22 August 2016
Erik Parner

PhD Course in Basic Biostatistics
Exercises, day 1

Exercise 1.1

In this exercise we will go through the statistical analysis made in the lecture[1]. Note that Stata commands are written with a special `font`, and enclosed with "-".
You will find the BMI data used at the lecture in the data file *bmiwomen.dta* and a file containing some Stata commands in *day1.do*.

1. Make a histogram with a normal curve and a Q-Q plot of the BMI measurements.

2. First make a short description of BMI using the command -`summarize`-, then a longer one using the option `detail` of the same command.

3. Use -`ci`- to find a 95% confidence interval for the mean BMI. Argue from the confidence interval, why a hypothesis that the mean BMI is equal to 24.1 kg/m2 must be rejected.

4. Use -`ttest`- to test the hypothesis that the mean BMI is equal to 24.1 kg/m2. Notice that you can also find the 95% confidence interval in this output.

In Stata you can also calculate confidence intervals and t-tests "by hand" directly from summary statistics.

5. Run the commands:

```
cii 85 24.99 3.58
```

and

```
ttesti 85 24.99 3.58 24.1
```

Compare with what you got in 3 and 4.

[1] Note that the first exercise is always a repetition of the analysis presented at the corresponding lecture. If you are confident that you do not need this repetition, feel free to skip it.

Exercise 1.2
The data set *trigly.dta* contains the cord blood serum triglyceride level for 427 babies from five samples:

```
table sample, c(count trigly mean trigly median trigly sd trigly) row

------------------------------------------------------------------
   sample |    N(trigly)  mean(trigly)   med(trigly)    sd(trigly)
----------+-------------------------------------------------------
        1 |           37      .4962162           .48      .2001437
        2 |           41      .4697561           .44      .1897563
        3 |           33      .4845454           .42      .1796366
        4 |           34      .5008823           .45        .22804
        5 |          282      .5058865           .46       .219138
          |
    Total |          427      .4995316           .46      .2121913
------------------------------------------------------------------
```

1. Find a 95% confidence interval for the mean based on sample number 5. Compare the interval with the interval based on the four other samples (see table 1.2 below).

2. Find a 95% prediction interval for a new observation based on all the data. Compare with the intervals based on each of the samples.

3. Find the percentage of babies with a value below the lower limit and the percentage of babies with a value above the upper limit in the prediction interval based on the whole sample (Hint: List the data after sorting them, or use the command -`count`- combined with an appropriate if-statement). Is it fair to interpret the computed intervals as 95%- prediction intervals?

4. Make a histogram and QQ-plot for the whole data set. Discuss the validity of the assumptions behind the analyses you have made.

Table 1.2		95% **C**onfidence **I**nterval		95% **P**rediction **I**nterval	
sample	mean	Lower limit	Upper limit	Lower limit	Upper limit
1	0.4962	0.429	0.563	0.104	0.888
2	0.4698	0.410	0.530	0.098	0.842
3	0.4845	0.421	0.548	0.132	0.837
4	0.5009	0.421	0.580	0.054	0.948
5	0.5059			0.076	0.935
total	0.4995	0.479	0.520		

Exercise 1.3

The data set *siblings.dta* contains information concerning the birth weight and sex of two siblings (not twins). A short summary:

```
codebook

--------------------------------------------------------------------------------
sex1st                                                       gender, first child
--------------------------------------------------------------------------------
                  type:  numeric (byte)
                 label:  sex
                 range:  [1,2]                        units:  1
         unique values:  2                        missing .:  0/1000
            tabulation:  Freq.   Numeric  Label
                           503         1  boy
                           497         2  girl

--------------------------------------------------------------------------------
sex2nd                                                      gender, second child
--------------------------------------------------------------------------------
                  type:  numeric (byte)
                 label:  sex
                 range:  [1,2]                        units:  1
         unique values:  2                        missing .:  0/1000
            tabulation:  Freq.   Numeric  Label
                           496         1  boy
                           504         2  girl
--------------------------------------------------------------------------------
weight1st                                           weight in grams, first child
--------------------------------------------------------------------------------
                  type:  numeric (int)
                 range:  [1185,5150]                  units:  1
         unique values:  276                      missing .:  0/1000
                  mean:   3500.93
              std. dev:   519.006

           percentiles:        10%       25%       50%       75%       90%
                              2900      3200      3530      3850      4100
--------------------------------------------------------------------------------
weight2nd                                          weight in grams, second child
--------------------------------------------------------------------------------
                  type:  numeric (int)
                 range:  [1474,5250]                  units:  1
         unique values:  299                      missing .:  0/1000
                  mean:    3674.9
              std. dev:   506.352

           percentiles:        10%       25%       50%       75%       90%
                              3060      3350      3650      4000    4312.5
```

In the following we will look at the difference between the birth weight of the second and the first child, i.e. you should generate a new variable defined as

```
generate wdif = weight2nd - weight1st
```

Let us now only consider siblings where both are male (Hint: drop the other observations from the dataset).

1. Describe the data.
   Make a dotplot, boxplot, histogram and QQ-plot of the differences in birth weight.
   Find summary statistics (number of observations, mean, sd, etc.).
   Make a short note of your findings. Do any observations stick out?

2. Find a 95% confidence interval of the mean difference in birth weight. Is the *mean* birth weight different for the two brothers? If so, how different?

3. Find a 95% prediction interval of the difference in birth weight. Write an interpretation of this interval.

4. Test the hypothesis of no difference in the mean birth weight for two brothers.

5. Write a summary of the analysis.

Now consider siblings who are brother and a sister, where the boy is the oldest (i.e. born first)

6. Repeat question 1 to 5.

7. Discuss the similarities and differences in what you found above. (We will return to this problem during exercises on day 2).

Exercise 1.4

We now return to the triglyceride data from exercise 1.2 above.
Generate a new variable containing the natural logarithms of the triglyceride levels and call it `lntri`.

1. Make a histogram and QQ-plot of the `lntri` variable for the whole data set. Compare with what you saw in question 4 in exercise 1.2.

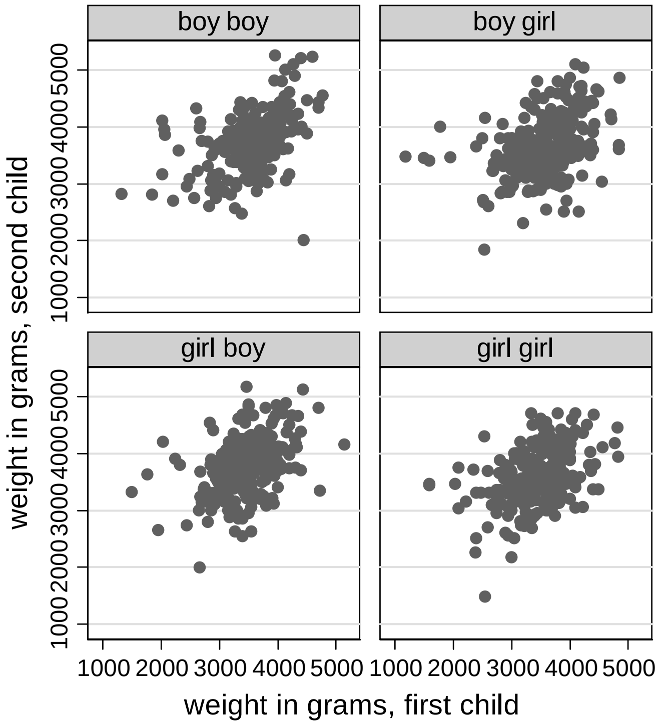
2. Find a 95% prediction interval for a new observation (of log triglyceride) based on all the data.

3. Find the percentage of new born with a log triglyceride value below the lower limit and the percentage of new born with a value above the upper limit in the prediction interval. Is it fair to term this a 95%- prediction interval?

Exercise 1.5

Now, back to the sibling data set. Run the following commands on the full dataset *sibling.dta* to generate the plot on the following page:

```
egen sex1sex2=group(sex1st sex2nd),label
scatter weight2nd weight1st,by(sex1sex2) aspect(1) scheme(s1mono)
```



Graphs by Gender of first and second child

1. Based on the plots above and your general knowledge, discuss why it is not reasonable to assume that the birth weight of two siblings are independent.