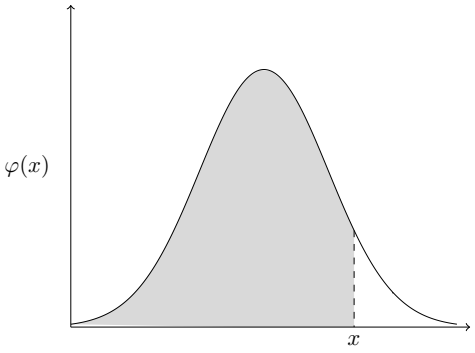
k) Which interval contains the middle 95% of the data? _____

l) Which interval contains the middle 99% of the data? _____

Table 2.1: Standard normal distribution

$\Phi(x) = P(Z < x)$ = proportion of area to the left of x (gray area)

x	$\Phi(x)$
$-\infty$	0[a]
-4.5	0.00001
-4	0.00003
-3.5	0.00023
-3	0.00135
-2.576	0.005
-2.5	0.00621
-2.326	0.01
-2	0.02275
-1.96	0.025
-1.645	0.05
-1.5	0.06681
-1	0.15865
-0.5	0.30854
0	0.5
0.5	0.69146
1	0.84135
1.5	0.93319
1.645	0.95
1.96	0.975



x	$\Phi(x)$
2	0.97725
2.326	0.99
2.5	0.99379
2.576	0.995
3	0.99865
3.5	0.99977
4	0.99997
4.5	0.99999
∞	1[a]

[a] limit

2.5.5 The results in a certain blood test performed in a medical laboratory are known to be normally distributed with $\mathcal{N}(60, 10^2)$

a) What percentage of the results are below 60? _____

What percentage of the results are above 60? _____

b) What percentage of the results are between 40 and 80? _____

What percentage of the results are below 40? _____

What percentage of the results are above 80? _____

c) The "healthy range" falls between 30 and 90.

What percentage of the results are between 30 and 90? _____

What percentage of the results are outside the healthy range of 30 to 90? _____

2.5.6 At an urban hospital the weights of new-born infants are normally distributed with $\mathcal{N}(3500, 400^2)$. Let X be the weight of a new-born picked at random. Find the following probabilities:

a) $P(X < 3500)$: _____

b) $P(3100 < X < 3900)$: _____

c) Determine the middle interval where 95% of the weights will fall.

_____

2.6 Homework

2.6.1 After performing a diagnostic test we have the following frequencies:

	standard test		
new test	+	-	Total
+	60	20	80
-	40	80	120
Total	100	100	200

Determine the asked proportions and give the name of the measure!

a) What is the proportion of people who test positive (new test +) for the disease among those who have the disease (standard test +)? _____

b) What is the proportion of people who test negative (new test -) for the disease among those who don't have the disease (standard test -)? _____

c) The proportion of the true positives among those who had positive results based on the new test?

_____

d) The proportion of the true negatives among those who had negative results based on the new test?

_____

2.6.2 A normal distribution has a mean of 30 and a standard deviation of 10. What proportion of the distribution is above 50?

2.6.3 A ball is drawn from a box containing 10 blue balls, 10 black and 5 green. What is the probability that the ball will be green given that it is not black?

3 Confidence intervals

3.1 Confidence interval for the population mean (μ) if the population standard deviation (σ) is known

3.1.1 Assume that the heights of first year pharmaceutical students are normally distributed with a mean of $\mu = 175$ and a standard deviation of $\sigma = 10$.

a) What percentage of heights are above 175 cm? _____

b) What percentage of heights are below 175 cm? _____

c) What percentage of heights are between 155 and 195? _____

d) What percentage of heights are below 155 cm? _____

e) Calculate the mean and the standard error of the mean of a sample of 36 cases derived from this population.

i) mean? _____

ii) standard error? _____

f) The mean of another random sample with 36 number of cases is 172.

i) Calculate the 95% confidence interval. _____

ii) What is the meaning of the 95% CI? _____

iii) Compare the population mean with the 95% CI calculated. Is the population mean included in the 95% CI? _____

3.2 Confidence interval for the population mean (μ) if the population standard deviation (σ) is unknown

3.2.1 (Example from Altman). In a trial we actually observed a mean serum albumin of 34.46 g/l with a standard deviation of 5.834 g/l from a sample of 21 patients with primary biliary cirrhosis.

a) Find the 95% confidence interval.

α : _____ n: _____

mean: _____ **SE:** _____

degrees of freedom: _____ t_α**:** _____

mean - t_α **SE:** _____ **mean +** t_α **SE:** _____

Confidence interval: _____

Meaning: P(< true population mean <) = 0.95

We can be 95% confident from this study that the true mean serum albumin among all such patients lies somewhere in the range _____ to _____ g/l, with 34.46 as our best estimate. This interpretation depends on the assumption that the sample of 21 patients is representative of all patients with the disease.

b) Find the 99% confidence interval and compare to the 95% CI.

α : _____ n: _____

mean: _____ **SE:** _____

degrees of freedom: _____ t_α**:** _____

mean - t_α **SE:** _____ **mean +** t_α **SE:** _____

Confidence interval: _____

Meaning: P(< true population mean <) = 0.99

c) Suppose that the above data were observed from a sample of 210 patients. Find the 95

α : _____ n: _____

mean: _____ **SE:** _____

degrees of freedom: _____ t_α**:** _____

mean - t_α **SE:** _____ **mean +** t_α **SE:** _____

Confidence interval: _____

3.2.2 In a study, systolic blood pressure of 10 healthy women was measured

Calculate the 95% confidence interval for the population mean, if the

- mean was 119, the standard error was 0.664

  _____

- mean was 119, the standard error was 2.1

  _____

Compare the length of this confidence intervals!

_____

3.2.3 Questions

a) Which is wider, a 95% or a 99% confidence interval?

_____

b) When you construct a 95% confidence interval, what are you 95% confident about?

_____

c) When computing a confidence interval, when do you use t and when do you use u?

_____

d) What does the confidence interval width depend on?

_____

3.3 Calculations with R

3.3.1 Calculate the following statistics for mass using the `quest2016.csv` !

Sample size: _____ **Mean:** _____

Standard deviation: _____ **Standard error:** _____

a) Assuming that the variable mass follows a normal distribution, determine the

i) 95% confidence interval: _____

ii) 99% confidence interval: _____

b) Can we state that the mean body mass in the population of students is 70?

_____

Explain. _____

3.3.2 Calculate the following statistics for height using the `quest2016.csv` !

Sample size: _____ **Mean:** _____

Standard deviation: _____ **Standard error:** _____

a) Assuming that the variable height follows a normal distribution, determine the

i) 95% confidence interval: _____

ii) 99% confidence interval: _____

b) Can we state that the mean height in the population of students is 170?

_____

Explain. _____

3.4 Homework

3.4.1 A researcher set a 95% confidence interval on the mean length of fish in a recreational lake and found it to be from 6.2 to 8.7 inches.

Which of the following is a proper interpretation of this interval?

(A) Of the fish in the recreational lake, 95% are between 6.2 and 8.7 inches long.

(B) We are 95% confident that the sample mean length of fish in the recreational lake is between 6.2 and 8.7 inches.

(C) We are 95% confident that the population mean length of fish in the recreational lake is between 6.2 and 8.7 inches.

(D) There is a 95% chance that a randomly selected fish from the recreational lake will be between 6.2 and 8.7 inches.

3.4.2 An industrial designer wants to determine the average amount of time it takes an adult to assemble an "easy to assemble" toy. A sample of 16 times yielded an average time of 19.92 minutes, with a sample standard deviation of 5.73 minutes. Assuming normality of assembly times, provide a 95% confidence interval for the mean assembly time.

4 One-sample, two-sample and paired *t*-tests

4.1 One-sample *t*-test for the mean of a normal population

4.1.1 The following are the systolic blood pressures (mm Hg) of $n = 9$ patients undergoing drug therapy for hypertension:

182.00 152.00 178.00 157.00 194.00 163.00 144.00 114.00 174.00

The mean = 162 mm Hg, the standard deviation SD = 23.92.

a) Find the standard error. _____

b) Find the 95% confidence interval for the population mean. _____

What is the meaning of this interval? _____

c) **We would like to test whether the sample is drawn from a population where $\mu = 130$.**

Find the null- and alternative hypothesis.

H_0: _____

H_A: _____

Based on the confidence interval Can we conclude with 95% confidence on the basis of these data that the population mean is different from 130? Explain your decision.

_____

Based on the *t* test-statistics Is the population mean significantly different from 130 at 5% level?

$t = \frac{\text{mean}-130}{SE} =$ _____

Compare its absolute value to the t-value of the table. _____

Explain your decision. _____

Based on the *p*-value The *p*-value given by R is $p = 0.004$.

Is there a significant difference from the hypothesized population mean 130 at 5% level?

_____

d) **We would like to test whether the sample is drawn from a population where $\mu = 150$.**

Find the null- and alternative hypothesis.

H_0: _____

H_A: _____

Based on the confidence interval Can we conclude with 95% confidence on the basis of these data that the population mean is different from 150? Explain your decision.

_____

Based on the *t* test-statistics Is the population mean significantly different from 150 at 5% level?

$t = \frac{\text{mean}-150}{SE} =$ _____

Compare its absolute value to the t-value of the table. _____

Explain your decision. _____

Based on the *p*-value The *p*-value given by R is $p = 0.171$.

Is there a significant difference from the hypothesized population mean 150 at 5% level?

_____

4.2 Paired *t*-test

4.2.1 The effect of saline on the blood PH was examined in a certain disease. The blood PH value was measured two times: before the treatment and 20 minutes later, after infusion of saline ($n = 18$). Is there a significant change in mean blood PH at 5% level?

0'	20'
7.43	7.43
7.39	7.39
7.37	7.38
7.43	7.42
7.39	7.39
7.36	7.41
7.38	7.38
7.39	7.39
7.34	7.41
7.32	7.35
7.40	7.39
7.32	7.33
7.42	7.39
7.42	7.4
7.37	7.36
7.37	7.39
7.39	7.37
7.43	7.48

Descriptive statistics			
	0'	20'	difference
mean	7.3844	7.3922	-0.00778
SD	0.03485	0.03264	0.02691

mean-SD diagram:

a) The name of the appropriate test: _____

b) H_0: _____

H_A: _____

c) t = _____ df = _____ critical t-value (t_α)= _____

d) Decision: _____

Conclusion: _____

e) Check your calculation using results of R.

```
> t.test(before,after,paired=TRUE)

        Paired t-test

data:  before and after
t = -1.2262, df = 17, p-value = 0.2368
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.021160128  0.005604573
sample estimates:
mean of the differences
           -0.007777778
```

i) Find the 95% confidence interval for the difference. _____

Decision based on the confidence interval: _____

ii) t = _____ df = _____

Decision based on t-value: _____

iii) p-value = _____

Decision based on p-value: _____

4.2.2 The effect of Na-lactate on the blood PH was examined in a certain disease. The blood PH value was measured two times: before the treatment and 20 minutes later, after infusion of Na-lactate (n = 20). Is there a significant change in mean blood PH at 5% level?

0'	20'
7.42	7.46
7.36	7.43
7.4	7.46
7.43	7.48
7.38	7.42
7.32	7.45
7.37	7.46
7.36	7.48
7.34	7.45
7.31	7.37
7.34	7.47
7.37	7.43
7.42	7.48
7.42	7.43
7.46	7.51
7.37	7.41
7.45	7.48
7.42	7.44
7.42	7.37
7.41	7.45

	Descriptive statistics		
	0'	20'	difference
mean	7.3885	7.4465	-0.058
SD	0.04258	0.03573	0.04336

mean-SD diagram:

a) The name of the appropriate test: _____

b) H_0: _____

H_A: _____

c) t = _____ df = _____ critical t-value (t_α)= _____

d) Decision: _____

Conclusion: _____

e) Check your calculation using results of R.

```
> t.test(before2,after2,paired=TRUE)

        Paired t-test

data:  before2 and after2
t = -5.9822, df = 19, p-value = 9.324e-06
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.07829262 -0.03770738
sample estimates:
mean of the differences
                 -0.058
```

i) Find the 95% confidence interval for the difference. _____

Decision based on the confidence interval: _____

ii) t = _____ df = _____

Decision based on t-value: _____

iii) p-value = _____

Decision based on p-value: _____

4.3 Two-sample *t*-test

4.3.1 In a medical study[1] two treatments were applied to back pain on 60-60 randomly assigned patients. The following table contains demographic and clinical characteristics. The continuous variables were compared using two-sample *t*-test. Check the *p*-values and the assumptions of the hypothesis test using the `independent_ttest_n_mean_sd.R` file!

		Group I (60)	Group II (60)	*P* value
Gender	Male	23% (14)	40% (24)	0.077
	Female	77% (46)	60% (36)	
Age	Mean ± SD	41.2 ± 11.9	42.7 ± 11.4	0.477
Weight	Mean ± SD	211.2 ± 60.9	168.6 ± 40.6	**0.000**
Height	Mean ± SD	65.8 ± 3.7	66.4 ± 4.1	0.430
Duration of Pain (months)	Mean ± SD	104.2 ± 106.5	129.0 ± 90.9	0.173
Onset of Pain	Gradual	67% (40)	70% (42)	0.845
	Injury	33% (20)	30% (18)	
Pain Distribution	Unilateral	20% (12)	25% (15)	0.662
	Bilateral	80% (48)	75% (45)	
Back Pain Distribution	Back pain only	15% (9)	20% (12)	0.849
	Back pain worse than leg pain	65% (39)	60% (36)	
	Leg pain worse than back pain	5% (3)	3% (2)	
	Both equal	15% (9)	17% (10)	
Numeric Rating Score	Mean ± SD	8.0 ± 1.0	7.7 ± .9	0.082
Oswestry Disability Index	Mean ± SD	30.7 ± 4.5	29.2 ± 5.2	0.096

*Multiple patients presented with disc herniation at more than one level.

4.3.1.1 Age variable

a) The name of the appropriate test: _____
Why we can't apply paired *t*-test? _____

b) H_0: _____
H_A: _____

c) Assumptions of the test: _____
Normality fulfillment: _____
Equal variances fulfillment: _____

d) *Descriptive statistics*: _____

e) Is there a significant difference? _____ Why? _____

f) Check the *p*-value with R. Was the chosen method appropriate? _____

4.3.1.2 Duration of Pain variable

a) The name of the appropriate test: _____
Why we can't apply paired *t*-test? _____

b) H_0: _____
H_A: _____

c) Assumptions of the test: _____
Normality fulfillment: _____
Equal variances fulfillment: _____

d) *Descriptive statistics*: _____

e) Is there a significant difference? _____ Why? _____

f) Check the *p*-value with R. Was the chosen method appropriate? _____

[1] Manchikanti L, Cash K, McManus C, Pampati V and Benyamin R. Randomized, Double-Blind, Active-Controlled Trial of Fluoroscopic Lumbar Interlaminar Epidural Injections in Chronic Axial or Discogenic Low Back Pain: Results of 2-Year Follow-Up. Pain Physician 2013; 16:E491-E504 ISSN 2150-1149.

4.3.2 Compare beers

Using `BEER.csv` datafile, compare the calorie content and prices of *LIGHT* and *NONLIGHT* american beers (variables LIGHT and CALORIES)! The population means of light and nonlight beers are the same?

4.3.2.1 Calories

a) Applied test:

b) H_0: _____ H_A: _____

c) Assumption(s) of the test: _____

Fulfillment(s)? _____

d) Descriptive statistics

	light	nonlight
sample size		
mean		
SD		
SE		

e) *Result*

Test-statistics: _____ df: _____ p-value: _____

Is there a significant difference?? _____ Why? _____

4.3.2.2 Prices

a) Applied test:

b) H_0: _____ H_A: _____

c) Assumption(s) of the test: _____

Fulfillment(s)? _____

d) Descriptive statistics

	light	nonlight
sample size		
mean		
SD		
SE		

e) *Result*

Test-statistics: _____ df: _____ p-value: _____

Is there a significant difference?? _____ Why? _____

4.4 Calculations with R

4.4.1 Open the file `befafter.csv`. A study was conducted to determine weight loss, body composition, etc. in obese women before and after 12 weeks of treatment with a very-low-calorie diet. Column BEFORE and AFTER contain weights of 9 women. We wish to know if these data provide sufficient evidence to allow us to conclude that the treatment is effective in causing weight reduction in obese women. Let $\alpha = 0.05$

a) The name of the appropriate test: _____

H_0: _____

H_A: _____

	mean	SD
Before		
After		
difference		

95% CI for the difference: _____

t = _____ degrees of freedom = _____ t_α = _____

p-value = _____

b) Decision: _____

4.4.2 Open the file `quest2016.csv` !

Compare the mean body mass of boys and girls (at 5% level).

a) The name of the appropriate test: _____

b) H_0: _____

H_A: _____

c) Assumptions: _____

	boys	girls
sample size		
mean		
SD		
SE		

Equality of variances $\alpha = 5\%$

a) *p*-value: _____

b) Decision about the equality of variances:

_____

Equality of population means $\alpha = 5\%$

a) $t =$ _____ df? _____ $p =$ _____

b) 95% CI of the difference: _____

c) Decision about the equality of population means: _____

Conclusion: _____

4.5 Homework

4.5.1 The body mass of 16 patients was measured before and after a special diet. The mean of the sample differences is 5 kg, the standard deviation of the differences is 2.5. Is there a significant change in body mass at 5% and at 1% level?

$\alpha = 5\%$

a) Appropriate test: _____

H_0 : _____ H_A: _____

b) t: _____

degrees of freedom: ____ critical value: ____

c) Decision: _____

Conclusion: _____

$\alpha = 1\%$

a) Appropriate test: _____

H_0 : _____ H_A: _____

b) t: _____

degrees of freedom: ____ critical value: ____

c) Decision: _____

Conclusion: _____

4.5.2 Open the file `LWTBWT.csv` and compare the mean body weight of newborn babies (variable BWT) by smoking habits of the mother (variable SMOKE 0 no, 1 yes)

4.5.3 Open the file `ANTHROPOMETRICS.csv` and compare the body height of boys and girls. Find other variables to be compared and find the appropriate test.

4.5.4 Open the file `CALC.csv`! Here systolic blood pressures are given before and after a calcium treatment in two groups. Find problems where paired *t*-tests can be used. Find problems where two-sample *t*-tests can be used.

4.5.5 Open the file `NEWDRUG.csv` and find problems where paired *t*-tests can be used. Find problems where two-sample *t*-tests can be used.

5 Correlation and regression

5.1 In a dataset measuring people's height and mass, let's examine the relationship between height and mass.

a) Based on the plot, what would you say about the linear relationship?

- direction: _____
- strength: _____

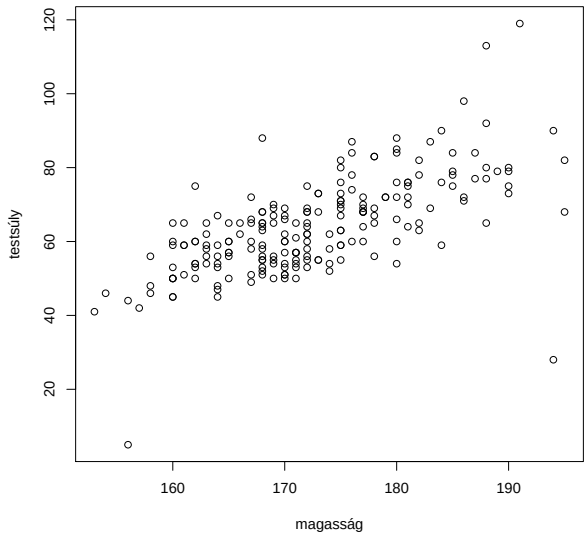
b) The correlation coefficient calculated from the data is $r = 0.6556$.

What does the ... of the correlation coefficient indicate?

- absolute value: _____
- sign: _____

c) How would you describe this linear relationship based on the correlation coefficient? _____

d) Why do we need to test whether the correlation coefficient is significant? _____



Significance test for the correlation coefficient ($n = 206$)

e) Null hypothesis: _____

Alternative hypothesis: _____

f) t-value of the correlation: _____

g) Degrees of freedom: _____ Critical value: (t_α): _____

h) Decision and conclusion: _____

i) The equation of the regression line: $y = 0,9698 \cdot x - 103,147$ (mass = $0,9698 \cdot$ height $- 103,147$)

Meaning: _____

What mass "belongs" to someone whose height is 160 cm, based on the regression line? _____

j) Coefficient of determination

$$r^2 = \frac{\text{squares}_{Regression}}{\text{squares}_{Total}} = 0.4299$$

Meaning: _____

```
> cor.test(indep,dep)

        Pearson's product-moment correlation

data:  indep and dep
t = 12.4003, df = 204, p-value < 2.2e-16
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.5699674 0.7271248
sample estimates:
      cor
0.6555883

> lm(dep~indep)

Call:
lm(formula = dep ~ indep)

Coefficients:
(Intercept)        indep
  -103.1470       0.9698

> fit = lm ( dep ~ indep )
> summary ( fit )

Call:
lm(formula = dep ~ indep)

Residuals:
    Min      1Q  Median      3Q     Max
-56.986  -5.681  -0.485   5.273  36.923

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -103.1470    13.5084  -7.636 8.51e-13 ***
indep          0.9698     0.0782  12.400  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 9.946 on 204 degrees of freedom
Multiple R-squared:  0.4298,    Adjusted R-squared:  0.427
F-statistic: 153.8 on 1 and 204 DF,  p-value: < 2.2e-16
```

Find these values in the R output.

Significance tests for the coefficients of the regression line

k) Null hypothesis of the regression coefficient: _____ *t*-value: _____ df: _____
p-value: _____ significance: _____

l) Null hypothesis of the constant term: _____ *t*-value: _____ df: _____
p-value: _____ significance: _____

5.2 A British survey examined the connection of money spent on alcohol and tobacco in 10 different regions of GBR. The correlation coefficient of money spent on alcohol and tobacco is $r = 0.784$. Test whether this is significantly different from 0 at 5% level. (`S:/R/alcohol.csv`)

Hand calculation

a) H_0: _____

H_A: _____

Assumption(s): _____

b) t-value: _____ df: _____ Critical value (t_α): _____

c) Decision: _____

Conclusion: _____

We ran the previous example in R. Interpret the results.

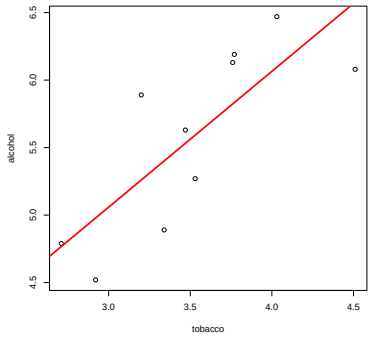
```
> cor.test(alcohol,tobacco)

        Pearson's product-moment correlation

data:  alcohol and tobacco
t = 3.5756, df = 8, p-value = 0.007234
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.3055382 0.9465163
sample estimates:
      cor
0.7842873
> lm(tobacco~alcohol)

Call:
lm(formula = tobacco ~ alcohol)

Coefficients:
(Intercept)      alcohol
     0.1082       0.6115
```



d) Correlation coefficient (r): _____ Coefficient of determination r^2: _____

e) t:_____ df: _____ p-value: _____

f) Conclusion to the direction and strength of the linear relationship: _____

g) Equation of the regression line: _____

h) Significance of the regression coefficients (use R to answer this question): _____

5.3 Is the correlation coefficient significantly different from 0 at 5% level?

H_0: _____ H_A: _____

5.3.1 Based on $n = 5$ pairs of observations, the correlation coefficient is $r = 0.8$.

a) t-value: _____ df: _____ Critical value (t_α): _____

b) Decision: _____

Interpretation (is there a contradiction?): _____

5.3.2 Based on $n = 500$ pairs of observations, the correlation coefficient is $r = 0.2$.

a) t-value: _____ df: _____ Critical value (t_α): _____

b) Decision: _____

Interpretation (is there a contradiction?): _____

5.3.3 Based on $n = 15$ pairs of observations, the correlation coefficient is $r = -0.8$.

a) t-value: _____ df: _____ Critical value (t_α): _____

b) Decision: _____

Interpretation: _____

5.4 Practice with R

5.4.1 Open the `quest2016.csv` data file. Examine the relationship between height and mass. Create a scatter plot.

a) Correlation coefficient r: _____ Coefficient of determination r^2: _____

b) t: _____ df: _____ p: _____

Decision: _____

c) Equation of the regression line: _____

d) Significance of the regression coefficients: _____

5.4.2 Open the data file `anthropometrics.csv`! Examine the connection between the first and second measurement of weight. Create a scatter plot.

a) Correlation coefficient r: _____ Coefficient of determination r^2: _____

b) t: _____ df: _____ p: _____

Decision: _____

c) Equation of the regression line: _____

d) Significance of the regression coefficients: _____

5.4.3 Open the data file `anthropometrics.csv`! Examine the linear association between height and hip circumference. Create a scatter plot.

a) Correlation coefficient r: _____ Coefficient of determination r^2: _____

b) t: _____ df: _____ p: _____

Decision: _____

c) Equation of the regression line: _____

d) Significance of the regression coefficients: _____

6 χ^2-test for independence

6.1 In a (hypothetical) study, the relationship between gender and participation in annual medical check-ups was examined. The results are summarized in the following 2×2 contingency table (observed frequencies).

Observed frequencies

GENDER	Medical check-ups within the last 1 year: yes	no	Total
male	15	40	55
female	25	20	45
Total	40	60	100

Expected frequencies

GENDER	Medical check-ups within the last 1 year: yes	no	Total
male			55
female			45
Total	40	60	100

Is participation in routine medical check-ups independent of gender at the 5% significance level?

a) *Participation rate among*

males: _____ females: _____

b) The name of the appropriate test _____

c) H_0: _____

H_A: _____

Assumption(s) of the test: _____

$$\chi^2 = \sum_i \frac{(\text{observed}_i - \text{expected}_i)^2}{\text{expected}_i} =$$

d) Degrees of freedom: (number of rows - 1)·(number of columns - 1)= _____ Critical value: _____

e) Decision: _____

Conclusion: _____

6.2 Two medicines are being compared regarding a particular side effect, 60 similar patients are split randomly into two groups, one on each drug. The results are presented in the observed frequencies table

Observed frequencies

DRUG	SIDE EFFECT: yes	no	Total
A	10	20	30
B	5	25	30
Total	15	45	60

Expected frequencies

DRUG	SIDE EFFECT: yes	no	Total
A			30
B			30
Total	15	45	60

Are drug type and the occurrence of side effect independent at the 5% level?

a) *Percentage of people who had side effect taking*

drug A: _____ drug B _____

b) The name of the appropriate test: _____

c) H_0: _____

H_A: _____

Assumption(s) of the test: _____

$$\chi^2 = \sum_i \frac{(\text{observed}_i - \text{expected}_i)^2}{\text{expected}_i} =$$

d) Degrees of freedom: (number of rows - 1)·(number of columns - 1)= _____ Critical value: _____

e) Decision: _____

Conclusion: _____

6.3 In a certain town, there are about 10 000 eligible voters. A random sample of 500 eligible voters was chosen to study the relationship between age and participation in the last election. The results are summarized in the following 3×2- contingency table

Observed frequencies

AGE GROUP	PARTICIPATION yes	no	Total
Under age 40	90	60	150
40 – 60	130	70	200
Above age 60	120	30	150
Total	340	160	500

Expected frequencies

AGE GROUP	PARTICIPATION yes	no	Total
Under age 40			150
40 – 60			200
Above age 60			150
Total	340	160	500

Is there a relationship between age group and having voted in the last election ($\alpha = 0.05$)?

a) *Participation rate*

Under age 40: _____ Between age 40 and 60: _____ Above age 60: _____

b) The name of the appropriate test: _____

c) H_0: _____

H_A: _____

Assumption(s) of the test: _____

$$\chi^2 = \sum_i \frac{(\text{observed}_i - \text{expected}_i)^2}{\text{expected}_i} =$$

d) Degrees of freedom: (number of rows - 1)·(number of columns - 1)= _____ Critical value: _____

e) Decision: _____

Conclusion: _____

6.4 The following observed frequency table shows the results of placebo and aspirin treatments in an experiment, with the number of people in each treatment group who did and did not develop thrombi. Decide whether the aspirin had an effect on thrombus formation.

Observed frequencies

	Developed thrombi? yes	Nem	Total
Placebo	16	4	20
Aspirin	6	14	20
Total	22	18	40

Expected frequencies

	Developed thrombi? yes	Nem	Total
Placebo			20
Aszpirin			20
Total	22	18	40

a) *The thrombus formation rate in the...*

placebo group: _____ aspirin group: _____

b) The name of the appropriate test: _____

c) H_0: _____

H_A: _____

Assumption(s) of the test: _____

d) Test-statistic:

$$\chi^2 = \sum_i \frac{(\text{observed}_i - \text{expected}_i)^2}{\text{expected}_i} =$$

e) Degrees of freedom: (number of rows - 1)·(number of columns - 1)= _____ Critical value: _____

f) Decision: _____

Conclusion: _____

6.5 Examine the relationship between the answers of pleased and gender. Are the answers to the question "Are you pleased with using a statistical software?" independent of gender? Use the `quest2016.csv` dataset!

Observed frequencies

GENDER	PLEASED yes	PLEASED no	Total
male			
female			
Total			

Expected frequencies

GENDER	PLEASED yes	PLEASED no	Total
male			
female			
Total			

a) The name of the appropriate test: _____

b) H_0: _____

H_A: _____

Assumption(s) of the test: _____

c) χ^2 = _____ Degrees of freedom= _____ p = _____

d) Fisher's exact p-value: _____

e) Decision: _____

Conclusion: _____

6.6 Examine whether the answers to the question "Is biostatistics difficult" depend on gender.

Observed frequencies

GENDER	DIFFICULT yes	DIFFICULT no	Total
male			
female			
Total			

Expected frequencies

GENDER	DIFFICULT yes	DIFFICULT no	Total
male			
female			
Total			

a) The name of the appropriate test: _____

b) H_0: _____

H_A: _____

Assumption(s) of the test: _____

c) χ^2 = _____ Degrees of freedom= _____ p = _____

d) Fisher's exact p-value: _____

e) Decision: _____

Conclusion: _____

6.7 Examine whether answers to the question "How do you like to eat?" depend on gender.

Observed frequencies

EATING	GENDER male	GENDER female	Total
I don't like to eat at all			
I don't like to eat			
Indifferent			
I like to eat			
I like to eat very much			
Total			

Expected frequencies

EATING	GENDER male	GENDER female	Total
I don't like to eat at all			
I don't like to eat			
Indifferent			
I like to eat			
I like to eat very much			
Total			

a) The name of the appropriate test: _____

b) H_0: _____

H_A: _____

Assumption(s) of the test: _____

c) χ^2 = _____ Degrees of freedom= _____ p = _____

d) Fisher's exact p-value: _____

e) Decision: _____

Conclusion: _____

7 Agreement, odds ratio and relative risk

7.1 Measure agreement (kappa)

Two people (MN és VU) qualified psychiatric patients. Do they agree?
Calculate the κ (kappa) statistics and interpret the reults!

Source: Wongpakaran et al.: A comparison of Cohen's Kappa and Gwet's AC1 when calculating inter-rater reliability coefficients: a study conducted with personality disorder samples, BMC Research Methodolgy 2013,13:61.
http://www.biomedcentral.com/1471-2288/13/61

7.1.1 Avoidant

$\kappa =$_____ agreement: _____

7.1.2 Dependent

$p_o = 0.9474$, $p_e = 0.8532$

$\kappa =$_____ agreement: _____

7.1.3 Depressive

$p_o = 0.8947$, $p_e = 0.7431$

$\kappa =$_____ agreement: _____

7.1.4 Schizoid

$p_o = 0.8421$, $p_e = 0.6371$

$\kappa =$_____ agreement: _____

7.1.5 Total antisocial

$p_o = 0.8947$, $p_e = 0.8116$

$\kappa =$_____ agreement: _____

		Rater VU	
	Rater MN	NO	YES
Avoidant	NO	14	0
	YES	0	5
Dependent	NO	17	1
	YES	0	1
Depressive	NO	15	1
	YES	1	2
Schizoid	NO	13	2
	YES	1	3
Total antisocial	NO	16	1
	YES	1	1

7.2 Odds ratio (OR)

Children's drug use and risk factors were examined in a retrospective (case-control) study. Examine and interpret the odds ratios in the following table and determine the significance based on the confidence intervals

Table 2. Results of the univariate analysis in the ever-smoked and regular-smoker groups

		Children	Drug users	OR (95% CI)	p value
Living in a block of flat	Yes	71	14	1.8 (0.9–3.7)	0.086
	No	263	31	1.0	
Age, years	17–18	107	23	2.3 (1.2–4.6)	0.014
	15–16	171	18		
Sociable delinquencies	Yes	129	28	3.4 (1.7–6.7)	<0.001
	No	186	14	1.0	
School performance	Poor	17	6	15.0 (2.7–84.5)	0.002
	Acceptable	117	17	4.8 (1.0–21.0)	0.044
	Good	144	20	4.4 (1.0–19.7)	0.050
	Very good	57	2	1.0	
Truancy from school	Yes	50	13	3.3 (1.5–7.3)	0.003
	No	210	20	1.0	

Source: Nyári T.A., Herédi K. and Parker L.: Addictive Behaviour of Adolescents in Secondary Schools in Hungary. Eur Addict Res 2005;11:38– 3 DOI:10.1159/00008141

Why did they calculate OR? _____

7.2.1 Living in a block flat vs drug usage

Table interpretation: in first row one can see that 14 children out 71 use drugs. In the second row 31 use druges out of 263. Create a 2×2 table and calculate the OR.

	DRUG		
Block flat	**yes**	**no**	Total
yes			71
no			263

a) The probability child use drug is

- when lives in block flat: _____
- when does not live in block flat: _____

b) Quotient (OR)= _____

Using the formula: _____

Interpretation: _____

c) 95% confidence interval: _____

Significance of odds ratio

d) H_0: _____

H_A: _____

e) Decision about significance based on CI: _____

Decision based on *p*-value: _____

7.2.2 School performance. Compare the good and very good category with the drug usage. Calculate OR!

SCHOOL PERFORMANCE	DRUG yes	DRUG no	Total
good			
very good			

a) The probability that a child use drug if s/he is

- good: _____
- very good: _____

b) Quotient (OR)= _____

Using the formula: _____

Interpretation: _____

c) 95% confidence interval: _____

Significance of odds ratio

d) H_0: _____

H_A: _____

e) Decision about significance based on CI: _____

Decision based on *p*-value: _____

7.2.3 Truancy from school and drug usage Create 2×2-es table and calculate OR.

TRUANCY	DRUG yes	DRUG no	Total
yes			
no			

a) The probability that a child use drug if he

- plays truant: _____
- does not play truant: _____

b) Quotient (OR)= _____

Using the formula: _____

Interpretation: _____

c) 95% confidence interval: _____

Significance of odds ratio

d) H_0: _____

H_A: _____

e) Decision about significance based on CI: _____

Decision based on *p*-value: _____

7.3 Relative risk (RR)

In a prospective study risks of respiratory complications were examined.

	History of respiratory problems*		RR (95% CI)	p value†	Absolute risk reduction (95% CI)
	No (n=7041)	Yes (n=2256)			
During surgery					
Bronchospasm	30 (0%)	133 (6%)	13·84 (9·34 to 20·50)	<0·0001	5·47% (4·49 to 6·45)
Laryngospasm	142 (2%)	180 (8%)	3·96 (3·19 to 4·90)	<0·0001	5·96% (4·80 to 7·13)
Cough	267 (4%)	286 (13%)	3·34 (2·85 to 3·92)	<0·0001	8·88% (7·44 to 10·33)
Desaturation <95%	373 (5%)	389 (17%)	3·25 (2·85 to 3·72)	<0·0001	11·94% (10·30 to 13·59)
Airway obstruction	130 (2%)	136 (6%)	3·27 (2·58 to 4·13)	<0·0001	4·18% (3·15 to 5·21)
Any‡	584 (8%)	595 (26%)	3·18 (2·87 to 3·53)	<0·0001	18·08% (16·15 to 20·01)

Source: Britta S von Ungern-Sternberg és mtsai: Risk assessment for respiratory complications in paediatricanaesthesia: a prospective cohort study. The Lancet, Vol 376 September 4, 2010, 773-783

Why did they calculate relative risk? _____

Examine using the table whether a respiratory problems (=yes) causes more frequent brochospasmus complication! Calculate a 2×2-es táble and RR!

RESPIRATORY PROBLEMS	**COMPLICATION** **yes**	**no**	Total
yes	133		2256
no	30		7041

a) The risk of having complication if patient

- had respiratory problems before: _____
- did not have respiratory problems before: _____

b) Quotient (RR)= _____

Using the formula: _____

Interpretation: _____

c) 95% confidence interval of the relative risk: _____

Significance of the relative risk

d) H_0: _____

H_A: _____

e) Decision about significance based on CI: _____

Decision based on *p*-value: _____

8 Type of problems not practiced during the semester

8.1 Fourfold (2×2) tables

1. In a study of measurement of agreement, the observed and expected probabilities were 0.85 and 0.5, respectively. Calculate the kappa statistic!
2. In a study 40 HPV positive tests of 50 abnormal cervical samples and 10 HPV positive tests of 60 normal cervical samples were detected. Calculate the odds ratio!
3. The risk of HPV infection for smokers was measured in a study. The calculated odds ratio was 1.58 with 95% CI [1.061 - 2.398]. We decide...
4. The risk of HPV infection for smokers was measured in a study. The calculated odds ratios was 1.58 with 95% CI [0.961 - 2.598]. We decide ...
5. In a study 8 HPV positive tests of 20 abnormal cervical samples and 10 HPV positive tests of 20 normal cervical samples were detected. Calculate the odds ratio!

8.2 Nonparametric tests

1. Find the ranks to the following data: 199, 126, 81, 68, 112, 112.

8.3 Survival analysis

1. In a study the average period of time of free of disease was 3.1 year and SE=0.44. Compare this result to the reference 2.2 years of survival rate using a 95% confidence interval!
2. In a study the average period of time of free of disease was 3.1 year and SE=0.44. Compare this result to the reference 2.2 years of survival rate! What is the null hypothesis ?
3. In a cohort study the first year the interval survival rate was 0.99. In the following four years the annual interval survival rates were 0.98, 0.97, 0.96 and 0.95, respectively. Calculate the 5-year cumulative survival rate!
4. In a cohort study the first year the interval survival rate was 0.90. In the following four years the annual interval survival rates were 0.90, 0.90, 0.90 and 0.90, respectively. Calculate the 5-year cumulative survival rate!

9 Summary of the methods

1. both are continuous (measured on the same subject):
   a) comparing the means of variables – the same thing about the same subjets, examining mean change: **paired t-test**
   b) examining relationships between variables: **correlation, regression**
2. one continuous dependent variable divided into independent groups according to another, categorical variable: (that is, comparing means of groups)
   a) number of groups=2: **two-sample t-test** (Independent t-test)
   b) number of groups>2: **One-way ANOVA** (Analysis of Variance)
3. both are categorical: examination of contingency tables, χ^2 **test**

In 1. and 2. we assumed that the samples come from normal distribution. If this assumption does not hold or we have ordered data, use **nonparametric methods based on ranks**.

9.1 Manual calculation

9.1.1 Effect of a new drug

To test the effect of a new drug, the temperature was measured on the same 10 patients before and after the treatment. The mean of the differences is = 0.6, the standard deviation of the differences is SD=0.4062. Test whether the drug is effective or not; i.e., test whether the change in mean is different from 0 at α=0.05 level.

a) To test the effect of the drug, what is the appropriate test?_____
   Assumption(s)?_____

b) State the null and the alternative hypothesis
   H_0: _____
   H_A: _____

c) Find the degrees of freedom: _____ standard error: _____
   test statistic: _____ critical value: _____

d) Decide whether the difference is significant or not: _____
   State a conclusion: _____

e) *Alternative solution using confidence interval*
   Find the 95% confidence interval: _____
   Decide whether the difference is significant or not: _____
   Conclusion: _____

9.1.2 Drug side effect

Two medicines are being compared regarding a particular side effect; 100 similar patients are split randomly into two groups, one of each group. In the group of drug A, 10 side effects were observed while in the group of drug B only 2 side effects were observed. Test whether drug and side effects are independent!

DRUG	SIDE EFFECTS yes	no	
A			
B			

a) Give the percentage of side effects in

group A: _____ group B: _____

b) What is the appropriate test?

_____

Assumption(s) of the test:

_____

c) State the null and the alternative hypothesis

H_0: _____
H_A: _____

d) Find the test statistic: _____

e) Find the degrees of freedom: _____ Critical value ($\alpha = 0.05$): _____

f) Decide whether the difference is significant or not:

_____

g) State a conclusion:

_____

9.1.3 Association

Based on 11 pairs of data, the coefficient of correlation is r=0.8. Is the correlation significant at 5% level? What is your opinion about the direction and about the strength of the association?

a) State the null and the alternative hypothesis

H_0: _____
H_A: _____

b) Find the test statistic: _____

c) Find the degrees of freedom: _____ Critical value: _____

d) Decision

_____

e) Interpretation

_____

9.2 Calculation using R

Open the file `quest2016.csv`. This file contains data of first year medical students.

9.2.1 Examine the relationship between ideal body mass and body mass

Examine the relationship between the body mass (mass) and the ideal body mass (ideal_mass)! Let the present mass be the independent variable. Is there a linear relationship between body mass at present and ideal body mass?

a) The name of the appropriate test: _____

Assumption(s) of the test: _____

b) State the null and the alternative hypothesis

H_0: _____

H_A: _____

c) $r =$ _____ it's meaning: _____

p-value: _____

d) Significance (decision): _____

Explain your result in the context of the problem: _____

_____

e) Equation of the line: _____

Based on the equation, find the ideal body mass to a 60 kg (actual) mass: _____

_____

9.2.2 Mean change of mass

Students were asked about they body mass at present (variable mass) and body weight 3 years ago (mass3). Find whether the mean change of mass is significant or not at 5% level.

a) The name of the appropriate test: _____

b) Assumption(s) of the test: _____

_____

c) State the null and the alternative hypothesis

H_0: _____

H_A: _____

d) *Descriptive statistics*

variable	mean	standard deviation	sample size
mass			
mass3			

e) *Result of the test*

test statistic: _____ degrees of freedom: _____ p -value: _____

Decision: _____

f) Explain your result in the context of the problem:

_____

9.2.3 Age of boys and girls

Compare the mean age of boys and girls! (variables age, gender)

a) The name of the appropriate test: _____

b) Assumption(s) of the test: _____

_____

c) State the null and the alternative hypothesis

H_0: _____
H_A: _____

d) *Descriptive statistics*

	mean	standard deviation	sample size
boy			
girl			

e) *Result of the test*

test statistic: _____ degrees of freedom: _____ p -value: _____

Decision: _____

f) Explain your result in the context of the problem:

_____

g) (Result of the test of equality of variances: _____)

9.2.4 Eye color and gender

Examine the relationship between the answers of eye color and gender. Is the gender and eye color of students independent (variable gender and eye)?

a) What is the appropriate test? _____

b) State the null and the alternative hypothesis

H_0: _____
H_A: _____

c) *Descriptive statistics:*

d) What is the assumption of the test? _____

Does it come true? _____

e) Find the degrees of freedom: _____ the test statistic: _____ critical value: _____ p-value: _____

f) State a conclusion: _____