# Biometry practical 5

## Illustrated (imperfect) practical guide

## Preparatory work

1. Open in *MS Excel* the questionary 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.

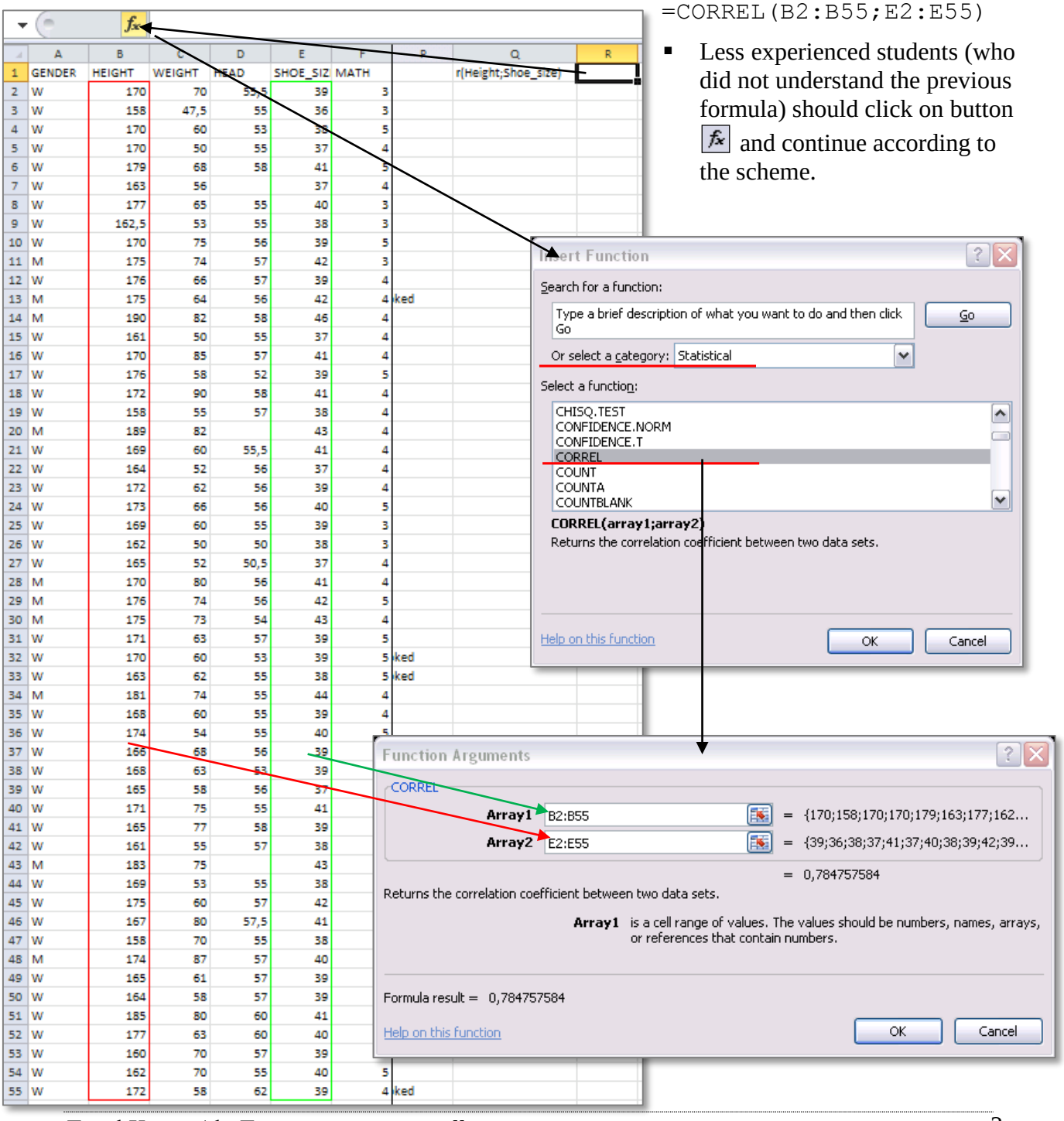
   For example, at the present moment the task is to calculate the correlation coefficient between height and shoe size – into *Excel* worksheet should be typed

   - '**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 $f_x$ and continue according to the scheme.



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.

| | |
|---|---|
| r(Height;Shoe_size) | 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 | |
| $H_0$: Height and shoe size **are not related** (or mathematically $r = 0$) | |
| $H_1$: Height and shoe size **are related** (or mathematically $r \neq 0$) | |

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

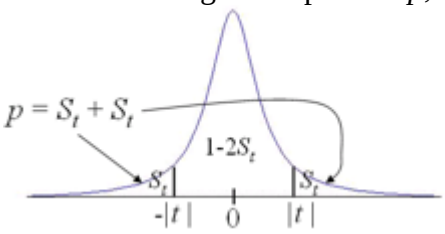
**Reminder from theory – hypothesis testing about correlation coefficient**

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

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

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)`.

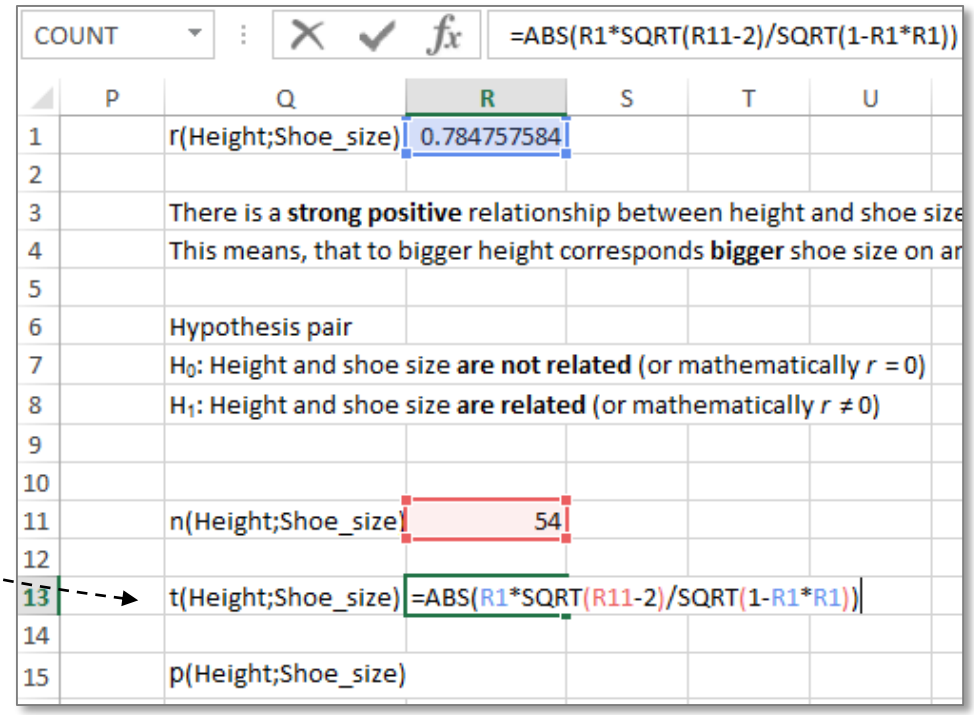
5. The evaluation of significance probability (p-value) is easier to perform, if all necessary intermediary quantities are pre-calculated and named in *Excel* worksheet.

   For example:

| | |
|---|---|
| r(Height;Shoe_size) | 0,784758 |
| There is a **strong positive** relationship between height and shoe size. | |
| This means, that to bigger height corresponds **bigger** shoe size on an average. | |
| Hypothesis pair | |
| $H_0$: Height and shoe size **are not related** (or mathematically $r = 0$) | |
| $H_1$: Height and shoe size **are related** (or mathematically $r \neq 0$) | |
| | |
| n(Height;Shoe_size) | |
| t(Height;Shoe_size) | |
| p(Height;Shoe_size) | |

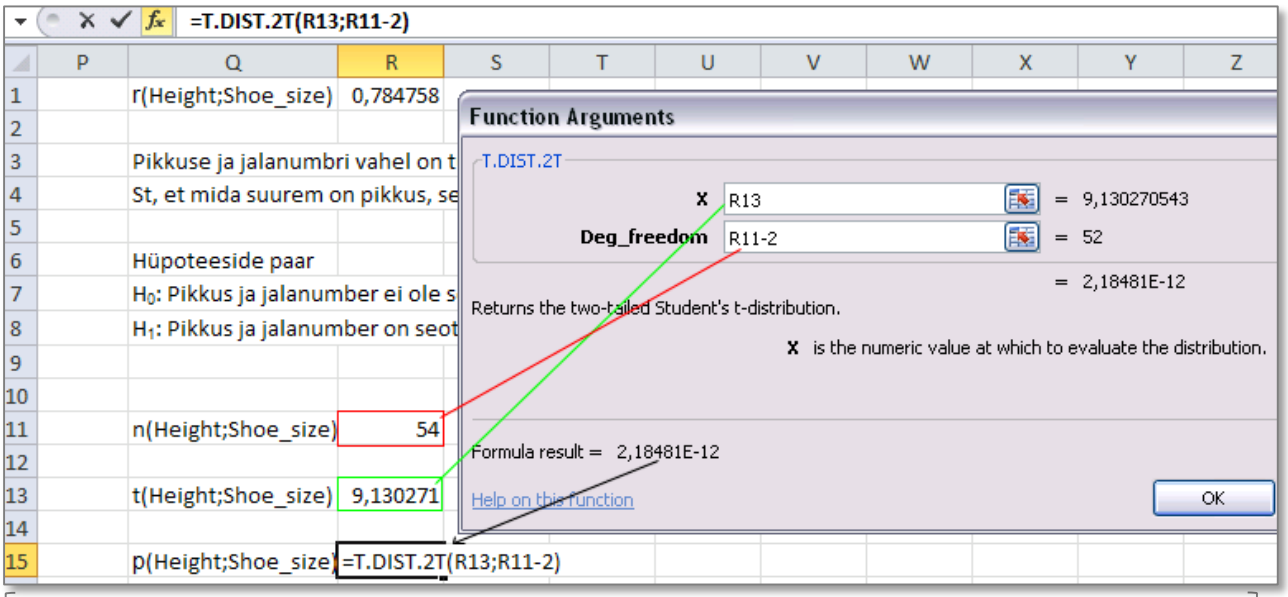
   a) Type 'n(Height;Shoe_size)' and count into following cell the number of students whose height and shoe size were both known (only these students are used by to calculate the correlation coefficient value).

b) 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 => $H_1$: students' height and shoe size are related

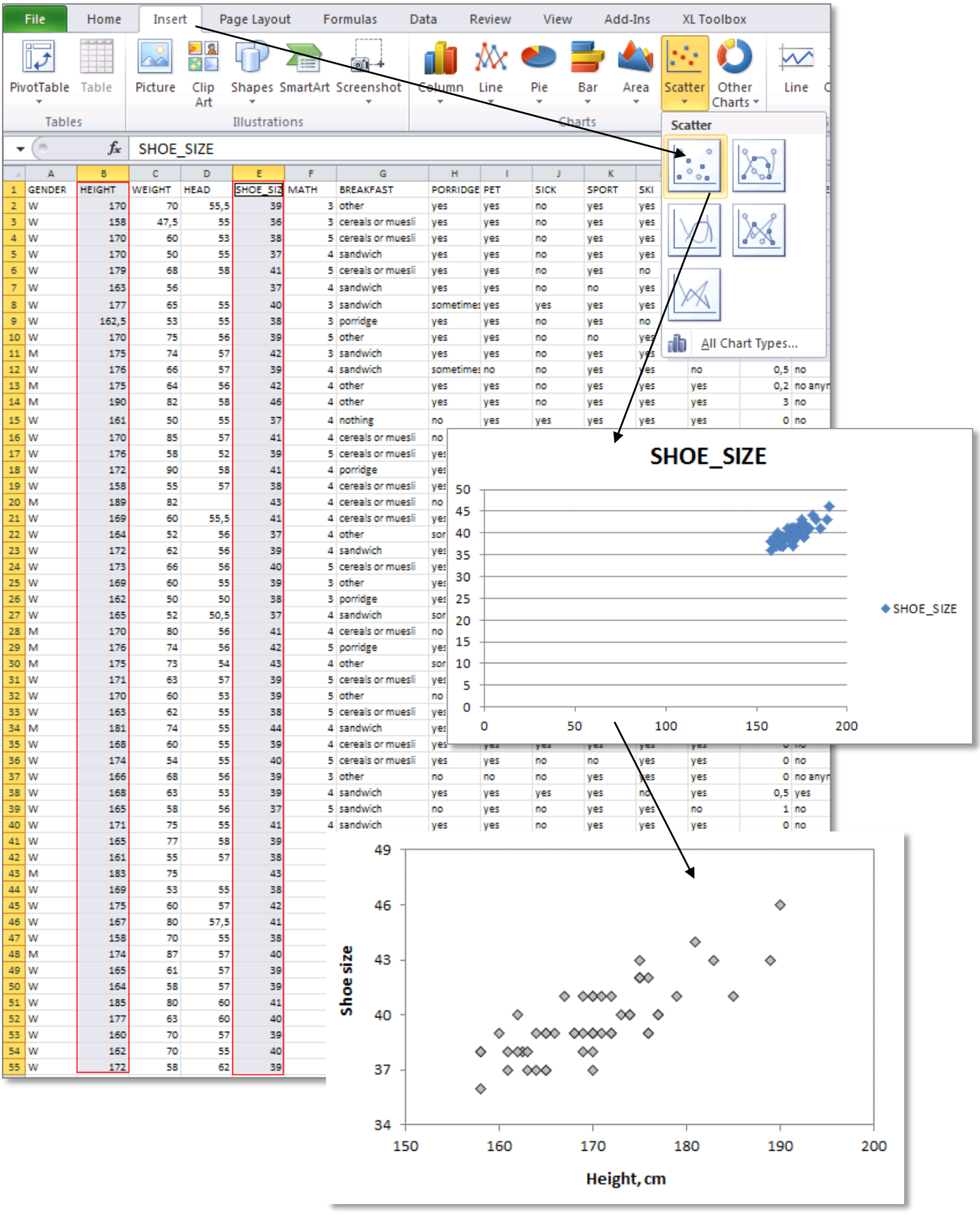
Remark. 2,18481E-12 = 2,18481…×$10^{-12}$

## 7. Write down the final conclusion.

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

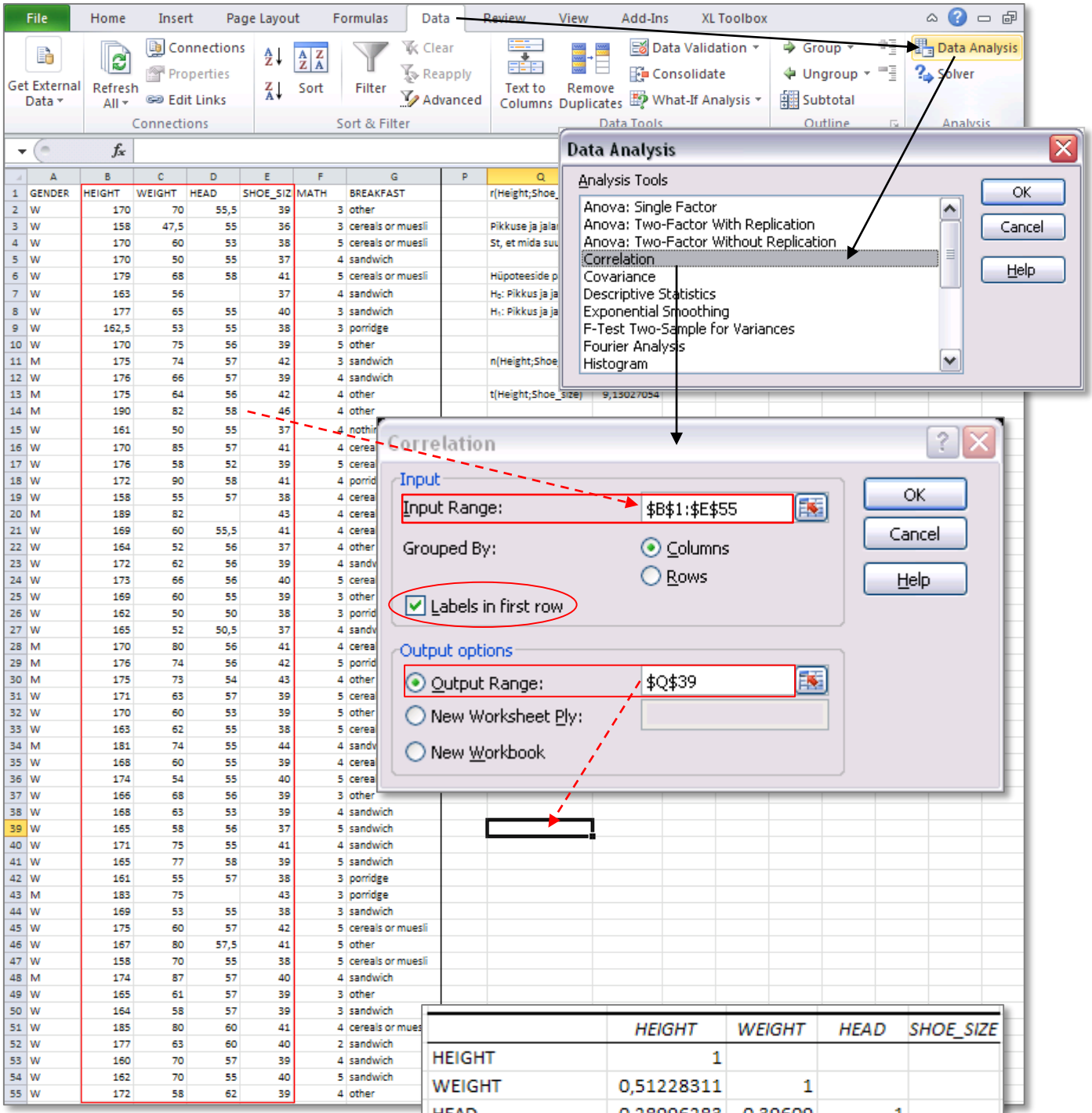
## Exercise 2 guide

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*).




Result:

| | 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 |

**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)!**