**Lab Assignment 3**

**Aims of Lab 3**

|  |  |
| --- | --- |
| **A** | **Computing new variables using "Compute"** |
| **B** | **Changing the coding of a variable using "Recode"** |
| **C** | **Importing (reading) data from a text file without columns** |
| **D1** | **Locating outliers using a boxplot** |
| **D2** | **Selecting and deleting cases** |
| **E** | **Computing confidence interval for population means** |
| **F** | **Testing a population mean using t-test** |

**Lab 3**

**> Load (open) the data file used last week, which contained information on the variable MLU.**

**A. COMPUTING NEW VARIABLES USING "COMPUTE"**

Remember that a variable is the output of one measurement (or experiment) on different subjects (called *cases*). So "height" or "weight" or "gender" or "score obtained on some test" or "native tongue" or "reaction time" are all variables. It is, however, often necessary to derive new variables based on the existing ones, such as the sum of the scores obtained on two different tests by each subject, the ratio of the correct sentences and of all sentences for each subject, or transforming a score into a grade. *Recoding*, to be introduced in the next section of this lab, is also a kind of variable transformation.

Now we take an example that should help us also better understand the concept of standard deviation (SD). SD is sometimes compared to the mean of the (absolute value of the) deviations. The latter can also be calculated with SPSS. Yet, since it is not a standard measure, we have to go through the steps of the calculation ourselves. First, we shall introduce a new variable based on MLU, which corresponds to the distance of each data point from the mean (called the *deviation* of each data point). Then, the mean of this second variable can be simply calculated using SPSS.

**> Compute the deviations using "TRANSFORM" and "COMPUTE".**  
Hint: First, enter the name of the new variable in "Target Variable", for instance, DEV. Copy MLU to the window "Numeric Expression", then type the minus sign '-', and finally enter the mean (calculated last week; using a dot and not a comma) in the same window. Subsequently, you will see a new column appearing in the Data Editor window, containing the deviation of each data point from the mean.

Check whether the sum (i.e., the mean) of the deviations is really 0, as mentioned earlier in the course. To do that, you need to change the variable being worked with in the "Analyze" - "Descriptive statistics" - "Frequencies" window.

Afterward, have another variable calculated again, called ABSDEV, which contains the absolute values of the deviations (that is, without the negative signs).

**> Use "COMPUTE" again to obtain the absolute deviations from the mean.**  
Hint: First, enter the name of the new column. Then choose the group "Arithmetic" within the "Function group". Find "Abs" within the window "Functions And Special Variables". Finally, put the variable DEV between the parentheses of 'Abs()'.

**> Now, have SPSS calculate the mean of the new variable ABSDEV (similarly to the way done in the previous lab).**

**\* 1. Copy the mean of ABSDEV to your report.  
\* 2. Compare the SD (calculated last week) with the mean of the deviations. For what two reasons (two differences in the way they are calculated) do they differ?**

**B. RECODING A VARIABLE**

A special type of variable transformation is called *recoding*, and it is used if the raw data have been collected using a different value set from what we need for statistical purposes. One might wish to change the units of measurement from inch to centimeter, or from fractions of seconds to milliseconds.

Another example is the recoding of nominal values to numbers: Even though it is good practice to use meaningful coding systems (strings such as "m" and "f" for gender, or "eur", "ame", "afr", "asi" and "aus" for continents of origin), some statistical packages (including SPSS) allow fewer manipulations and analyzes for data encoded thus. Therefore, we may prefer to recode "m" as "1" and "f" as "2", etc. – keeping always in mind that the numerical values should not be seen as real numbers (no order between them, and no arithmetical manipulations).

We are now interested in knowing how many long MLU's there are in the text. We define an MLU as "long" if it contains more than six words. In the present case, a sample of 20 utterances, you probably would not use SPSS, but in the case of 1000 utterances the story becomes quite different... Therefore, we are going to introduce a new variable LONG\_MLU derived from MLU: LONG\_MLU is 0 if the MLU is 6 or less, and 1 otherwise. The process of changing the values of a variable in this manner is called recoding, which is especially useful in the case of questionnaires.

**> Create a new variable LONG\_MLU from the variable MLU that is 1 for original ("old") values greater than 6, and 0 else.**  
Hint: "Transform", "Recode". Always choose "Into Different Variables", otherwise you lose your original data, and you won't be able to check your computations. Copy MLU to the window, and enter the name LONG\_MLU as Output Variable. Click on "Change" to have this name in the window. Afterward, use "Old and New Values" to provide the original and the corresponding recoded values: enter an old and a corresponding new value, click on "Add", and repeat this procedure for all values. If the formula is okay in the window, click on "Continue", then on "OK".

**\* 3. Create a histogram of LONG\_MLU, and copy it to your report.**

**> For the next task, open a new data file, and close the old data file.**

**C. IMPORTING (READING) DATA FROM A TEXT FILE**

The subjects of an experiment read sentences on the screen of a computer, word by word. Each time the subject has read the word he or she presses a key. The previous word disappears and the next one becomes visible. The time elapsed between pressing the keys is the time needed by the subject to read the word.

The following values are the time in milliseconds needed to read 24 words (Source: Edith Kaan and Laurie Stowe, Developing an Experiment, 1995. Techniques and Design, Klapper vakgroep Taalwetenschappen, Rijksuniversiteit Groningen):

450 390 467 654 30 542 334 432 421 357 497 493 550 549 467 575 578 342 446 547 534 495 979 479.

The data can be found here: [words.txt](https://www.let.rug.nl/nerbonne/teach/Intro-Stats-LOT/labs/data/words.txt).

**> Place your mouse above the link and click on the right button. Choose 'Save Link As... '.  
> Save this file in your own SPSS-lab folder (directory).  
> Have a look at the structure of this file: What does it contain? How is it organized? For instance, are values delimited by some special character, such as by a space, or each value is in a new line? Does the file contain information describing the content of the file (name of the variable(s), description, source of the data, etc.)?  
> Import this file to SPSS using "File", "Read Text Data". Find the text file just being saved and open it.**

You are now offered the *Text Import Wizard* of SPSS, which is going to help you open the file.

**> Answer the questions of *Text Import Wizard*.**  
Hints: This text file does not have a Predefined Format. That is, the variables are not found in a specific column, but the values are simply delimited by a space. The file does not contain any variable name. Each case consists of a single observation (a single value). Therefore, you have to choose 'A specific number of variables represents a case' and set it to 1.

If you wish, you can also define the name of the variable, but you can do that also later.

**> Use the name RDT for the variable. Then, go to "Variable view" and use the field "label" to explain what the abbreviation RDT stands for: "reading time per word".** Observe that you will be shown the label and not the variable name in different reports returned by SPSS.

If the data import is successful, you have a variable (column) with 24 numbers.

**> Set in the Variable View the number of decimals for this variable to 0 (as the reading time has been measured with the precision of 1 msec, so the values are always integer).**

**> Save these data as a usual data file, that is, in the native SPSS format .sav.**

**D. LOCATING OUTLIERS USING A BOXPLOT.**

**\* 4. Create a histogram including a Normal curve, as well as a boxplot of RDT. Copy it to your report.  
\* 5. You can find two outliers among your data. Which are they, and what kind of explanation(s) could you provide to explain them?  
\* 6. In case you decide to remove these cases from your data set, do you expect the mean or the standard deviation to change more? Why?**

**> Remove these cases from your data file by selecting the corresponding rows (click on the gray case number on the left), and then press the DELETE key.**

**> Calculate the mean and the SD again by creating a new histogram.**

**\* 7. What can you observe, as compared to your previous results?**

From now onward we shall work on these data with the outliers being removed.

**E. CALCULATING A CONFIDENCE INTERVAL FOR THE MEAN**

**\* 8. You know the size of the sample, and you know its standard deviation. What is the standard error, then? Calculate it both by hand (give details of your calculation in the report) and let SPSS calculate it for you. Are the two values the same?**  
Hint: "Analyze", "Descriptive Statistics", "Frequencies". Choose "Statistics" and SE. Do not forget to turn off "Display Frequency Table".

Now we turn to [Table D](http://bcs.whfreeman.com/ips6e/content/cat_050/ips6e_table-d.pdf) of Moore and McCabe. Having calculated the standard error, let us find the confidence interval for the mean of the variable RDT. Let us set the confidence level to *C* = 95%.

**> Determine the *degree of freedom* (*df*) of the sample.  
> Use Table D to determine the z\* and the t\* corresponding to the level of confidence *C*.**

**\* 9. Determine the confidence interval for the mean of the sample using the Student-t-statistic. Provide details of your calculations in your report.  
\* 10. What is the meaning of this confidence interval?  
\* 11. Why have we used the t-statistic and not the z-statistic?  
\* 12. Suppose we know that the population standard deviation happens to be the same as the standard deviation of the sample. Determine the confidence interval using the z-statistic for this case.**

**\* 13. Now have the confidence interval calculated for you by SPSS. Copy the values returned by SPSS to your report. Is it different from your calculations?**  
Hint: "Analyze", "Compare Means", "One Sample T-test".

The last two columns of the table present the lower bound and higher bound of the confidence interval as a difference from the test value. If you set the test value to 0, then the last two columns will give you simply the bounds of the confidence interval. If, however, you set the test value to the sample mean, then the last two columns will show you how much you have to add to, and detract from the sample mean to find the confidence interval, in which the population mean lies with the given confidence level.

In "Options" you can set the confidence level.

**> Repeat the procedure of having SPSS compute the confidence interval with a t-test, but this time with a confidence level of 99%.**

**\* 14. Add again to your report the higher and lower values between which the population mean must lie. Why *and* how is this confidence interval different from the previously calculated one?**

**F. TESTING A HYPOTHESIS ON POPULATION MEANS USING A T-TEST(OneSample).**

Suppose there are two competing theories about reading. They associate reading with two different neural mechanisms and therefore they have two different predictions about reading speed of the particular words employed in this experiment. Theory FRT ("fast reading theory") predicts that the average time needed to read these words is at most 440 msec, whereas theory SRT ("slow reading theory") predicts a reading time of at least 505 msec (always including the time needed to press the button).

Are your data able to refute or corroborate any of these theories at a significance level of *alpha* = 5%? A hint: use the above theories as null hypotheses; so you ask whether you can refute them, or whether your data are consistent with them (hence, they corroborate them). Please provide your calculations both by hand, and using SPSS.

**\* 15. In each of the two cases, what is the null hypothesis exactly, and what is the alternative hypothesis? (In words/one full sentence, please.)  
\* 16. Are you using a one-sided or a two-sided test?  
\* 17. Perform the test for both cases by hand, and describe the steps of your calculation.  
\* 18. Let SPSS calculate the test for you and copy the results.  
\* 19. What is the meaning of the P-value in each of the two cases? (Hint: the probability of exactly what is it?) Please write one full sentence for each case in your report.  
\* 20. For each case, provide *the* key sentence summarizing the results of the statistical analysis, as it is done in scientific papers. That is, either "*Based on our data, we can reject the null-hypothesis at a significance level alpha = 0.05, that is, we can conclude that [the alternative hypothesis in words] is true (t = ..., df = ..., P = ...)*" or "*our data do not provide sufficient evidence to reject the null hypothesis, that is, to conclude that...*".**