

## Biometry practical 5

### Illustrated (imperfect) practical guide

#### Preparatory work

1. Open in *MS Excel* the questionnaire data (file analysed already in previous practicals),
  2. insert new worksheet, rename it as 'Praks5' (or 'Practical5') and
  3. make a copy of the data table (from worksheet 'Andmed') and paste it into the upper left corner of the new worksheet.
- 

#### Exercise 1.

**Are the students' height and shoe size related?** Study this using *MS Excel* functions.

- Calculate the correlation coefficient between variables 'HEIGHT' and 'SHOE\_SIZE';
- describe the relationship on the basis of calculated coefficient;
- test the statistical significance of the relationship:
  - formulate the null- and alternative hypothesis,
  - test, which of these hypothesis is true (find the sample size  $n$  and teststatistic  $t$ , and calculate on the basis of these values significance probability  $p$ ),
  - phrase the final conclusion.

#### Exercise 2.

Illustrate the relationship between variables 'HEIGHT' and 'SHOE\_SIZE' with scatterplot.

#### Exercise 3.

- Calculate correlation coefficients between all continuous variables in dataset (height -- shoe size) using statistical procedure *Correlation (Data-tab -> Data analysis... -> Correlation)*.
  - Between which variables is the strongest relationship? But the weakest?
  - Describe some correlations (write down the sentences describing both the strength and the direction of relationships).
-

## Exercise 1 guide

1. As the result of *MS Excel* functions is usually only one non-commented value, it is useful to write down before calculations what will be calculated.

For example, at the present moment the task is to calculate the correlation coefficient between height and shoe size – into *Excel* worksheet should be typed


- **'Linear correlation coefficient between height and shoe size'**
- or more shortly **'r(Height;Shoe\_size)'**, as the linear correlation coefficient is usually denoted with letter **'r'**.

After that put the cursor into empty cell where you want to calculate the correlation coefficient.

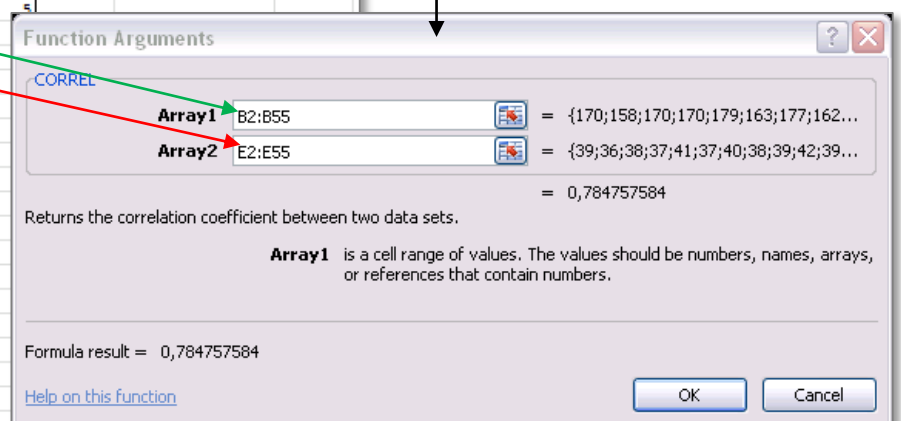
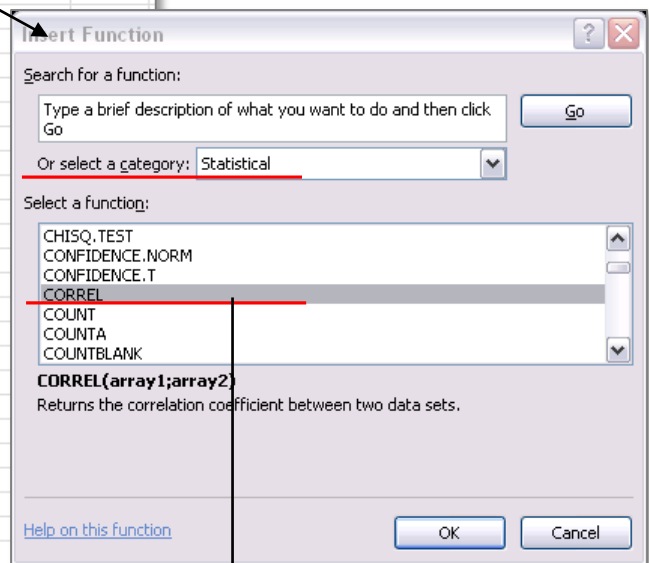
2. Linear correlation coefficient is calculable with function CORREL, which has two arguments – the range of values of the first variable and the range of values of the second variable.

- More experienced *Excel* users can type the appropriate command yourself:

=CORREL (B2 : B55 ; E2 : E55)

- Less experienced students (who did not understand the previous formula) should click on button  and continue according to the scheme.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R
1	GENDER	HEIGHT	WEIGHT	HEAD	SHOE_SIZ	MATH												r(Height;Shoe_size)
2	W	170	70	55,5	39													
3	W	158	47,5	55	36													
4	W	170	60	53	38													
5	W	170	50	55	37													
6	W	179	68	58	41													
7	W	163	56	37	4													
8	W	177	65	55	40													
9	W	162,5	53	55	38													
10	W	170	75	56	39													
11	M	175	74	57	42													
12	W	176	66	57	39													
13	M	175	64	56	42	4	ked											
14	M	190	82	58	46													
15	W	161	50	55	37													
16	W	170	85	57	41													
17	W	176	58	52	39													
18	W	172	90	58	41													
19	W	158	55	57	38													
20	M	189	82	43	4													
21	W	169	60	55,5	41													
22	W	164	52	56	37													
23	W	172	62	56	39													
24	W	173	66	56	40													
25	W	169	60	55	39													
26	W	162	50	50	38													
27	W	165	52	50,5	37													
28	M	170	80	56	41													
29	M	176	74	56	42													
30	M	175	73	54	43													
31	W	171	63	57	39													
32	W	170	60	53	39													
33	W	163	62	55	38													
34	M	181	74	55	44													
35	W	168	60	55	39													
36	W	174	54	55	40													
37	W	166	68	56	39													
38	W	168	63	53	39													
39	W	165	58	56	37													
40	W	171	75	55	41													
41	W	165	77	58	39													
42	W	161	55	57	38													
43	M	183	75	43														
44	W	169	53	55	38													
45	W	175	60	57	42													
46	W	167	80	57,5	41													
47	W	158	70	55	38													
48	M	174	87	57	40													
49	W	165	61	57	39													
50	W	164	58	57	39													
51	W	185	80	60	41													
52	W	177	63	60	40													
53	W	160	70	57	39													
54	W	162	70	55	40													
55	W	172	58	62	39													



3. Describe the relationship between students' height and shoe size:

- how strong (weak / intermediate / strong),
- Positive or negative (what this positive or negative means?).

NB! This conclusion follows from the positivity/negativity of the relationship! Only the word "bigger" or "smaller" is necessary to fill the gap in text.

<b>r(Height;Shoe_size)</b>	0,784758
There is a ..... relationship between height and shoe size.	
This means, that to bigger height corresponds ..... shoe size on an average.	
Hypothesis pair	
H <sub>0</sub> : Height and shoe size are not related (or mathematically $r = 0$ )	
H <sub>1</sub> : Height and shoe size are related (or mathematically $r \neq 0$ )	

4. Write down the hypothesis pair also in text form.

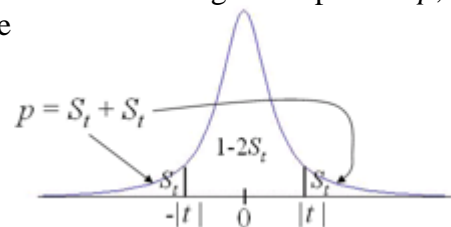
**Reminder from theory – hypothesis testing about correlation coefficient**

To test in *Excel*, is the correlation coefficient different from zero (is the relationship statistically significant), at first the absolute value of teststatistic (which in case of null hypothesis follows the t-distribution) must be calculated by the formula

$$t = r\sqrt{n-2} / \sqrt{1-r^2} \underset{H_0}{\sim} t_{n-2}.$$

Quantity  $r$  in this formula is the correlation coefficient and  $n$  is the sample size (number of students whose height and shoe size were both known).

The decision, which of the hypothesis is true, will be made according to the p-value  $p$ , which is calculated as the sum of the areas under the tails of teststatistic's distribution (denoted as  $S_t$  in figure).



In *Excel* the p-value is calculable with function `T.DIST.2T (ABS (t) ; n-2)`.

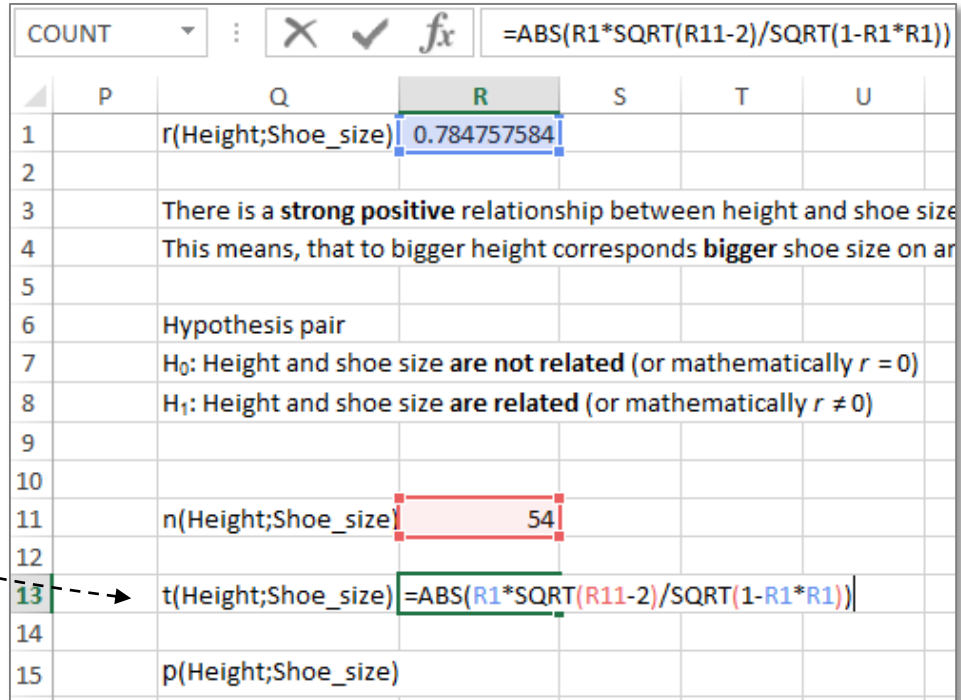
5. The evaluation of significance probability (p-value) is easier to perform, if all necessary intermediary quantities are pre-calculated and named in *Excel* worksheet.

For example:

- a) Type 'n(Height;Shoe\_size)' and count into following cell the number of students whose height and shoe size were both known (only these students are used by to calculate the correlation coefficient value).

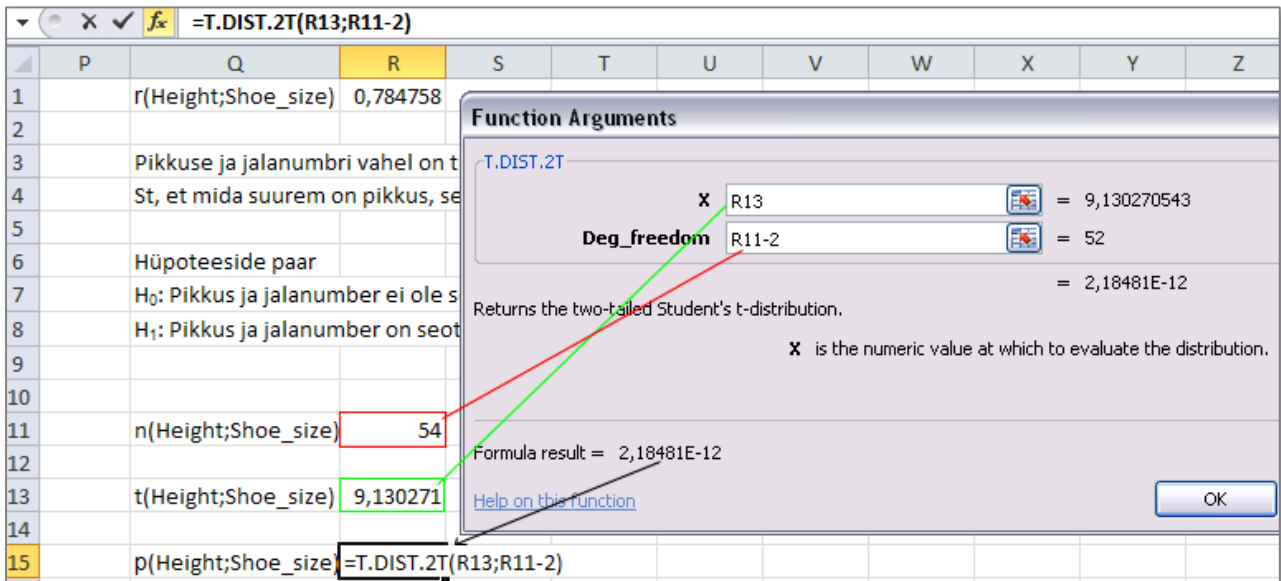
<b>r(Height;Shoe_size)</b>	0,784758
There is a <b>strong positive</b> relationship between height and shoe size.	
This means, that to bigger height corresponds <b>bigger</b> shoe size on an average.	
Hypothesis pair	
H <sub>0</sub> : Height and shoe size are not related (or mathematically $r = 0$ )	
H <sub>1</sub> : Height and shoe size are related (or mathematically $r \neq 0$ )	
<b>n(Height;Shoe_size)</b>	
<b>t(Height;Shoe_size)</b>	
<b>p(Height;Shoe_size)</b>	

b) Type behind the cell 't(Height;Shoe\_size)' formula to calculate absolute value of teststatistic:



c) Input behind the cell 'p(Height;Shoe\_size)' function T.DIST.2T with two arguments:

- Absolute value of teststatistic  $|t|$  and
- (number of observations) – 2, the parameter of the corresponding t-distribution:  $(n - 2)$ .



**NB!** In older *Excel* versions there is no function T.DIST.2T and function TDIST must be used. It has three arguments: the first two are the same as in function T.DIST.2T ( $|t|$  and  $n-2$ ), the third argument is number 2 (it determines, that two-side hypothesis  $r \neq 0$  is tested).

**6. Make a formal decision, which of the hypothesis is right and why.**

A'la: p(Height;Shoe\_size) 2,18E-12 < 0,05 => H<sub>1</sub>: students' height and shoe size are related

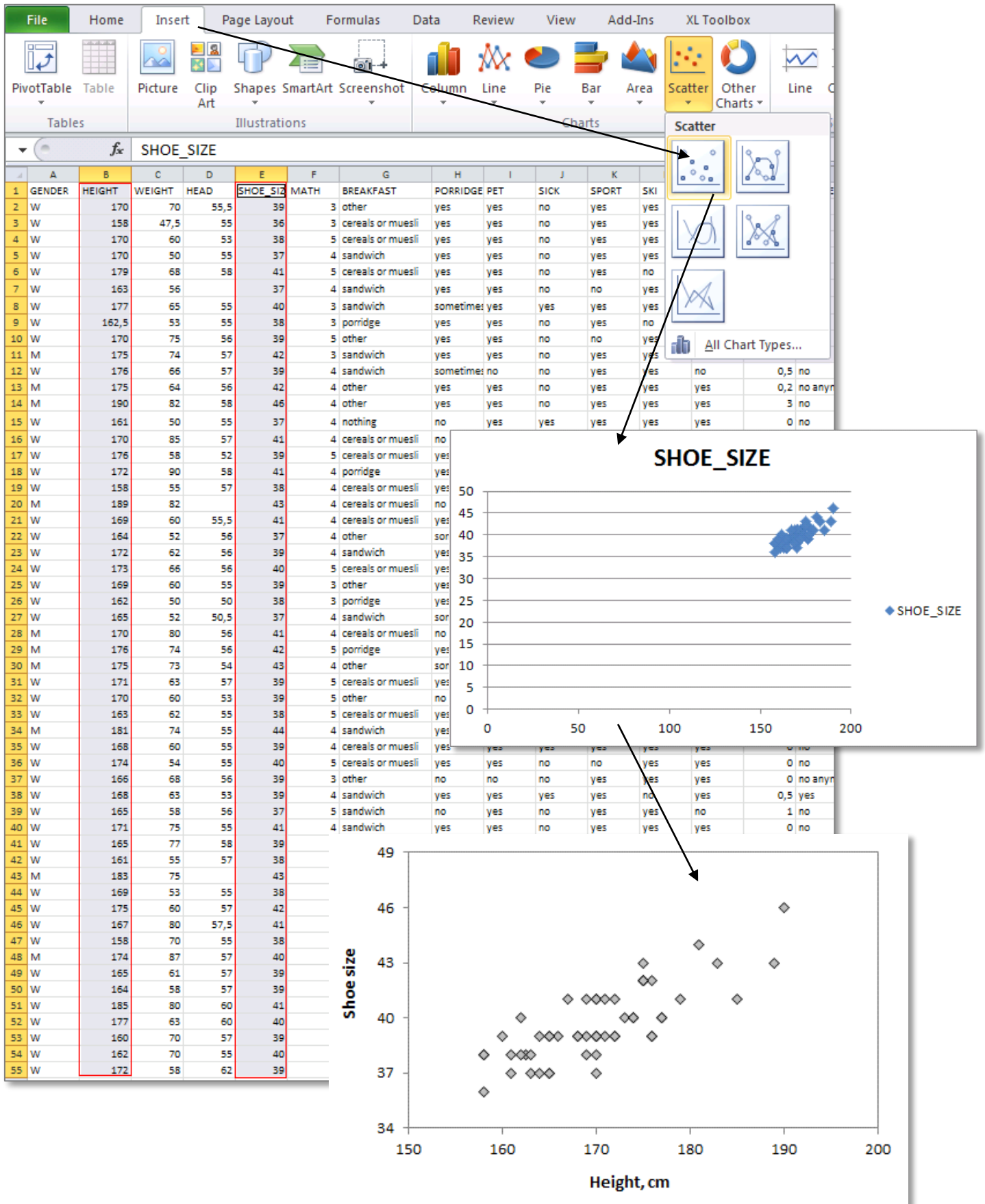
Remark. 2,18481E-12 = 2,18481... × 10<sup>-12</sup>

**7. Write down the final conclusion.**

A'la: between students height and shoe size there is **strong positive statistically significant** relationship ( $r = 0.785, p < 0.001$ ).

## Exercise 2 guide

Illustrate the relationship between variables 'HEIGHT' and 'SHOE\_SIZE' with scatterplot.



### Exercise 3 guide

1. Calculate correlation coefficients between all continuous variables in dataset (height – shoe size) using statistical procedure *Correlation* (Data-tab -> Data analysis... -> Correlation).

The screenshot shows the Excel interface with the Data Analysis toolpak installed. The Data Analysis dialog box is open, and 'Correlation' is selected. The Correlation dialog box is also open, showing the input range as '\$B\$1:\$E\$55' and the output range as '\$Q\$39'. The 'Labels in first row' checkbox is checked. The resulting correlation matrix is shown below:

	HEIGHT	WEIGHT	HEAD	SHOE_SIZE
HEIGHT	1			
WEIGHT	0,51228311	1		
HEAD	0,28996283	0,39609	1	
SHOE_SIZE	0,78475758	0,69898	0,29501	1

Result:

2. a) Between which variables is the strongest relationship? But the weakest?  
 b) Is the shoe size more related with height or weight?  
 c) With which body measurement has the strongest relationship head circumference?

**Describe some correlations (write down the sentences describing both the strength and the direction of relationships)!**