Biometry practical 4

Illustrated (imperfect) practical guide

Preparatory work

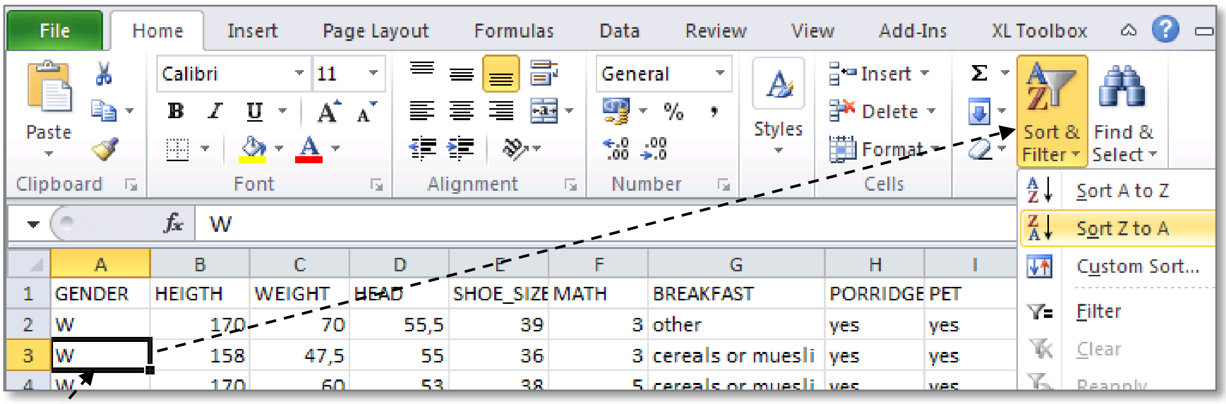
1. Open in *MS Excel* the questionary data (file analysed already in previous practical),
2. insert new worksheet, rename new worksheet to 'Praks4' (or 'Practical4') and
3. make a copy of the data table (from worksheet 'Andmed') and paste it into the upper left corner of the new worksheet.

Exercise 1.

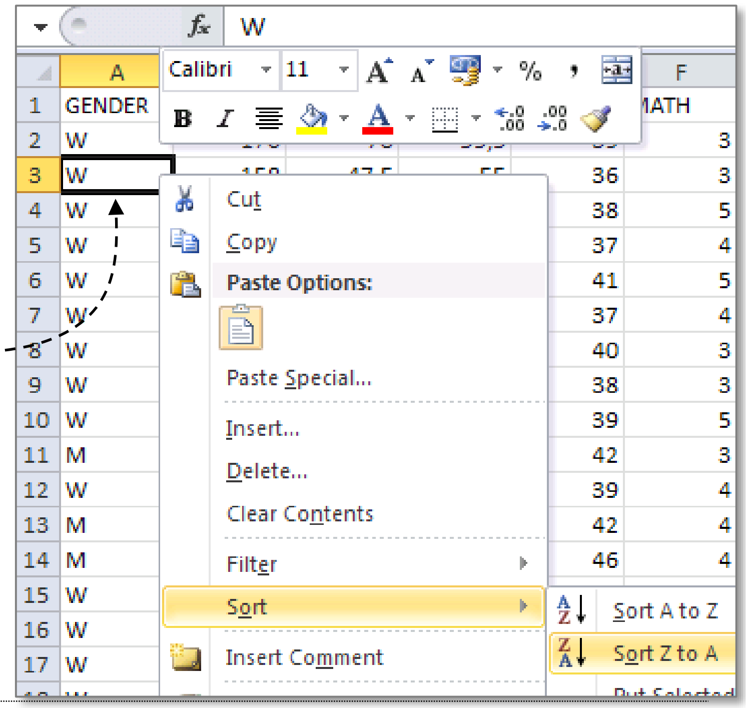
Suppose that the first year students of Institute of Veterinary Medicine and Animal Sciences (students in our dataset) are just a random sample from all first year students of Estonian University of Life Sciences. Knowing, that the average height of Estonian women is 169 cm, test the hypothesis: is the average height of first year female students of Estonian University of Life Sciences different from Estonian average 169 cm?

Guide

1. Sort the datatable by column 'GENDER'.



1) For example, by putting the cursor into column 'GENDER' and selecting from *Home*-tab commands
Sort & Filter ->
Sort Z to A
(to get women before men).



2) Or clicking with right mouse button on an arbitrary cell in column 'GENDER' and selecting from the drop down menu commands
Sort -> Sort Z to A.

2. Find the number of girls and their average height and standard deviation of height using functions COUNT, AVERAGE and STDEV.S.

=STDEV.S(B2:B45)

	A	B	C	L	M	N	O	P	Q	R	S
1	GENDER	HEIGTH	WEIGHT	SKI	CAR	BEER	SMOKE		**Neidude pikkus**		
2	W	170	70	yes	yes	0	no		Vaatluste arv	44	
3	W	158	47,5	yes	yes	0	no		Keskmine	168,2	
4	W	170	60	yes	no	0	no		Standardhälve	=STDEV.S(B2:B45)	
5	W	170	50	yes	no	0	no				
6	W	179	68	no	no	0	no				
7	W	163	56	yes	no	0	no				
8	W	177	65	yes	no	0	no				
9	W	162,5	53	no	no	0,25	no				
10	W	170	75	yes	yes	0	no				
11	W	176	66	yes	no	0,5	no				
12	W	161	50	yes	yes	0	no				
13	W	170	85	no	no	0	no				
14	W	176	58	no	no	0,5	no				
15	W	172	90	no	no	0	no				
16	W	158	55	yes	yes	0	yes				
17	W	169	60	yes	no	1	yes				
18	W	164	52	yes	no	0	no				
19	W	172	62	no	no	0	no				
20	W	173	66	yes	no	0	no				
21	W	169	60	no	yes	0	no				
22	W	162	50	yes	yes	0	no				
23	W	165	52	yes		0	no				
24	W	171	63	yes	yes	0	no				
25	W	170	60	no	yes	0	no anymore, but I\'ve smoked				
26	W	163	62	yes	yes	0	no anymore, but I\'ve smoked				
27	W	168	60	yes	yes	0	no				
28	W	174	54	yes	yes	0	no				
29	W	166	68	yes	yes	0	no anymore, but I\'ve smoked				
30	W	168	63	no	yes	0,5	yes				
31	W	165	58	yes	no	1	no				
32	W	171	75	yes	yes	0	no				
33	W	165	77	no	yes	0	no				
34	W	161	55	yes	yes	0	no anymore, but I\'ve smoked				
35	W	169	53	yes	yes	0	no				
36	W	175	60	no	no	0,5	no				
37	W	167	80	yes	yes	2	no				
38	W	158	70	no	no	0	no				
39	W	165	61	no	yes	0,5	yes				
40	W	164	58	no	yes	0	no anymore, but I\'ve smoked				
41	W	185	80	no	yes	0	no				
42	W	177	63	yes	no	0	no				
43	W	160	70	no	yes	0	no anymore, but I\'ve smoked				
44	W	162	70	no	no	2	no				
45	W	172	58	no	no	0	no anymore, but I\'ve smoked				
46	M	175	74	yes	no	1	no				
47	M	175	64	yes	yes	0,2	no anymore, but I\'ve smoked				

Girls' height	
Sample size	44
Average	168.2386
Standard deviation	6.076536

So, there are 44 girls with average height 168.2 cm and standard deviation 6.1 cm;

– this says, that the average difference of girls' height from 168.2 cm is 6.1 cm; or, expecting that the height follows normal distribution, then according to the properties of normal distribution

*) approximately 68.3% of the heights of first year female students is in interval 168.2±6.1 cm ($\bar{x} \pm s$) and

*) approximately 95.5% of the heights of first year female students is in interval 168.2±12.2 cm ($\bar{x} \pm 2s$).

3. Formulate the hypothesis pair and write it down.

For example:

Girls' height	
Sample size	44
Average	168.2386
Standard deviation	6.076536

H_0: the heights of first year female students correspond to Estonian standard (169 cm)
H_1: the heights of first year female students does not correspond to Estonian standard (169 cm)

H_0: the average height of first year female students does not differ from 169 cm
H_1: the average height of first year female students differs from 169 cm

H_0: $\mu_F = 169$ μ_F - average height of first year female students
H_1: $\mu_F \neq 170$

Remainder from theory – relationships between hypothesis testing and confidence intervals

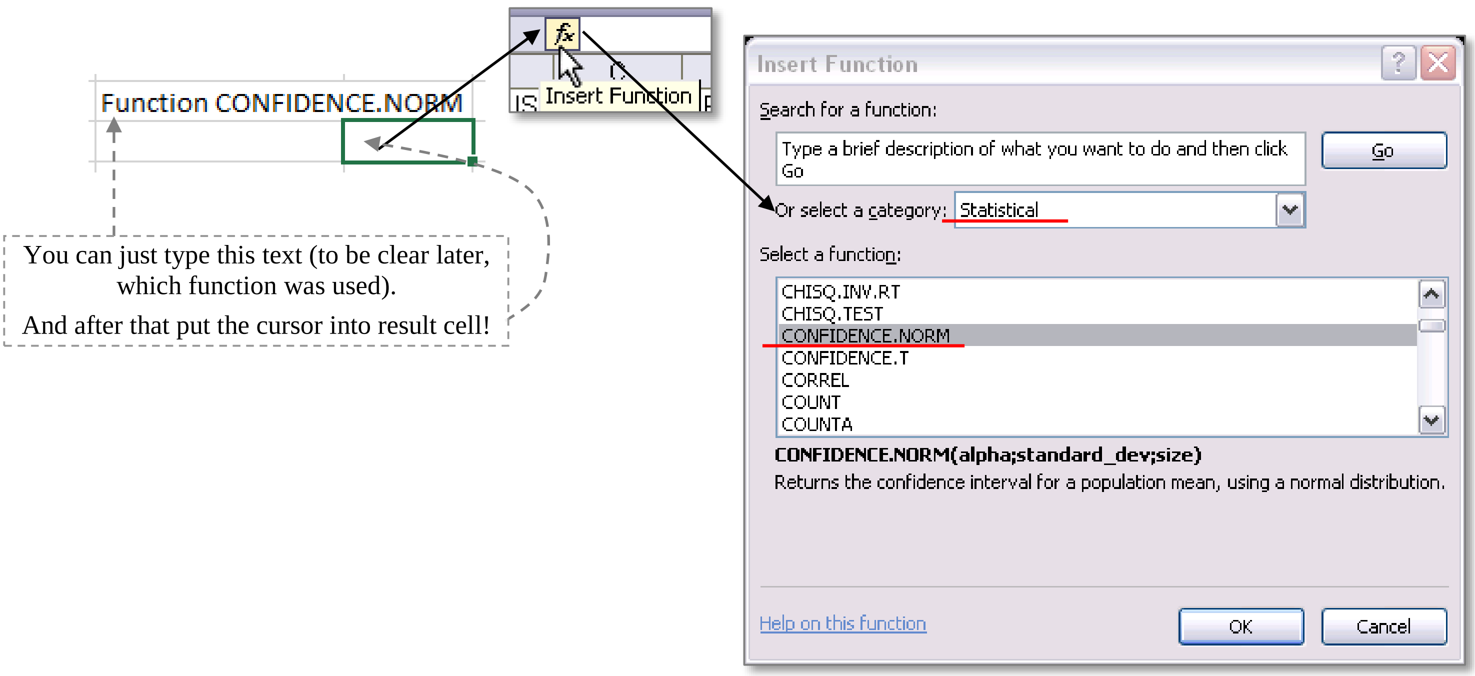
- If the task of hypothesis testing is to compare some estimated parameter with constant value, the decision is often made based on the confidence interval of the studied parameter:
  - if the **constant is between** confidence limits, **then there is no reason to reject null hypothesis** and it can be concluded **that the studied parameter does not differ from given constant**;
  - if constant is outside confidence interval, then the **studied parameter is differ from given constant**.
- For example, if you want to compare the average value with given constant (does the data correspond to standard), the hypothesis pair is:

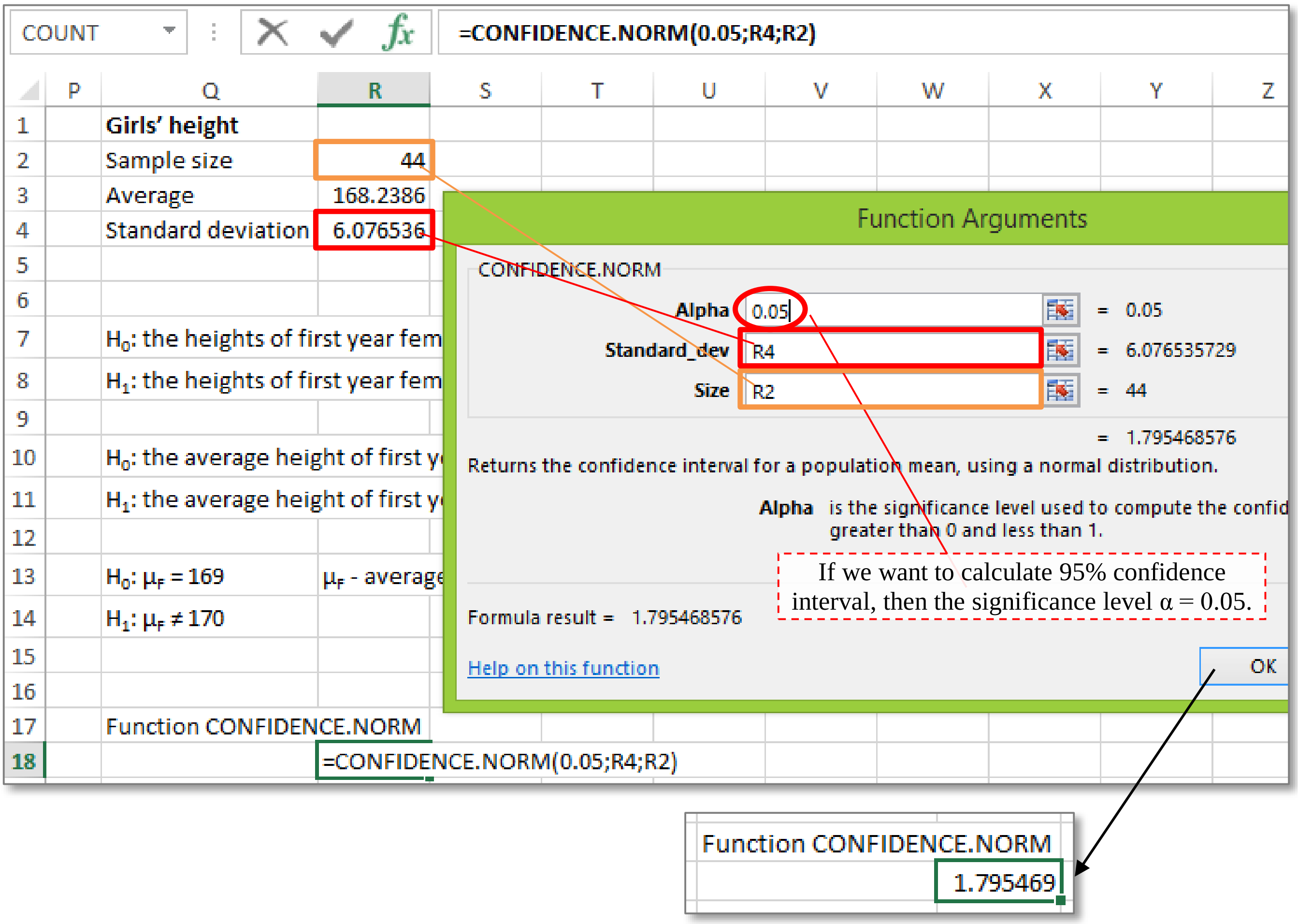
$$H_0: \mu = c \text{ and } H_1: \mu \neq c.$$

if $c \in [\underline{\mu}, \overline{\mu}]$, then $H_0: \mu = c$ is true; if $c \notin [\underline{\mu}, \overline{\mu}]$, then $H_1: \mu \neq c$ is true.

4. Calculate the half of the 95% confidence interval using functions CONFIDENCE.NORM and CONFIDENCE.T:

a) function CONFIDENCE.NORM
(this function has 3 arguments:
significance level α, standard deviation and number of female students in dataset);

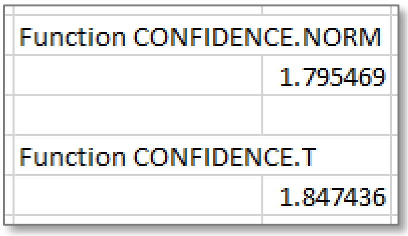




b) function CONFIDENCE.T

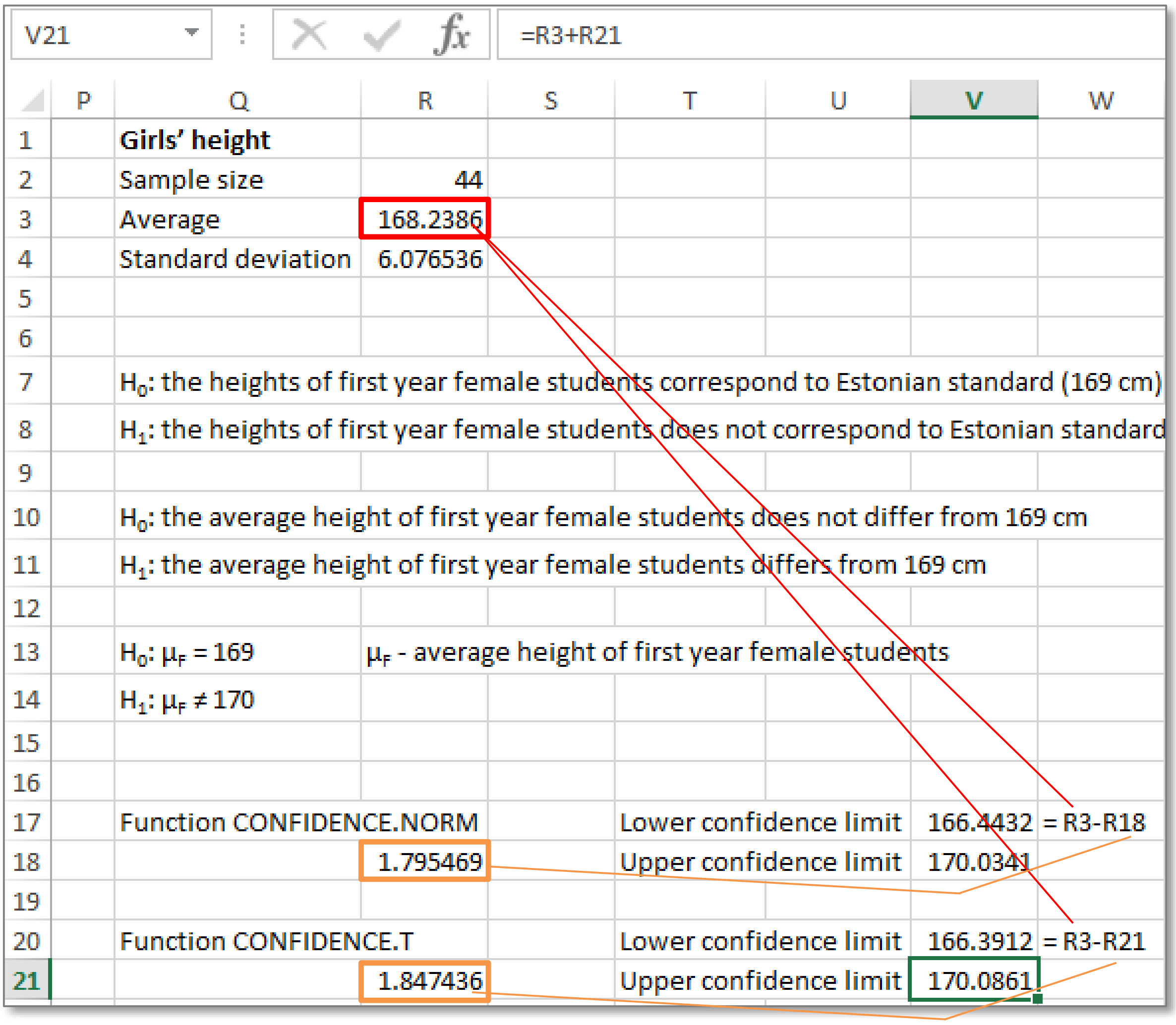
(arguments of this function are similar to the function CONFIDENCE.NORM).

Result:



Function CONFIDENCE.NORM	
	1.795469
Function CONFIDENCE.T	
	1.847436

5. Calculate the lower and upper confidence limits based on both functions.



V21 | =R3+R21

	P	Q	R	S	T	U	V	W
1		**Girls' height**						
2		Sample size	44					
3		Average	168.2386					
4		Standard deviation	6.076536					
5								
6								
7		H_0: the heights of first year female students correspond to Estonian standard (169 cm)						
8		H_1: the heights of first year female students does not correspond to Estonian standard						
9								
10		H_0: the average height of first year female students does not differ from 169 cm						
11		H_1: the average height of first year female students differs from 169 cm						
12								
13		H_0: $\mu_F = 169$	μ_F - average height of first year female students					
14		H_1: $\mu_F \neq 170$						
15								
16								
17		Function CONFIDENCE.NORM			Lower confidence limit		166.4432	= R3-R18
18			1.795469		Upper confidence limit		170.0341	
19								
20		Function CONFIDENCE.T			Lower confidence limit		166.3912	= R3-R21
21			1.847436		Upper confidence limit		170.0861	

Which of these 95% confidence intervals is wider? Why?

Answer.
The confidence interval got with function CONFIDENCE.T is slightly wider.

The reason is, that function CONFIDENCE.T calculates the confidence limits based on t-distribution following the formula $\bar{x} \pm t_{1-\alpha/2,n-1} \, s/\sqrt{n}$,

but function CONFIDENCE.NORM calculates **asymptotic** (approximate) confidence interval (whereby the accuracy is increasing when the sample size is increasing) based on normal distribution: $\bar{x} \pm z_{1-\alpha/2} \, s/\sqrt{n}$, got range is little under estimated in case of small data set.

Parameters $t_{1-\alpha/2,n-1} = t_{0,975;43} = 2,017$ and $z_{1-\alpha/2} = z_{0,975} = 1,96$ are 97.5%-points (values, from which the bigger values can occur only with probability 0.025) of t-distribution and standard normal distribution, respectively, the first of these quantities is calculable in Excel 2010 with function =T.INV(0,975;43) and the second with function =NORM.S.INV(0,975).

NB! In the older *Excel* versions there is no functions CONFIDENCE.T and CONFIDENCE.NORM. There is only function CONFIDENCE, which is equivalent to functions CONFIDENCE.NORM,
to calculate the confidence interval based on t-distribution the corresponding option *Confidence Level for Mean* of procedure *Descriptive Statistics* can be used.

6. Make a final conclusion and write it down – does the average height of first year female students differ from Estonian average (169 cm)? The answer must contain also the argumentation, why you made this decision.

Example. As the average height of Estonian women (169 cm) is between 95% confidence limits of the first year female students average height: 169∈(166,4; 170,1), then there is no reason to reject nullhypothesis H_0: the average heights of first year female students does not differ from 169 cm.

7. Supplementary task.

The average height of women over the world is 154 cm. Can you conclude that the average height of first year female students in Estonian University of Life Sciences differs from the world average?

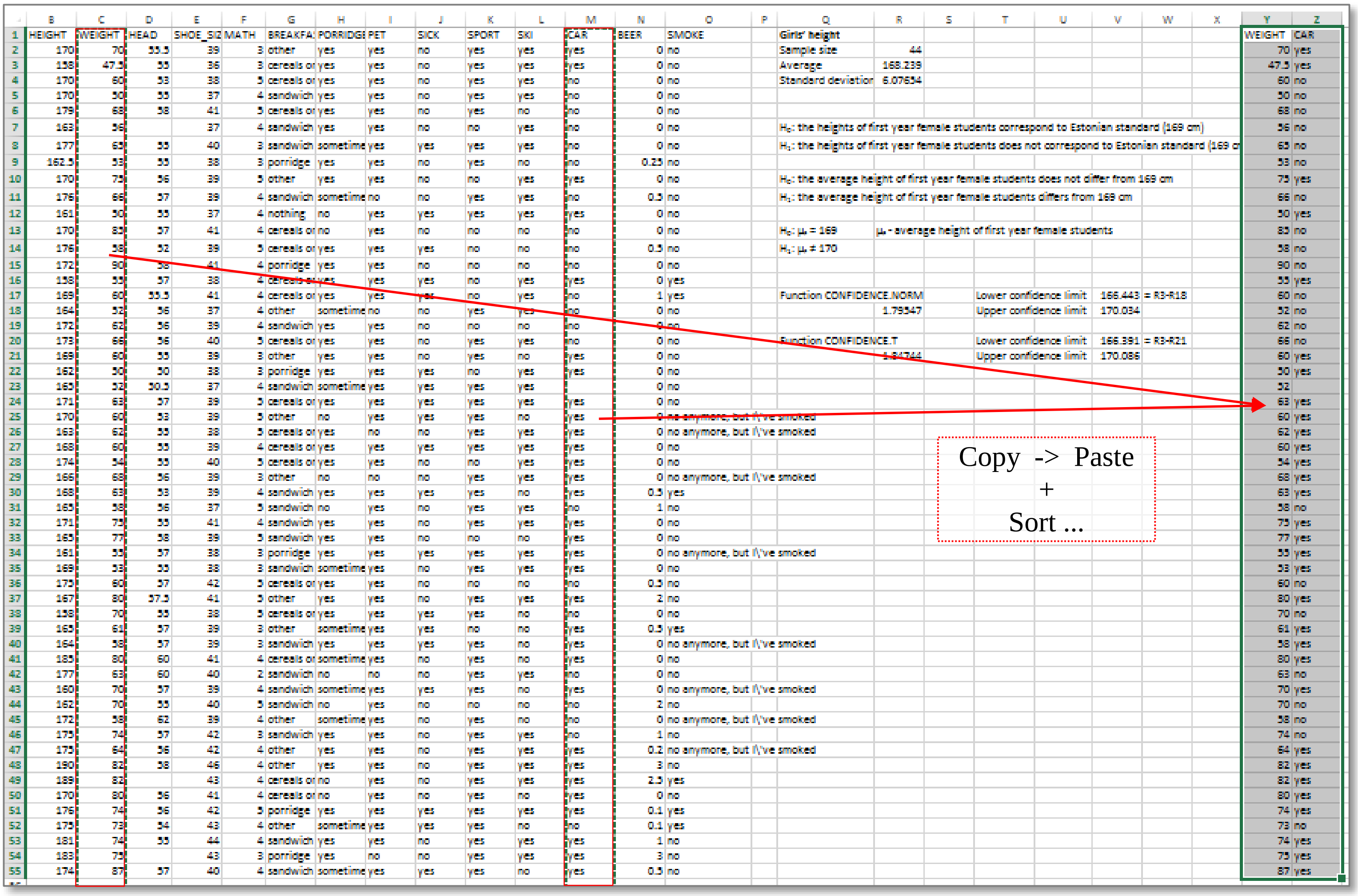
NB! You don't need to make any additional calculations. The decision can be made just based on already calculated confidence interval.

Exercise 2.

Are weights of students owning and not owning a car different?

Guide

1. Make an additional table containing only columns 'WEIGHT' and 'CAR' and sort it by column 'CAR'.



2. Calculate the average and standard deviation of weights depending on the owning of car (NB! Omit the student, who does not know has she or he a car or not).

You can use corresponding <u>functions</u> or <u>PivotTable</u>. If you wish, you can try both variants.

WEIGHT	CAR
60	no
50	no
68	no
56	no
65	no
53	no
66	no
85	no
58	no
90	no
60	no
52	no
62	no
66	no
58	no
60	no
70	no
63	no
70	no
58	no
74	no
73	no
70	yes
47.5	yes
75	yes
50	yes
55	yes
60	yes
50	yes
63	yes
60	yes
62	yes
60	yes
54	yes
68	yes
63	yes
75	yes
77	yes
55	yes
53	yes
80	yes
61	yes
58	yes
80	yes
70	yes
64	yes
82	yes
82	yes
80	yes

	CAR	
	No	Yes
No of students	22	31
Average	64.40909091	66.59677419
Standard deviation	9.945849055	11.10886985

	Column Labels		
Values	no	yes	Grand Total
Count of WEIGHT	22	31	53
Average of WEIGHT2	64.40909091	66.59677419	65.68867925
StdDev of WEIGHT3	9.945849055	11.10886985	10.59854236

3. Formulate the hypothesis pair and write it down.

t-test	
H_0: The **average weights** of students owning and not owning a car **are not different**	
H_1: The **average weights** of students owning and not owning a car **are different**	
or	
H_0: $\mu_{No}=\mu_{Yes}$	μ_{No} - average weights of students not owning a car
H_1: $\mu_{No}\neq\mu_{Yes}$	μ_{Yes} - average weights of students owning a car

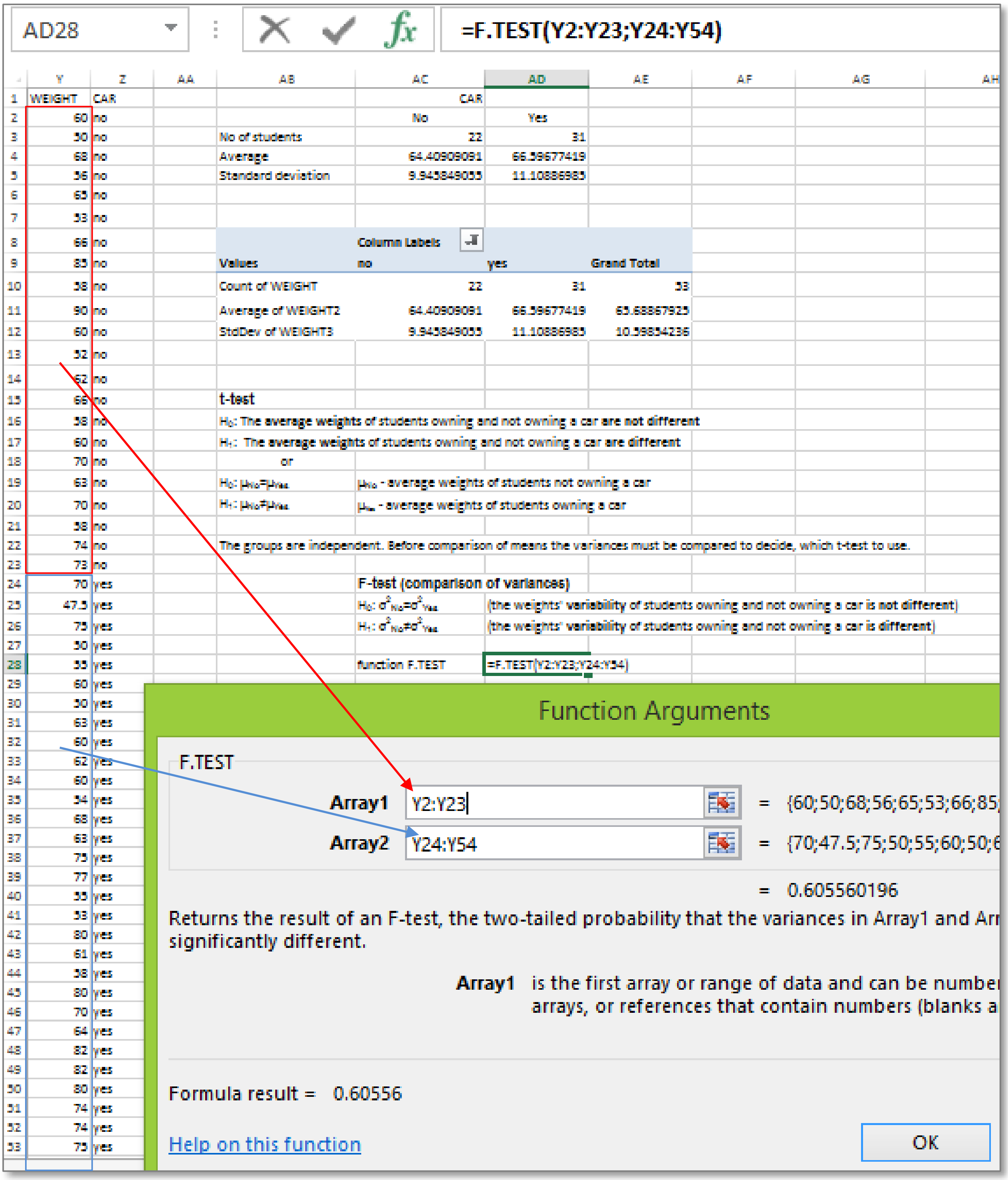
4. Which t-test to use?
NB! There are three types of t-tests, look at page 12 (step 7b).

- As compared **groups are independent** (there are different students in groups), before the comparison of means the **variances must be compared** to decide, which t-test to use (this, which assumes equal variances, or this, which assumes unequal variances**)**.
- **To compare variances the F-test can be used**.

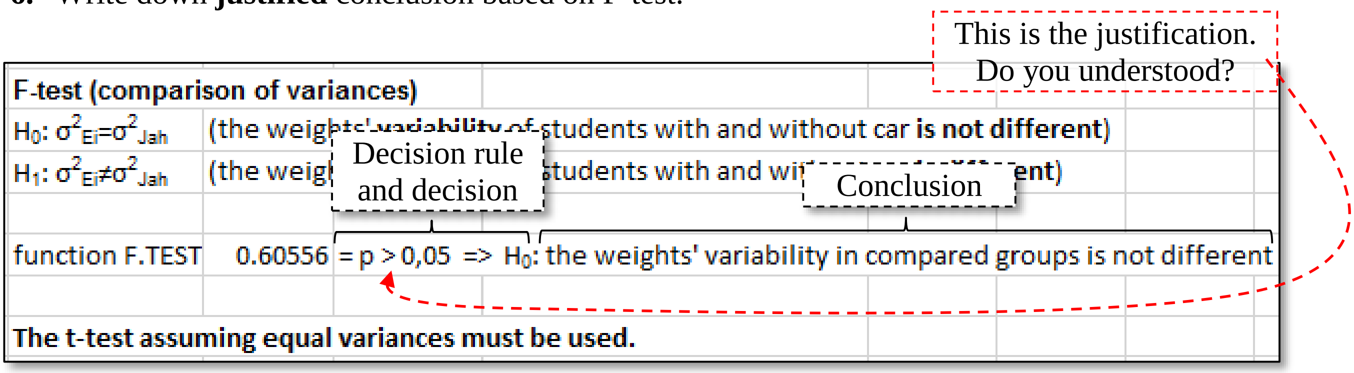
5. To decide is the weights' variability of students with and without car equal or not, formulate the **hypothesis pair for variances comparison** and **perform F-test (function F.TEST**).

NB! There is also statistical procedure *F-test* (*Data*-tab -> *Data Analysis... -> F-Test Two-Sample for Variances*), but this tests only one side hypothesis and can't be directly applied to decide about equality of variances.

t-test		
H_0: The **average weights** of students owning and not owning a car **are not different**		
H_1: The **average weights** of students owning and not owning a car **are different**		
or		
H_0: $\mu_{No}=\mu_{Yes}$	μ_{No} - average weights of students not owning a car	
H_1: $\mu_{No}\neq\mu_{Yes}$	μ_{Yes} - average weights of students owning a car	
The groups are independent. Before comparison of means the variances must be compared to decide, which t-test to use.		
	F-test (comparison of variances)	
	H_0: $\sigma^2_{No}=\sigma^2_{Yes}$	(the weights' **variability** of students owning and not owning a car **is not different**)
	H_1: $\sigma^2_{No}\neq\sigma^2_{Yes}$	(the weights' **variability** of students owning and not owning a car **is different**)



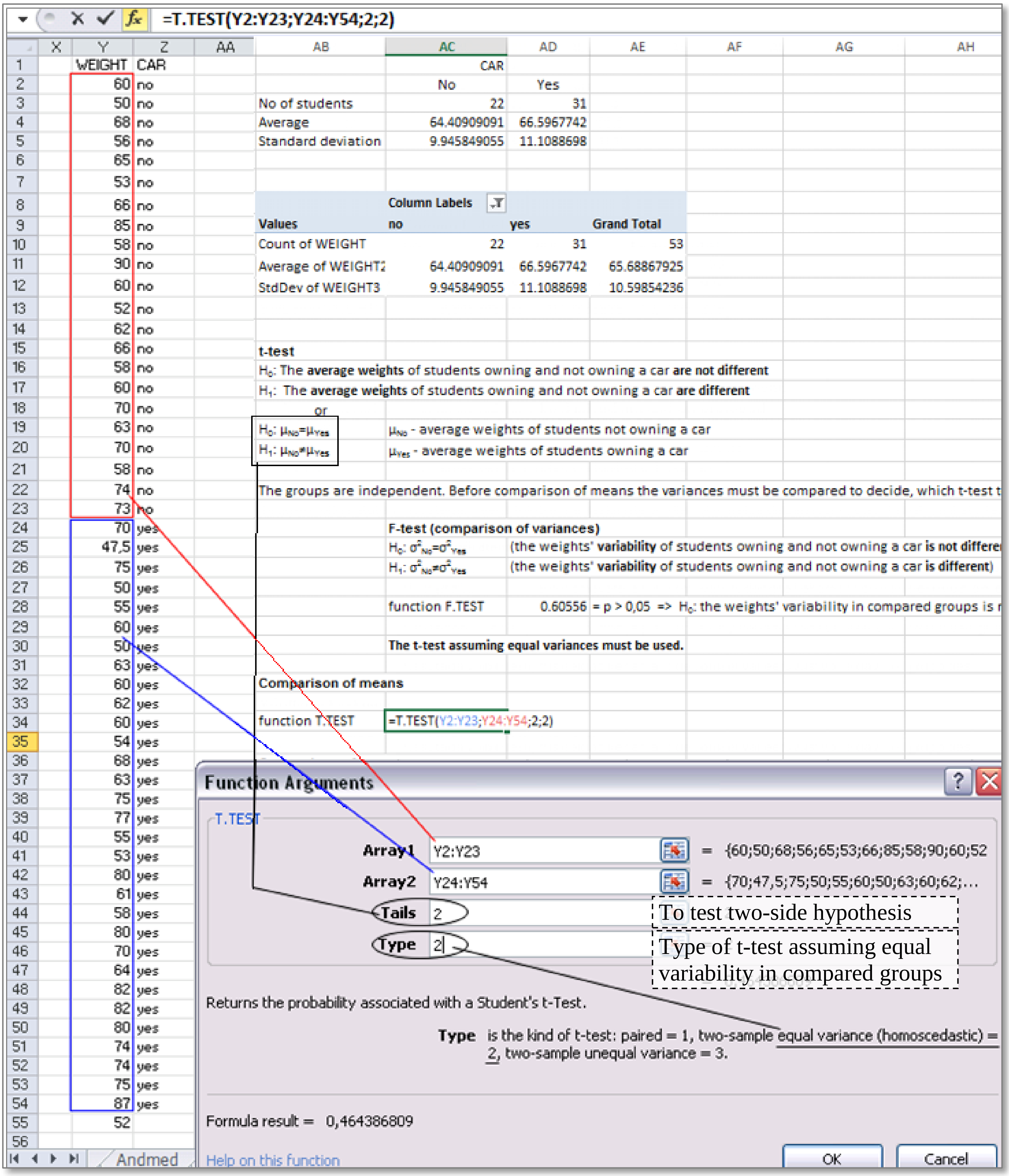
6. Write down **justified** conclusion based on F-test.



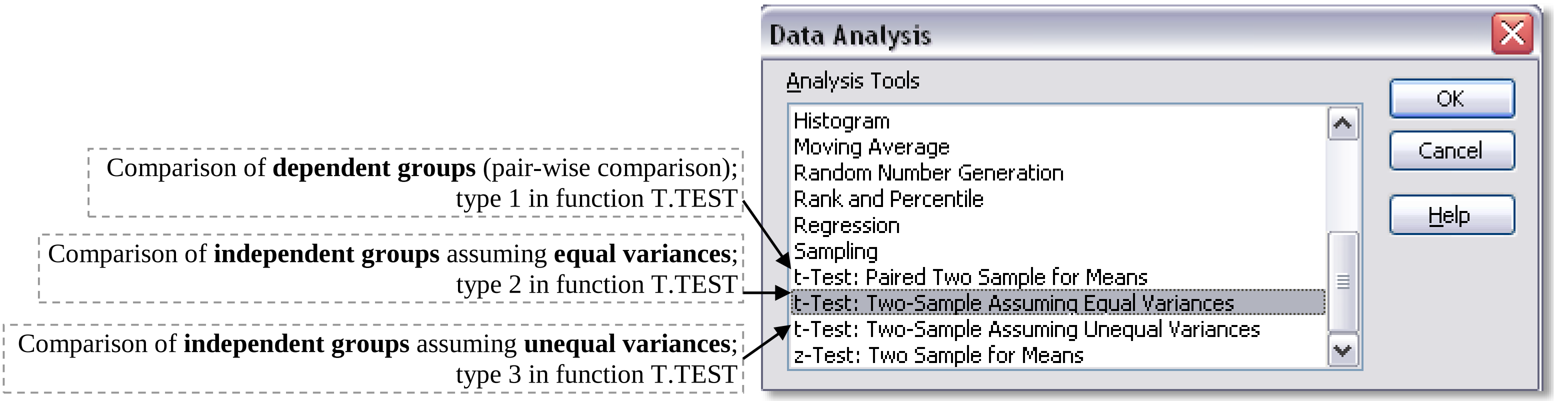
7. Perform the t-test to compare average weights.

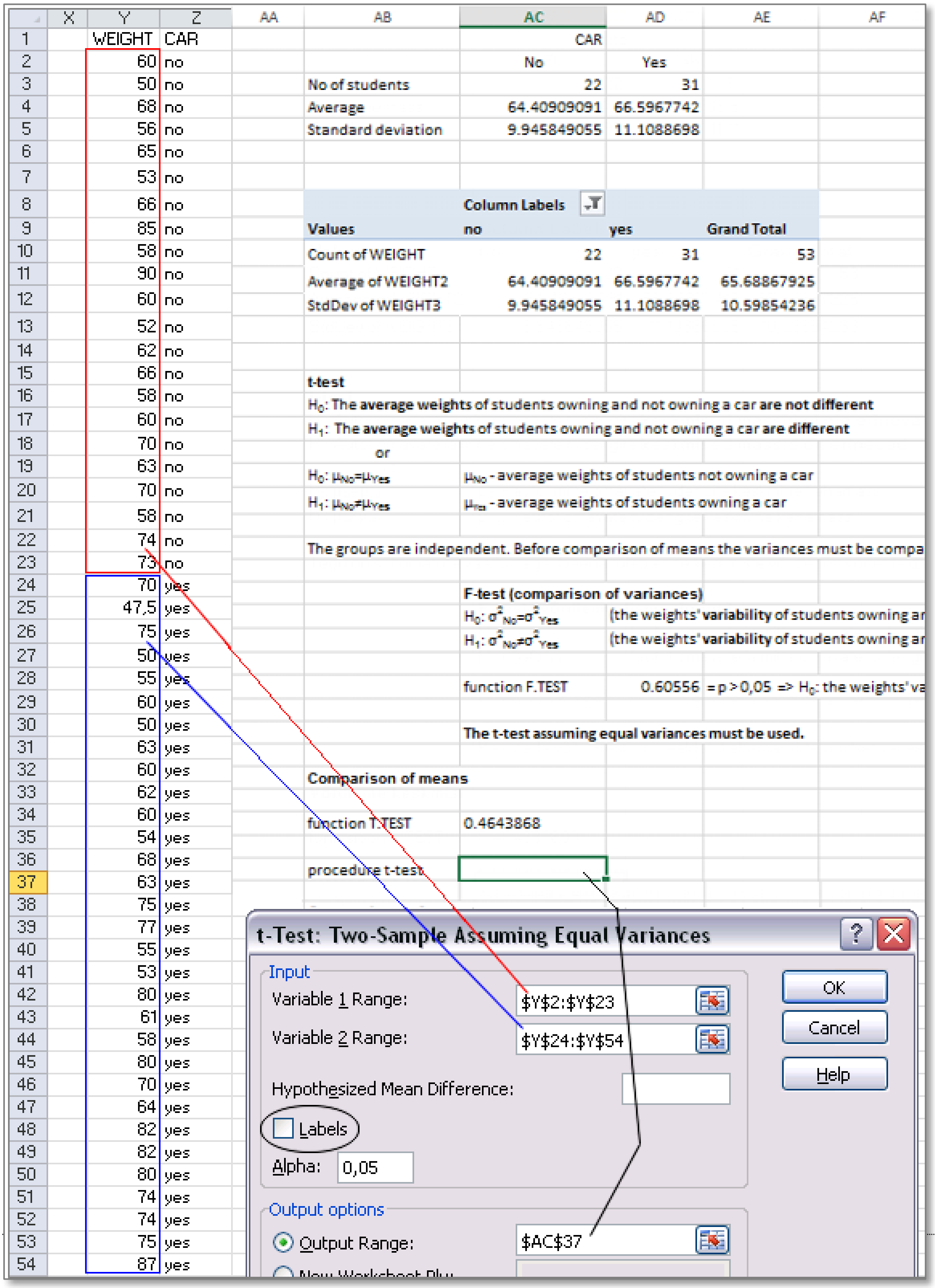
Using both

a) function T.TEST:



b) corresponding statistical procedure (*Data*-tab -> *Data Analysis...* -> *t-Test: ...*):





8. Write down the final conclusion, justify it(!).

Comparison of means

function T.TEST 0,4643868 = $p > 0{,}05$ => H_0: the **average weights** of students with and without car **are not different**

procedure t-test: t-Test: Two-Sample Assuming Equal Variances

	Variable 1	*Variable 2*
Mean	64,40909	66,596774
Variance	98,91991	123,40699
Observations	22	31
Pooled Varianc	113,3241	
Hypothesized M	0	
df	51	
t Stat	-0,73719	
P(T<=t) one-tail	0,232193	
t Critical one-ta	1,675285	
P(T<=t) two-tail	0,464387	= $p > 0{,}05$ => H_0: the **average weights** of students with and without car **are not different**
t Critical two-ta	2,007584	

Other possibilities for the final conclusion:

- „**the average weights of students owning and not owning a car are not statistically significantly different (p > 0,05)**“,
  this is a little more scientific and accurate conclusion;
- „the students’ weight does not depend on owning of a car“ (this is little differently phrased conclusion but also correct).

Remark. Significance probability (p-value) $p = 0.464$ is showing that

- the probability to make a mistake concluding that the average weights are different is 46.4%;
- assuming, that in population there is no difference between compared groups, the probability to get the sample with observed difference just by chance is 0.464.

As from one hand the probability to make a wrong conclusion that the groups are different is too big (formally the traditional limit is 0.05) and from other hand the probability to get observed difference just by chance is also big (once again, the formal limit is 0.05), there is no reason to reject null hypothesis about equality of average weights.

9. But what are showing the other quantities calculated by procedure *t-Test: Two-Sample Assuming Equal Variances*?

- The first part of output contains basic descriptive statistics of compared groups (mean, variance and number of observations; NB! Be careful and don't mix up the compared groups; you may want to write instead abstract names 'Variable 1' and 'Variable 2' the real group names):

	Car='No' *Variable 1*	Car='Yes' *Variable 2*
Mean	64,40909091	66,59677419
Variance	98,91991342	123,4069892
Observations	22	31

- The other output parts are already related with hypothesis testing:

Pooled Variance	113,3240757	Common variance calculated assuming the equal variances in compared groups
Hypothesized Mean D	0	
df	51	
t Stat	-0,737186899	Empirical (sample based) value of t-statistic
P(T<=t) one-tail	0,232193405	p-value corresponding to one-side hypothesis
t Critical one-tail	1,67528495	Critical value of t-statistic corresponding to one-side hypothesis
P(T<=t) two-tail	0,464386809	p-value corresponding to two-side hypothesis
t Critical two-tail	2,00758377	Critical value of t-statistic corresponding to two-side hypothesis

Two-side (or two-tailed) hypothesis means testing "non-equal *against* equal":

H_0: $\mu_{No} = \mu_{Yes}$,
H_1: $\mu_{No} \neq \mu_{Yes}$.

One-side (or one-tailed) hypothesis means testing "less than *against* more or equal than" (or "more than *against* less or equal than"; in which direction to test depends on data, for example Excel performs one-side t-test testing always is bigger average also statistically significantly bigger):

H_0: $\mu_{No} \geq \mu_{Yes}$,
H_1: $\mu_{No} < \mu_{Yes}$.

Instead of p-value the final conclusion can be made also based on **comparing the empirical value of teststatistic $|t|$ with its critical value** $t_{critical}$. If the empirical value is smaller than critical value, $|t| = 1,68 < 2,01 = t_{critical}$, then there is no reason to reject the null hypothesis (the empirical value of teststatistic is in interval, where it should be with 95% probability if the null hypothesis is true – so, our data does not allow to reject null hypothesis).

10. Can you reject the null hypothesis of one-side t-test?

How about the conclusion of this one-side test (how to phrase it)?