Instructions for computer-use in solving the proposed practical activities

Downloading *EpiInfo* (for home use)

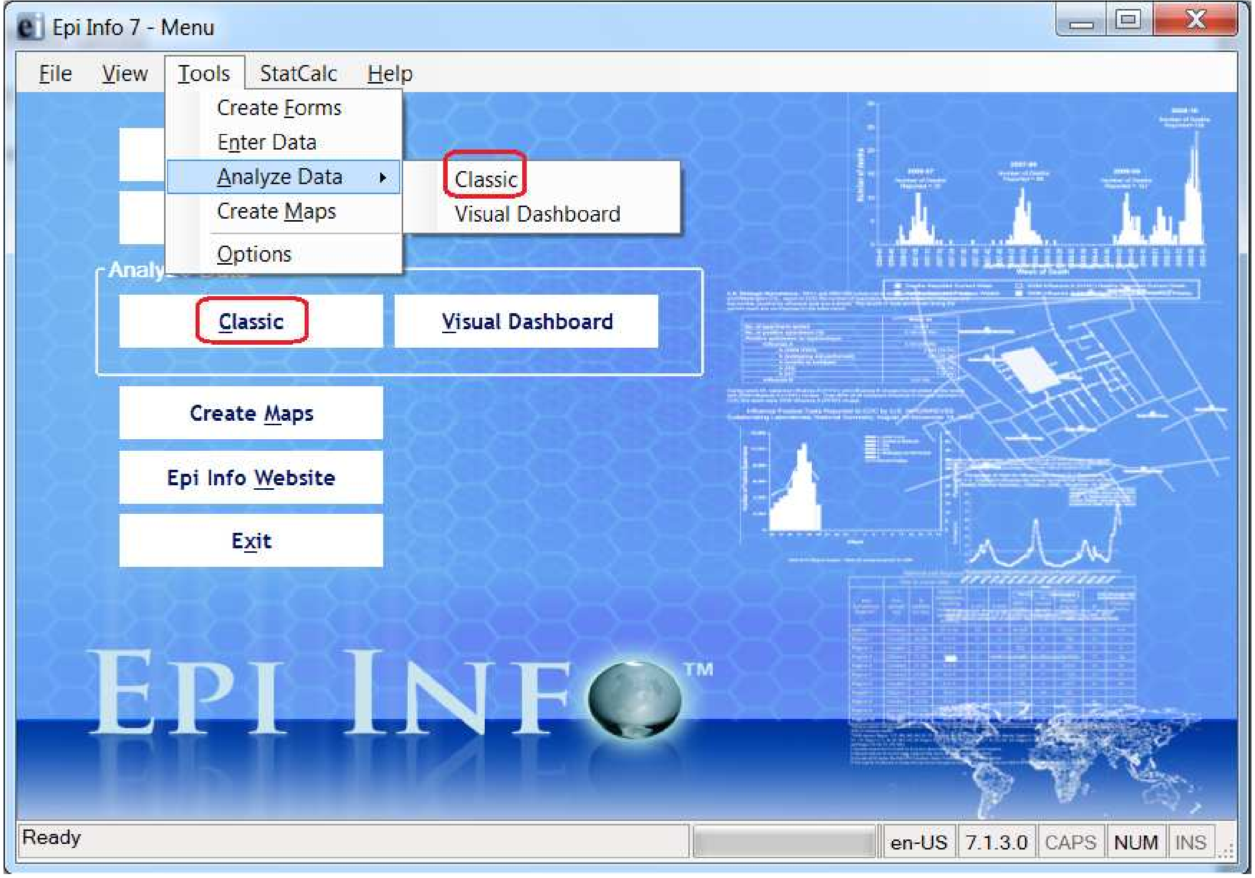
Use an Internet browser to navigate to: http://www.cdc.gov/epiinfo/installation.htm and follow the installation instructions.

Starting Epi Info

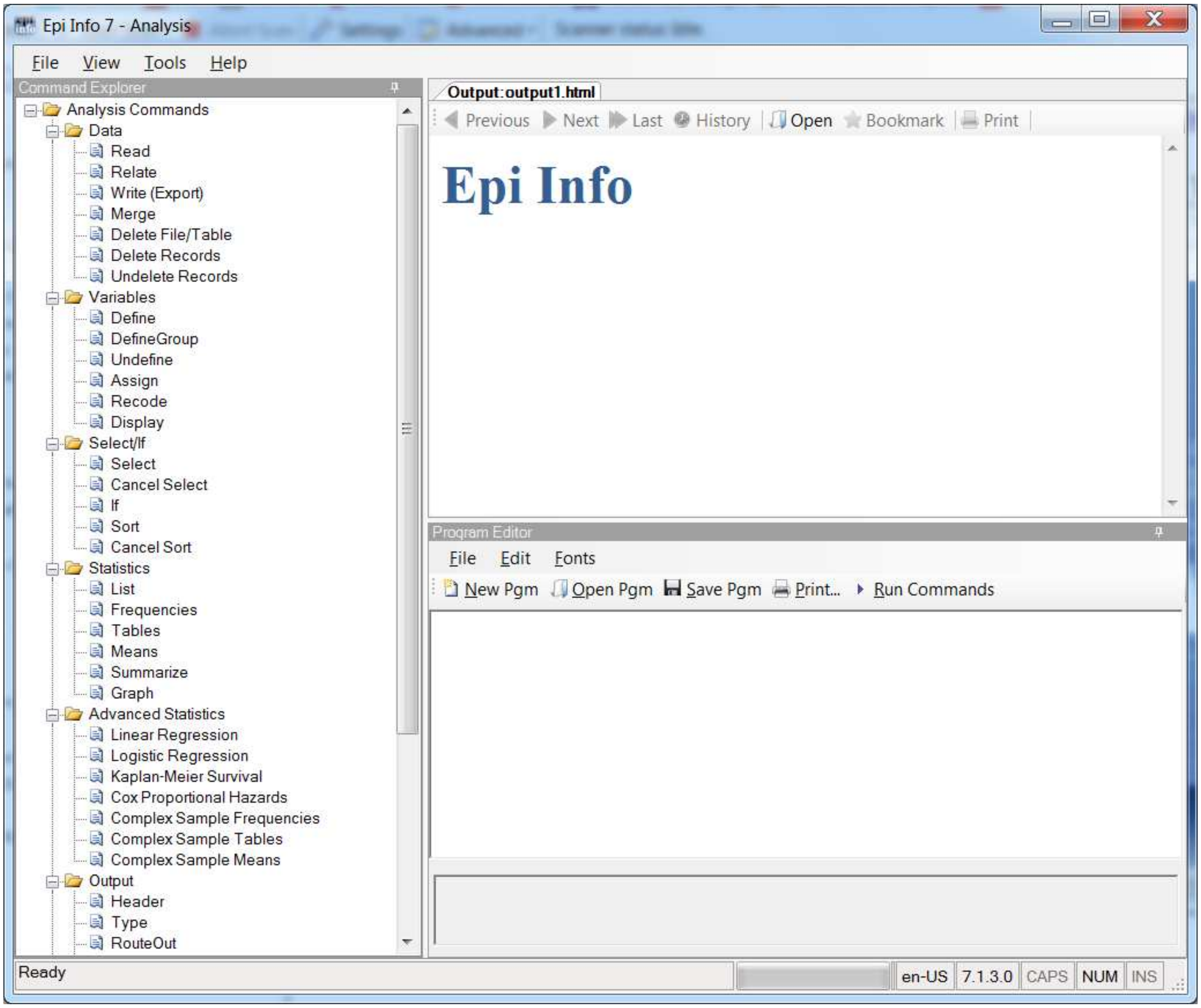
Start button, shortcut **Epi Info**^{TM} **7** or **Start button**, **All Programs**, **CDC**, **Epi Info 7**, **Epi Info**^{TM} **7**

Analyzing data in Epi Info

Click on **Classic** in the **Analyze Data section** or using the menu.

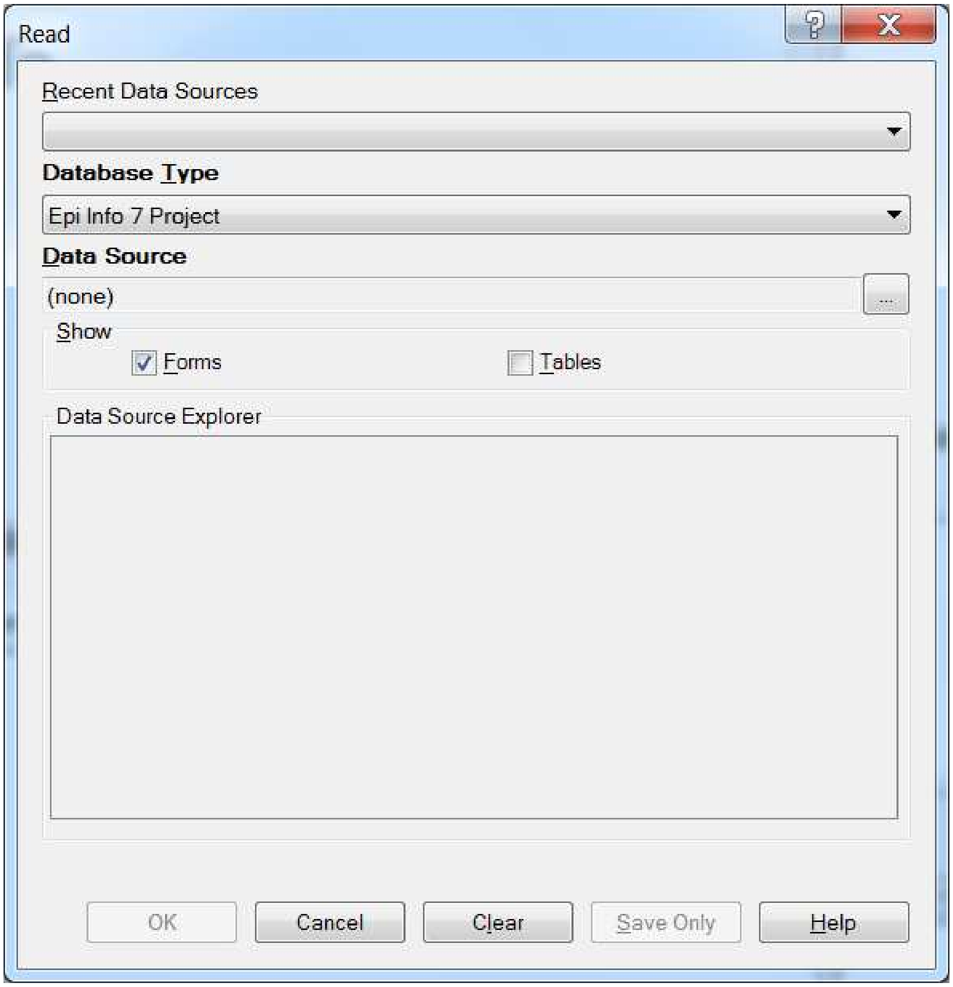


The data analysis window:



Importing an Excel file in Epi Info

In the commands section - **Command Explorer –** left hand side of the window, **Analysis Commands**, in the section : **Data** – click on **Read**. The window for importing data will open. Select the type of file to import.



Database Type

Epi Info 7 Project

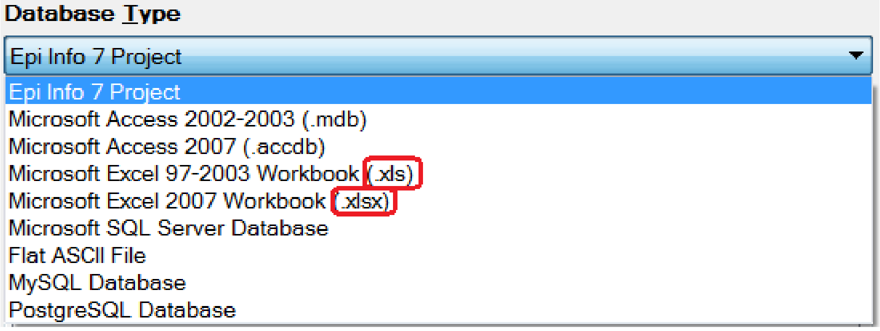
Data Source

(none)

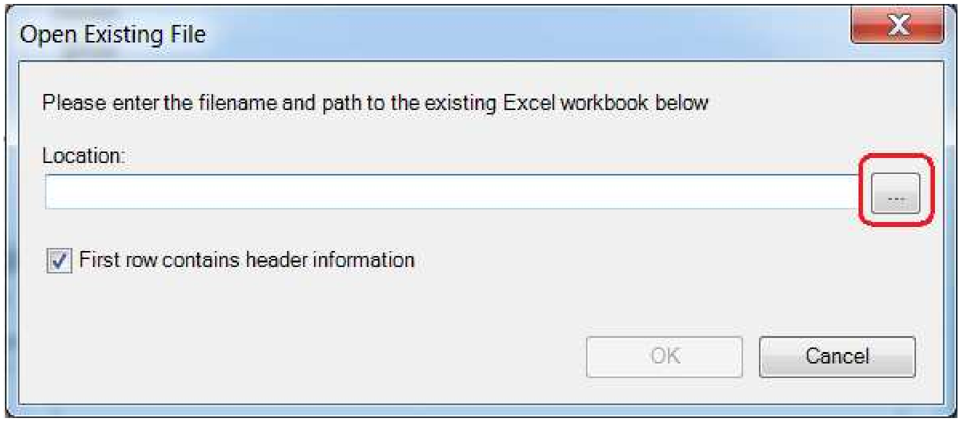
Show

☑ Forms □ Tables

If the **Excel** file has the **.xlsx**, extension we choose **Microsoft Excel 2007 Workbook (.xlsx)**, if the file has the **.xls** extension, we choose **Microsoft Excel 97-2003 Workbook (.xls)**.


Database Type
Epi Info 7 Project
Epi Info 7 Project
Microsoft Access 2002-2003 (.mdb)
Microsoft Access 2007 (.accdb)
Microsoft Excel 97-2003 Workbook (.xls)
Microsoft Excel 2007 Workbook (.xlsx)
Microsoft SQL Server Database
Flat ASCII File
MySQL Database
PostgreSQL Database

After this we search for the file to import after clicking on the button ... corresponding to **Data Source**. In the window that will open we click the button ... of the **Location** field, and we browse for the Excel file. We select the file and click **OK**, letting checked the option : **First row contains header information**.

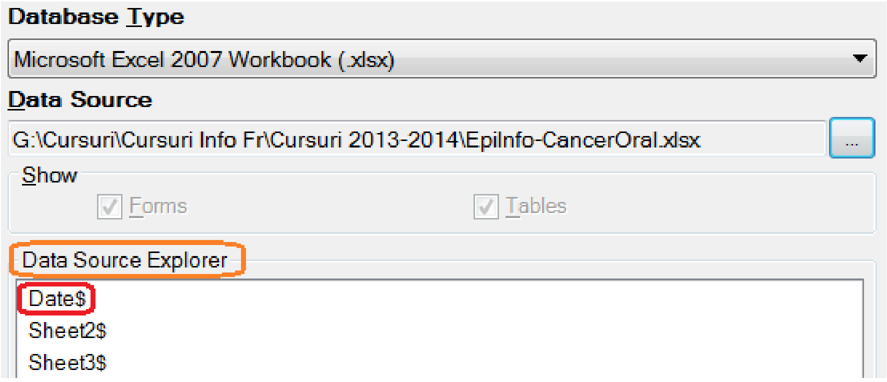


Please enter the filename and path to the existing Excel workbook below

Location:

☑ First row contains header information

After this EpiInfo shows all the worksheets in the file, in the **Data Source Explorer** section. Select the worksheet that contains the data and click the OK button to import it.



Database Type

Microsoft Excel 2007 Workbook (.xlsx)

Data Source

G:\Cursuri\Cursuri Info Fr\Cursuri 2013-2014\EpiInfo-CancerOral.xlsx

Show

☑ Forms ☑ Tables

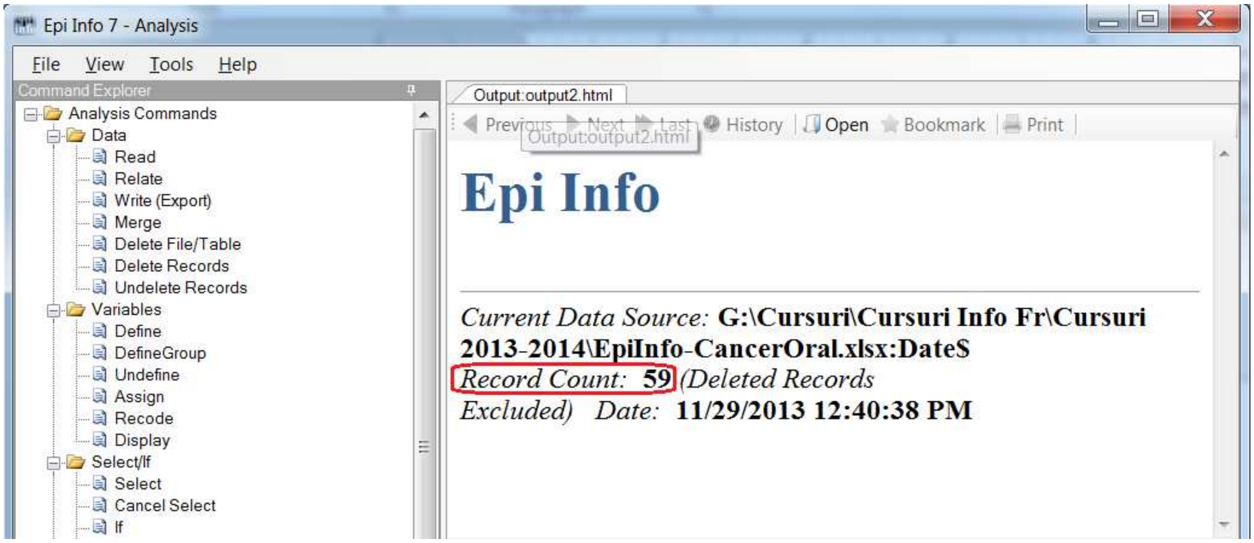
Data Source Explorer

Date$

Sheet2$

Sheet3$

After importing Epi Info shows in the results section (**Output**), the imported file name and the number of records in the file (**Record Count**).



Epi Info

Current Data Source: **G:\Cursuri\Cursuri Info Fr\Cursuri 2013-2014\EpiInfo-CancerOral.xlsx:Date$**

Record Count: **59** *(Deleted Records Excluded)* *Date:* **11/29/2013 12:40:38 PM**

Activating the *Data Analysis* module in *Microsoft Excel*

Click on a void cell, then click on ***Add-Ins*** in the ***Tools*** menu. Check the box next to ***Analysis ToolPack*** and then click OK. Select another empty cell, then search for ***Data Analysis*** in the ***Tools*** menu.

If ***Data Analysis*** does not appear in the ***Tools*** menu, despite being checked in ***Add-Ins***, uncheck the box in ***Add-Ins*** and repeat the above procedure.

Descriptive statistics

Qualitative (categorical) data:

Frequency tables

Use the ***COUNTIF*** function in ***Microsoft Excel*** to count how many times each value taken by a variable appears in the database (its absolute frequency).

E.g. to find out how many female persons (coded with F in the file) are in the sample, a void cell at the future location of the frequency table should contain a similar formula to =COUNTIF(A2:A58, "F"), if data regarding gender was recorded in cells A2 to A58. A correct frequency table should look like this:

Table 1. Gender distribution in the studied sample

Gender	**Number of subjects**
Male	20
Female	37
Total	57

Note that any table has to be labeled on top of it, using a clear and precise title. Select the table, right click on this selection and choose ***Caption*** in order to label the table. Row and column labels should be visible and easily understandable by the reader, with no need to search for further explanations in order to understand the content of the table.

Pie charts

Follow the instructions above to create a frequency table using ***COUNTIF***.

Select only the cells containing the absolute frequencies and their labels (do not select the total or column labels). Use ***Insert - Graph*** and select ***Pie***. Click ***Next***. In the ***Chart Options*** window click on the tab ***Data Labels*** and tick ***Percentage***. Continue and finish the chart wizard. *A correct pie chart should look like this:*

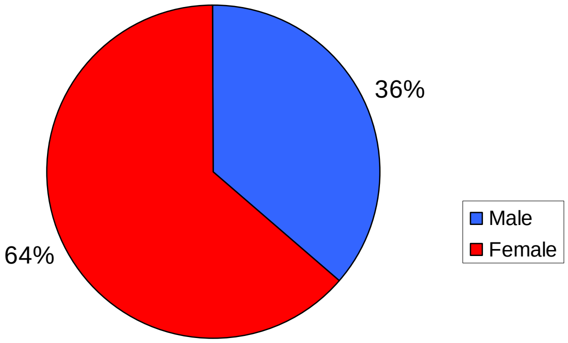


Figure 1. Gender distribution in the studied sample

Note that pie charts have to be labeled using visible percentages.

If you plan to use the chart in a ***PowerPoint*** presentation, make sure to label it on top, using a clear and precise chart title (what, how and for which subjects has been represented?).

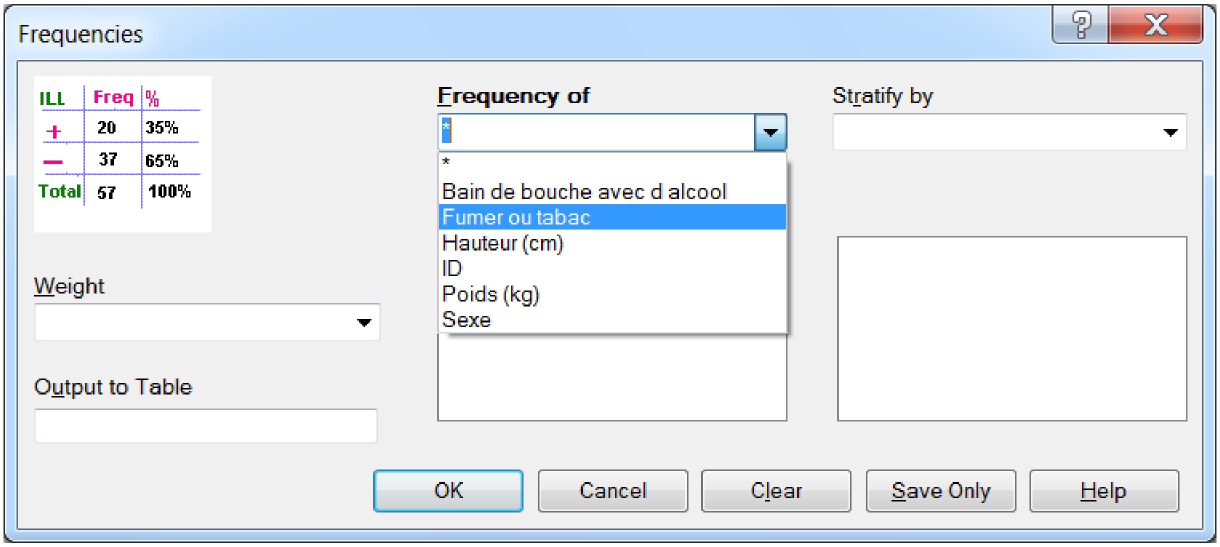
If you plan to use the chart as a figure in a ***Word*** document, erase the chart title in ***Excel*** but remember to label the chart in ***Microsoft Word***: select the figure, right click on this selection and use ***Caption***.

All labels and legend entries should be visible and easily understandable by the reader, with no need to search for further explanations in order to understand the content of the figure.

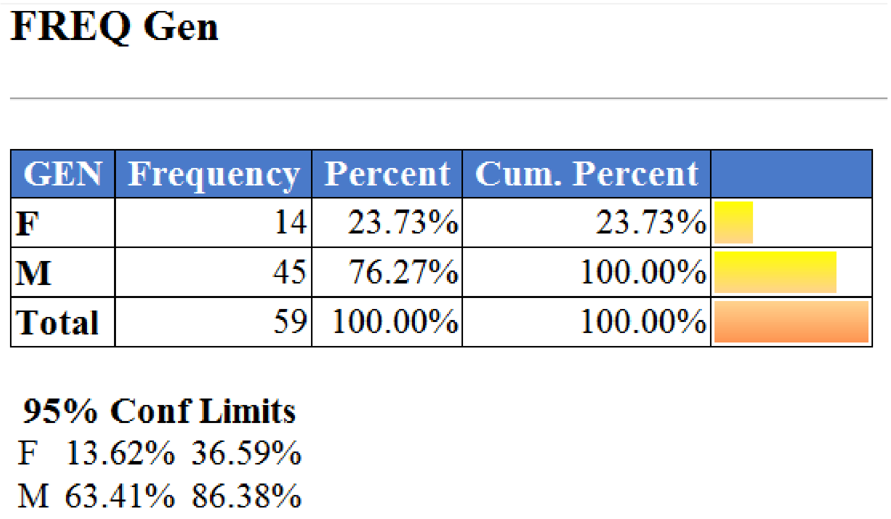
Frequency tables in Epi Info

In Command Explorer section **Statistics** we choose **Frequencies** .

In the new window we select the variable of interest in the list of **Frequency of** and we click the boutton **OK**.



In the Output we get the frequency table and the corresponding confidence intervals.



FREQ Gen

GEN	Frequency	Percent	Cum. Percent	
F	14	23.73%	23.73%	
M	45	76.27%	100.00%	
Total	59	100.00%	100.00%	

```
 95% Conf Limits
F   13.62%  36.59%
M  63.41%  86.38%
```

Contingency tables

In ***Microsoft Excel***, select any cell containing data. Then, click in the menu bar ***Data*** – ***Pivot Table*** – ***Pivot Chart Report***. Work your way through the wizard and obtain a new worksheet containing an empty pivot table and a field list.

Drag and drop the field representing a prognostic factor (the risk factor, the new diagnostic test or the new treatment, depending on the given research scenario) to the area labeled ***Drop Row Fields Here.*** Drag and drop the field representing an outcome (the disease, the reference diagnostic test or the treatment response, depending on the given research scenario) to the area labeled ***Drop Column Fields Here***. Finally, drag and drop any of the formerly used fields to the area labeled ***Drop Data Fields Here***.

Rename row and column labels so that they are easily understandable by the reader, with no need to search for further explanations in order to understand the content of the table (*e.g. If male gender was coded as m, rename the corresponding row label: male*)

Right click on a row label and select order, to correct the row order in your contingency table. Right click on a column label and select order, to correct the column order in your contingency table.

After inserting the contingency table into your ***Word*** document, remember to label it using ***Caption*** and a correct title.

The column chart associated to a contingency table

After creating a pivot contingency table, select ***Insert - Chart*** from the menu bar.

To hide the chart buttons right click the button ***Count of*** and select ***Hide Pivot Chart Field Buttons***.

To show frequency labels, right click the empty chart area towards the upper left corner, select ***Chart Options*** and, in the ***Data Labels*** tab, tick ***Percentage*** or ***Value***.

Then, switch to the ***Titles*** tab and define clear and precise titles for your chart axes, including the units of measurement between brackets, where necessary.

After inserting the chart into your ***Word*** document, remember to label it using ***Caption*** and a correct title.

Quantitative data:

Individual description of quantitative variables

Mean, median, standard deviation, 95% confidence interval for means

In ***Microsoft Excel***, use ***Tools*** – ***Data Analysis*** – ***Descriptive Statistics*** to simultaneously compute the most important descriptive parameters for selected quantitative variables.

In the ***Descriptive Statistics*** window tick options ***Summary Statistics*** and ***Confidence Level for Mean***.

To find the lower limit of the 95% confidence interval, compute ***Mean*** minus ***Confidence Level (95%)***.

To find the upper limit of the 95% confidence interval, compute ***Mean*** plus ***Confidence Level (95%)***.

Frequency table and Histogram

In ***Microsoft Excel***, use ***Tools*** – ***Data Analysis*** – ***Descriptive Statistics*** to compute minimum, maximum and range for the desired quantitative variable.

Choose a convenient bin size for the variable of interest (a round-figure for which 7-10 non-overlapping intervals of that similar size will cover the whole variable range).

Label a void column as “***Bin** Variable name (units of measurement)*” on the same worksheet as the variable of interest.

Below this label, insert the value for minimum+ the chosen bin size.

Use ***Edit – Fill – Series*** (select options: in Columns, Step value=bin size, Stop value=maximum-bin size) to complete the column containing the bin values for your variable.

Now use ***Tools – Data Analysis – Histogram***:

For ***Input Range*** select the range of cells containing the quantitative variable for which you want to plot a frequency table and histogram. For ***Bin Range*** select the newly created column. In both cases, include the column labels in your selection and tick ***Labels***.

In ***New Worksheet Ply*** write a suggestive name for the worksheet that will contain the frequency table and histogram for your variable.

In order to display the histogram you need to select ***Chart Output***.

After pressing the OK button, both frequency table and histogram will appear in a raw, unfinished form.

In order to be comprehensible, both the frequency table and the histogram need adjustments:

1. replace the upper limit of each bin shown in the brute table with the corresponding bin interval
2. delete the chart legend since its information is redundant and only takes up space
3. delete the title ***Histogram***, since you will label the figure in ***Microsoft Word***, using ***Caption***
4. resize the chart area as needed, in order to have a clear view over your histogram
5. eliminate spaces between columns by right clicking any of the columns and using ***Format Data Series – Options*** and adjusting the ***Gap Width***
6. verify the content and font size of all labels, to make sure that your histogram is easily understandable by anyone who reads your work.

A correct frequency table and a correct histogram should look like this:

Table 2. Weight distribution in the studied sample

Weight intervals (kg)	*No. of subjects*
<=40	21
(40-50]	144
(50-60]	297
(60-70]	240
(70-80]	145
(80-90]	79
(90-100]	45
>100	29

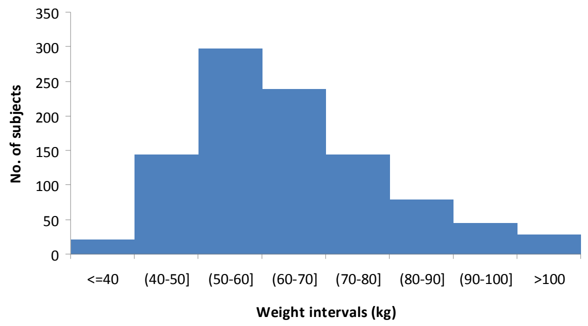


Figure 2. Histogram of weight in the studied sample

Description of a potential relation between two quantitative variables

Scatter chart

In ***Microsoft Excel***, select the columns containing the two quantitative variables, including their labels. Then select ***Insert – Chart*** and choose ***XY (Scatter)***.

Advance to step 3 of the chart wizard and define correct titles for both X and Y axes. Do not forget to specify after the title of each axis the corresponding units of measurement, between round brackets.

Then click on the ***Legend*** tab and uncheck the ***Show legend*** box, since no useful information derives from a legend when investigating only two variables at once.

In the ***Titles*** tab write the precise title of each axis, including the units of measurement in parentheses.

If you plan to use the chart in a ***PowerPoint*** presentation, make sure to label it on top, using a clear and precise chart title (what, how and for which subjects has been represented?).

If you plan to use the chart as a figure in a ***Word*** document, erase the chart title in ***Excel*** but remember to label the chart in ***Microsoft Word***: select the figure, right click on this selection and use ***Caption***.

All labels should be visible and easily understandable by the reader, with no need to search for further explanations in order to understand the content of the figure.

After finishing the chart wizard, right-click on any point from the data cloud and select ***Add Trendline***. The most common trend of data clouds is a linear one. In the ***Options*** tab check ***Display equation on chart*** and ***Display R-squared value on chart***.

To highlight the trendline using a contrasting color, right-click on the trendline and use ***Format Trendline***.

To highlight the trendline labels using a contrasting color, right-click on the label box and use ***Format Data Labels***.

A correct scatter chart should look like this:

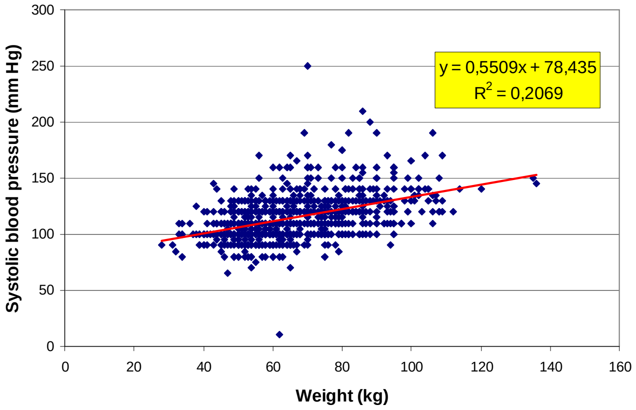


Figure 3. Relation between weight and systolic blood pressure for all subjects included in the studied sample

Survival data:

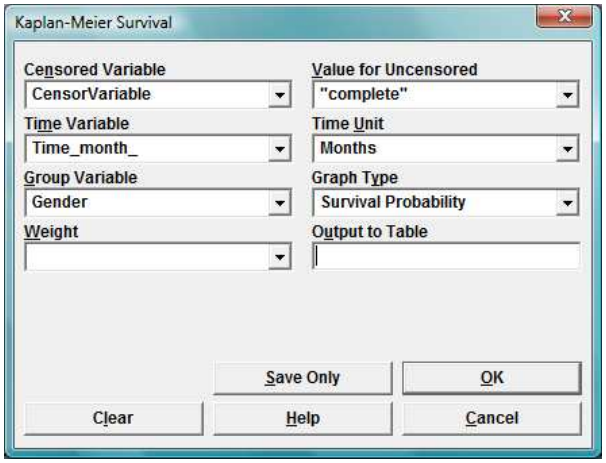
Median of survival time

In ***Microsoft Excel***, use ***Tools*** – ***Data Analysis*** – ***Descriptive Statistics***. In the ***Descriptive Statistics*** window check ***Summary Statistics***.

Survival probability chart

In the ***Analysis*** module of ***EpiInfo*** click ***Kaplan-Meyer Survival***, from the left panel.

Complete the dialog box as seen in the image below:


Kaplan-Meier Survival

Censored Variable: CensorVariable

Value for Uncensored: "complete"

Time Variable: Time_month_

Time Unit: Months

Group Variable: Gender

Graph Type: Survival Probability

Weight:

Output to Table:

Change the ***Group Variable*** as needed for your comparison.

Data Analysis

Performing a Student test (t-test) in Excel

Before performing the test you need to sort your data according to the groups that you wish to compare. (e.g. if you wish to compare cholesterol values of males with cholesterol values of females, you need to sort your data by gender).

To sort your data, click on any cell inside your data range, then use ***Data*** – ***Sort***.

If the groups that you wish to compare are independent (e.g. comparing cholesterol values of women with those of men), use ***Tools*** – ***Data Analysis*** – ***t-Test: Two-Sample Assuming Unequal Variances.***

If the groups that you wish to compare are paired (e.g. comparing cholesterol values of the same subjects, before and after treatment), use ***Tools*** – ***Data Analysis*** – ***t-Test: Paired Two Sample for Means.***

In the test window, select for ***Variable 1 Range*** the cells containing the quantitative variable corresponding to the first group (e.g. initial cholesterol values for women) and for ***Variable 2 Range*** the cells containing the quantitative variable corresponding to the second group (e.g. initial cholesterol values for men), without selecting the column label. Pay attention not to select the grouping variable (e.g. *Gender*) instead of the corresponding quantitative variable that you wish to compare.

Since the null hypothesis (H_0) of your reasoning states the absence of difference between mean values of the compared variable, introduce 0 in the ***Hypothesized mean difference*** box.

Give a title to the future worksheet that will contain the test results, by entering a suggestive name in the ***New Worksheet Ply*** box.

Immediatly after pressing OK, rename the generic lables ***Variable 1*** and ***Variable 2*** using suggestive labels: include information regarding both the quantitative variable you have compared and the grouping variable that sets them apart. This will allow you to easily interpret your test results later on.

The two-tailed p-value rendered by the test shows the statistical significance of the investigated difference between the mean values of the compared groups.

If the rendered two-tailed p-value includes the letter E followed by a negative figure this means in fact a very low (i.e. significant) p-value (e.g. p = 3,22342E-6 = 3,22342 x 10^{-6} = 0,0000032234).

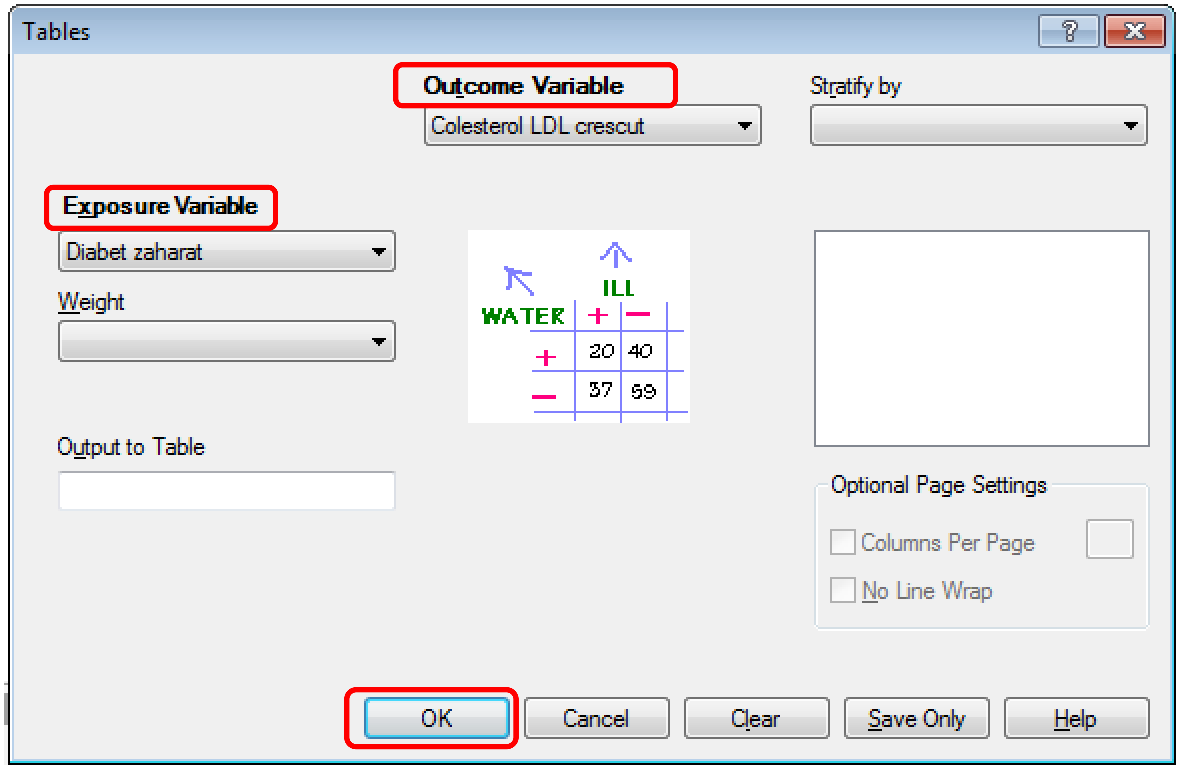
The test results also include the mean values of the compared variables. By subtracting them, you will be able to evaluate the difference between mean values, thus appraising the clinical significance of this difference.

Computing the contingency table, Risk Ratio (RR) and Odds Ratio (OR) in EpiInfo

Performing a Chi-square (χ^2) test in EpiInfo

In the ***Analysis*** module of ***EpiInfo*** click ***Tables***, from the left panel.

In the dialog window that opens, select from the drop-down lists the ***Exposure Variable*** (e.g. the risk factor, treatment, etc.) and the ***Outcome Variable*** (e.g. the disease suspected to be an outcome of the risk factor, the improvement of health suspected to be an outcome of the treatment, etc.), then press OK.



Tables

Outcome Variable: Colesterol LDL crescut

Stratify by:

Exposure Variable: Diabet zaharat

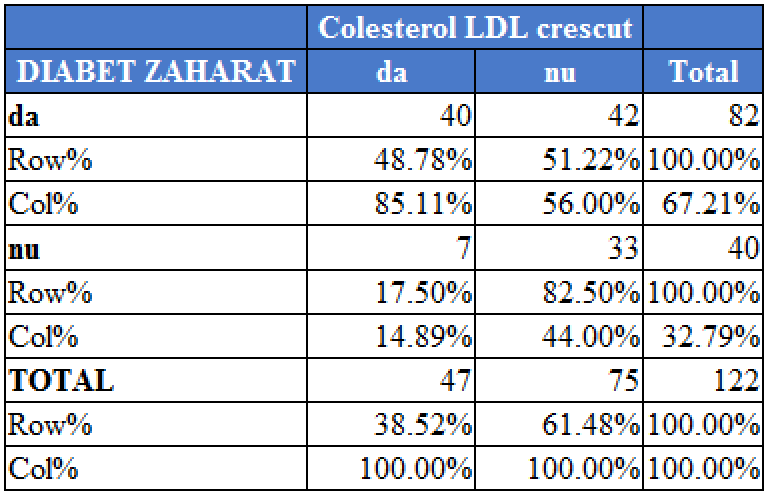
Weight:

Output to Table:

Optional Page Settings

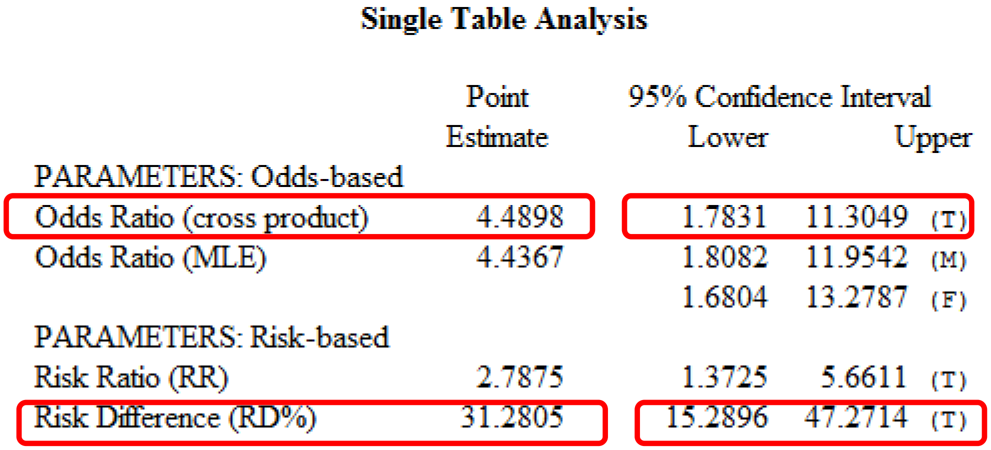
□ Columns Per Page

□ No Line Wrap



	Colesterol LDL crescut		
DIABET ZAHARAT	da	nu	Total
da	40	42	82
Row%	48.78%	51.22%	100.00%
Col%	85.11%	56.00%	67.21%
nu	7	33	40
Row%	17.50%	82.50%	100.00%
Col%	14.89%	44.00%	32.79%
TOTAL	47	75	122
Row%	38.52%	61.48%	100.00%
Col%	100.00%	100.00%	100.00%

Depending on the type of data collection used in your research scenario, interpret only the appropriate indicator (RR / OR) and its ***95% Confidence Interval*** displayed to the right of the ***Point Estimate*** of each indicator.

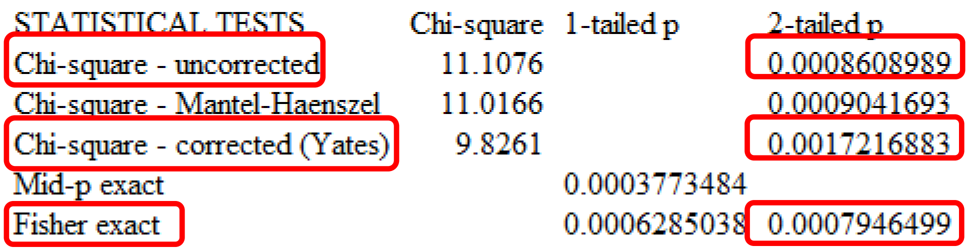


```
                         Single Table Analysis

                                  Point      95% Confidence Interval
                                 Estimate      Lower        Upper
PARAMETERS: Odds-based
Odds Ratio (cross product)        4.4898       1.7831     11.3049  (T)
Odds Ratio (MLE)                  4.4367       1.8082     11.9542  (M)
                                               1.6804     13.2787  (F)
PARAMETERS: Risk-based
Risk Ratio (RR)                   2.7875       1.3725      5.6611  (T)
Risk Difference (RD%)            31.2805      15.2896     47.2714  (T)
```

In most cases, the two-tailed p-value rendered by the ***Chi-square - uncorrected*** test shows the statistical significance of the investigated difference between the frequency distribution in the compared groups. Yet, sometimes, when one or more expected frequencies are lower than 5, a message will be displayed below the test results, telling you to interpret the p-value rendered by the ***Fisher exact*** test.

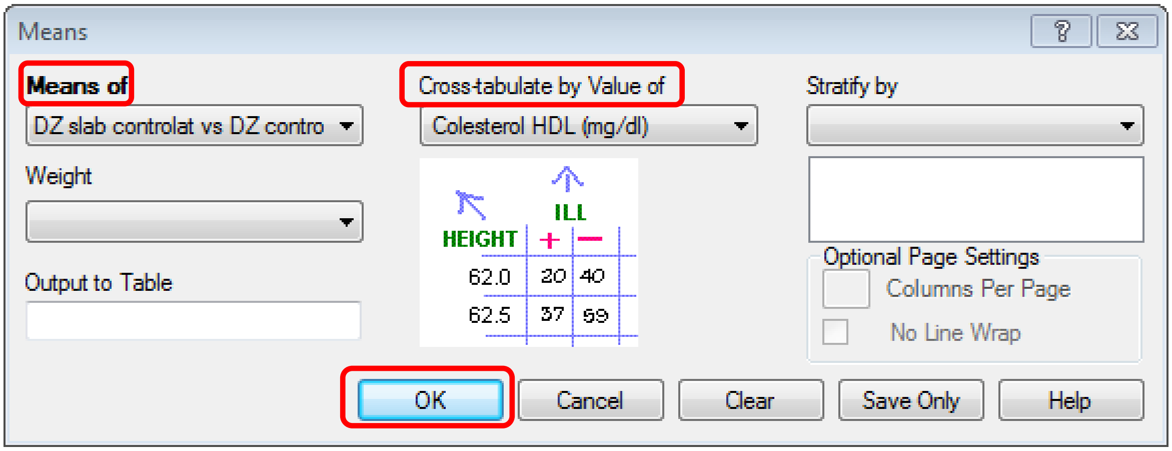
a.



```
STATISTICAL TESTS              Chi-square   1-tailed p      2-tailed p
Chi-square - uncorrected        11.1076                     0.0008608989
Chi-square - Mantel-Haenszel    11.0166                     0.0009041693
Chi-square - corrected (Yates)   9.8261                     0.0017216883
Mid-p exact                                 0.0003773484
Fisher exact                                0.0006285038    0.0007946499
```

Comparing quantitative data (Test student/ANOVA/ ...) in Epi Info

In **Command Explorer, Statistics** section we choose the command **Means**.



Means

Means of: DZ slab controlat vs DZ contro

Cross-tabulate by Value of: Colesterol HDL (mg/dl)

Stratify by:

Weight:

Output to Table:

Optional Page Settings

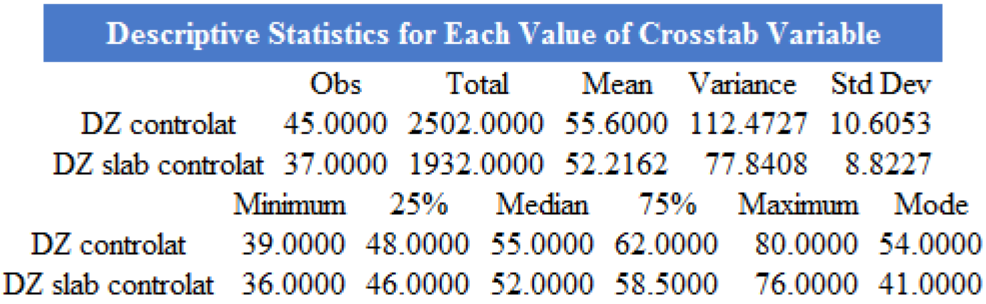
□ Columns Per Page

□ No Line Wrap

In the opened window choose the quantitative variable from the list from **Means of** and the grouping variable in the list from **Cross-tabulate by Value of**, then press the OK button.

In the window with the results (output), we have:

1. **Descriptive statistics** for groups:

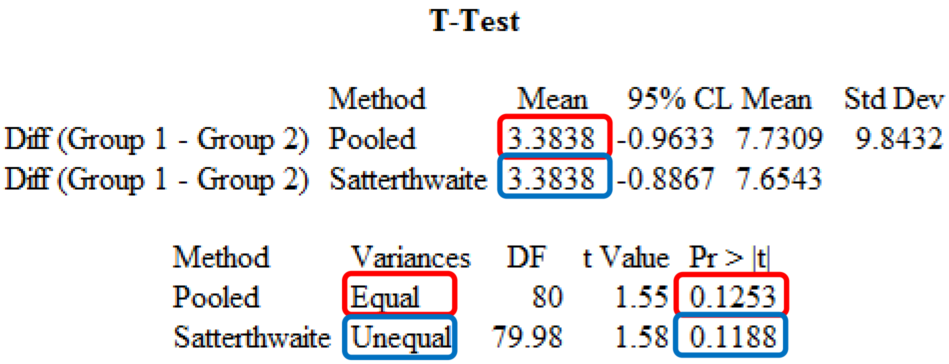


Descriptive Statistics for Each Value of Crosstab Variable

	Obs	Total	Mean	Variance	Std Dev
DZ controlat	45.0000	2502.0000	55.6000	112.4727	10.6053
DZ slab controlat	37.0000	1932.0000	52.2162	77.8408	8.8227

	Minimum	25%	Median	75%	Maximum	Mode
DZ controlat	39.0000	48.0000	55.0000	62.0000	80.0000	54.0000
DZ slab controlat	36.0000	46.0000	52.0000	58.5000	76.0000	41.0000

1. **The result of the t test** (Student) to compare the means of two independent samples with equal variances (**Pooled**) or unequal (**Unequal**) variances

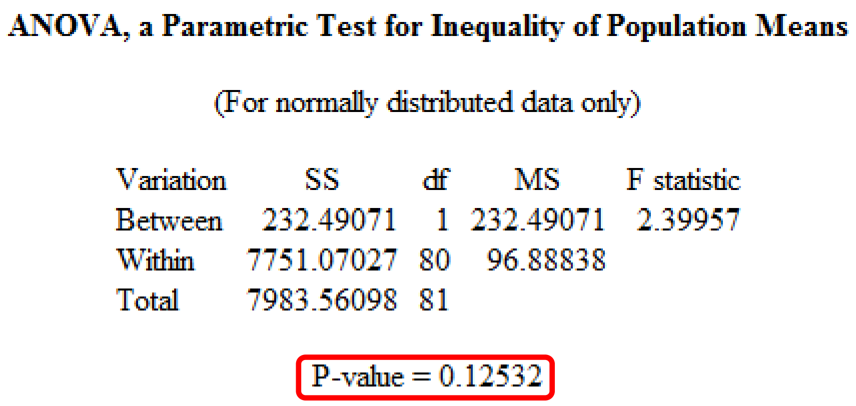


T-Test

	Method	Mean	95% CL Mean		Std Dev
Diff (Group 1 - Group 2)	Pooled	3.3838	-0.9633	7.7309	9.8432
Diff (Group 1 - Group 2)	Satterthwaite	3.3838	-0.8867	7.6543	

Method	Variances	DF	t Value	Pr > |t|
Pooled	Equal	80	1.55	0.1253
Satterthwaite	Unequal	79.98	1.58	0.1188

2. **The result of ANOVA test** to compare the averages of three or more independent samples with equal variances:



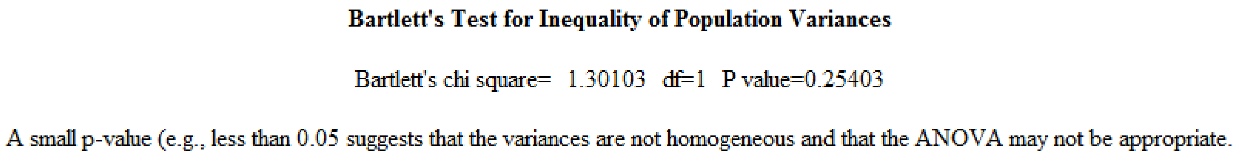
```
ANOVA, a Parametric Test for Inequality of Population Means

              (For normally distributed data only)

        Variation    SS          df   MS          F statistic
        Between      232.49071    1   232.49071   2.39957
        Within      7751.07027   80    96.88838
        Total       7983.56098   81

                    P-value = 0.12532
```

3. The result of the Bartlett test to compare the variances of two independent samples:

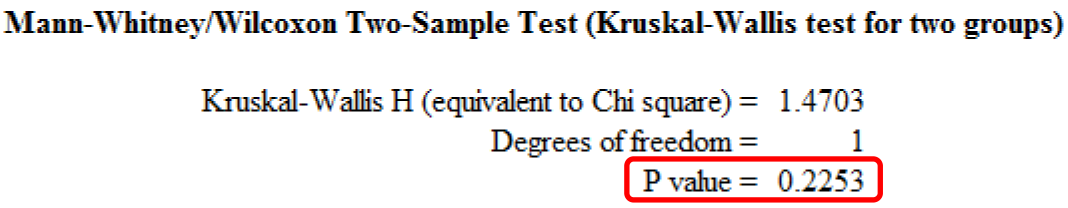


```
              Bartlett's Test for Inequality of Population Variances

              Bartlett's chi square=   1.30103   df=1   P value=0.25403

A small p-value (e.g., less than 0.05 suggests that the variances are not homogeneous and that the ANOVA may not be appropriate.
```

4. The results of nonparametric tests for comparing two independent samples (**Mann-Whitney test / Wilcoxon Two-Sample**) or more than two independent samples (**Kruskal-Wallis test**):



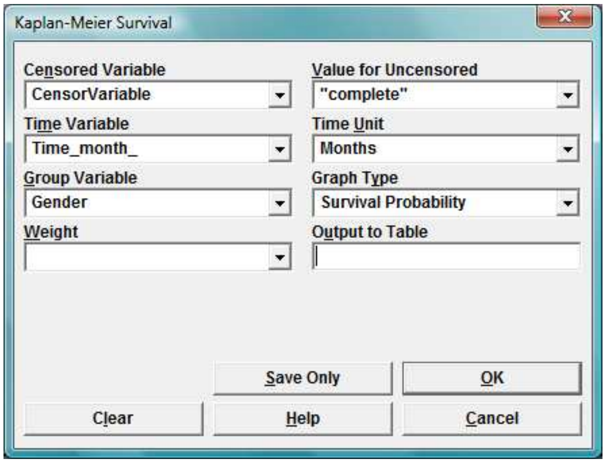
```
Mann-Whitney/Wilcoxon Two-Sample Test (Kruskal-Wallis test for two groups)

          Kruskal-Wallis H (equivalent to Chi square) =   1.4703
                                  Degrees of freedom =        1
                                             P value =   0.2253
```

Performing a Log-rank test for survival analysis in EpiInfo

In the ***Analysis*** module of ***EpiInfo*** click ***Kaplan-Meyer Survival***, from the left panel.

Complete the dialog box as seen in the image below:



Kaplan-Meier Survival

Censored Variable: CensorVariable

Value for Uncensored: "complete"

Time Variable: Time_month_

Time Unit: Months

Group Variable: Gender

Graph Type: Survival Probability

Weight:

Output to Table:

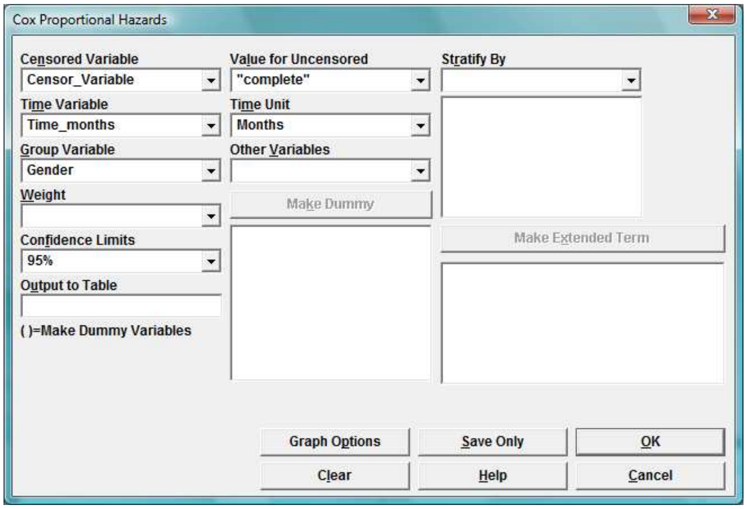
Change the ***Group Variable*** as needed for your comparison.

Below the survival chart you will find the result of the Log-rank test.

Performing a Cox Regression and computing the Hazard Ratio in EpiInfo

In the ***Analysis*** module of ***EpiInfo*** click ***Cox proportional hazard***, from the left panel.

Complete the dialog box as seen in the image below:



Cox Proportional Hazards

Censored Variable: Censor_Variable

Value for Uncensored: "complete"

Stratify By:

Time Variable: Time_months

Time Unit: Months

Group Variable: Gender

Other Variables:

Weight:

Confidence Limits: 95%

Output to Table:

( )=Make Dummy Variables

Change the ***Group Variable*** as needed for your comparison.

The Hazard Ratio (HR), its 95% confidence interval and the statistical significance of the Cox regression model will be listed below the regression model chart.

Performing a Simple Linear Regression in Microsoft Excel

Use ***Data Analysis*** – ***Regression*** from the ***Tools*** menu in ***Microsoft Excel***.

For ***Input Y Range*** select the cell range that contains the dependent variable (**y**) in your sample, the one you want to predict using a simple linear regression.

For ***Input X Range*** select the cell range that contains the independent variable (**x**) in your sample, the one you want to use in order to predict the dependent variable (**y**), using a simple linear regression.

Make sure to include in your selection the cells containing labels for both the dependent and the independent variable and check the ***Labels*** box. Also check the ***Confidence Level*** box for a 95% CI and enter a suggestive name for the new worksheet where your simple linear regression will be saved.

Performing a Multiple Linear Regression in Microsoft Excel

Use ***Data Analysis*** – ***Regression*** from the ***Tools*** menu in ***Microsoft Excel***.

For ***Input Y Range*** select the cell range that contains the dependent variable (**y**) in your sample, the one you want to predict using a multiple linear regression.

For ***Input X Range*** select the contiguous cell range that contains the independent variables ($\mathbf{x_1, x_2, \ldots , x_n}$) in your sample, the ones you want to use in order to predict the dependent variable (**y**), using a multiple linear regression. If the independent variables do not form a contiguous cell range, cut isolated variables before using ***Regression,*** and insert them into adjacent columns in order to form a contiguous cell range.

Make sure to include in your selection the cells containing labels, for all dependent and independent variables and check the ***Labels*** box. Also check the ***Confidence Level*** box for a 95% CI and enter a suggestive name for the new worksheet where your multiple linear regression will be saved.