Practical "Handling data, sampling and estimation"
P1

1. Dataset *serum.sav* contains data on serum triglyceride concentration in cord blood for a sample of 282 babies (Bland and Altman, 1996). To analyze variable *serumtrigl* go to *Analyze* $\rightarrow$ *Descriptive Statistics* $\rightarrow$ *Explore*. Move variable *serumtrigl* to the *Dependent List* box. There are default options selected for both *Statistics* and *Plots*. Within *Plots*, click on the *Histogram* checkbox, which is not automatically selected. Run the analysis. You will obtain some plots together with the following table:

Table 1. Descriptives for serumtrigl

Descriptives

			Statistic	Std. Error
Serum triglyceride concentration in cord blood	Mean		.4827	.01973
	95% Confidence Interval for Mean	Lower Bound	.4439	
		Upper Bound	.5216	
	5% Trimmed Mean		.4510	
	Median		.3797	
	Variance		.110	
	Std. Deviation		.33133	
	Minimum		.08	
	Maximum		1.97	
	Range		1.89	
	Interquartile Range		.34	
	Skewness		1.541	.145
	Kurtosis		2.630	.289

The 95% confidence interval reported corresponds to the confidence interval for the population mean serum triglyceride concentration. We saw in the theory slides that, for a normal distribution of the population data, or a large sample size, this is:

$$\bar{x} - t_{n-1,0.975}\frac{s}{\sqrt{n}}; \bar{x} + t_{n-1,0.975}\frac{s}{\sqrt{n}}$$

with $t_{n-1,0.975}$ the 97.5% percentile of the Student's t with $n-1$ degrees of freedom (n is the sample size).

(a) Obtain the standard error reported in the table using the sample size and the standard deviation of the sample.

(b) Looking at the histogram and the box plot, what would you say about the skewness of the distribution? You can additionally ask for the best fitting normal distribution in the histogram if you go to *Histogram* under the *Graphs* menu and click on the *Display Normal Curve* checkbox.

(c) Even though the sample size is large enough for the sample mean distribution to be normal (Central Limit Theorem), a transformation would be convenient to improve the approximation of the sample distribution to a normal distribution. Why would a logarithmic transformation be useful?

(d) Create a new variable (e.g. *logserumtrigl*) with the log transformed *serumtrigl* (use e.g. a base 10 logarithm). To do so, go to *Transform* $\rightarrow$ *Compute Variable*. Repeat the same exploratory analysis you did for variable *serumtrigl*. What can you say about the histogram and box plot for *logserumtrigl*? What happens if you use a natural logarithm instead?

2. Dataset *btgdiabet.sav* contains data on urinary β-thromboglobulin (β-TG) excretion (measured in ng/100 ml creatinine per day) in 12 normal subjects and in 12 diabetic patients. (Kirkwood 2003).

(a) Analyze variable *btg* using the *Explore* option. Do you notice anything peculiar in the data by looking at the output table? In that case, obtain a frequency table of variable *btg* to see if there are more peculiarities. Fix this/these value/s and save the modified file under a new name.

(b) Analyze variable *btg* separately by group (move variable *group* to the *Factor List* box). Is the distribution of the sample skewed for any of the two groups? Are the standard deviations of the two groups similar?

(c) If we had to analyze this data using a method that requires homogeneous variances and normality, would you transform the data first to apply this method? If so, which transformation would be the most suitable? Redo the exploratory analysis with the transformed data and discuss what has changed, in case a transformation was deemed necessary.

3. We randomly pick 10 people. What is the probability that we have 7 women and 3 men?

4. Suppose the average number of patients who receive radiotherapy within one day is 8. What is the probability that tomorrow less than 4 patients will receive the treatment?