Introduction

The table below shows the birthweight in kilograms of 14 children born to mothers who smoked during pregnancy and 15 children born to mothers who did not smoke during their pregnancy. The data are stored in the Stata file **preg.dta** which can be downloaded from the AT Moodle page.

smoke	Mother's smoking status (smoker/non-smoker)
bwt	Birthweight of baby, kg

Mothers who smoked		Mothers who did not smoke	
3.08	3.06	3.99	3.93
2.74	2.75	4.29	2.54
2.90	3.2	3.70	4.03
3.17	3.13	3.73	3.06
3.35	2.98	3.29	3.54
3.02	3.21	2.95	3.61
3.23		4.18	2.91
3.18		2.71	

5.1 Data management and exploration

a. Launch Stata, change the working directory to the AT folder and load the data set **preg.dta**. Explore the distribution of each variable. Are there any missing or unusual values? What are the numeric codes for smokers and non-smokers?

5.2 Comparison of two population means

a. Fill in the following table using `summarize` with the `bysort` prefix, or `tabstat`, e.g.

```
bysort smokeyn:summarize weightkg
tabstat weight , by(smokeyn) stat(n mean var)
```

	Smoking group (i = 1)	Non-smoking group (i = 2)
Sample size, n_i		
Sample mean, $\bar{x}_i$		
Sample variance, $\hat{\sigma}_i^2$		
Estimated pooled variance $\hat{\sigma}^2$		

b. Using the results in the table, **by hand**, carry out a two-sample t-test and provide a 95% CI for the difference in the means assuming a common variance. Use Stata to check your result.

```
ttest bwt, by(smoke)
```

Interpret your result.

c. What are the assumptions of the test in (b)? Do you think that they are satisfied?

d. How would you carry out, by hand, a test of the assumption of equal variances? Carry out the test in Stata:

```
sdtest bwt, by(smoke)
```

What do you conclude?

e. As there is doubt about the assumption of the equality of variances, use Stata to carry out a procedure to compare the means that does not make this assumption.

```
ttest bwt, by(smoke) unequal
```

How have the results changed?

f. Create a box-plot of birth-weight by mothers smoking status in Stata.

```
graph box bwt , over(smoke)
```

Explain how the appearance of the box-plot relates to the results of your analysis.

g. Note that the largest observation in the group that did not smoke was 4.29 kg. Investigate what happens to the result of your t-test when this is changed to 5.29 kg, then to 6.29 kg and so on up to 14.29 kg. Before doing this think about what you might expect to happen. Does the behaviour differ if tests assuming i) equal, ii) unequal variance are used?

To change highest value in Stata:

```
replace bwt=5.29 if bwt==float(4.29)
```

then,

```
replace bwt=6.29 if bwt==float(5.29)
```

etc.

5.3 Comparison of population proportions

Here we will define a low birth weight baby as birthweight < 3kg and investigate whether there is any association between mother's smoking and low birthweight.

a. Generate a binary variable called lbw that takes the value 1 if birthweight is less than 3kg and 0 otherwise.

```
gen lbw=(weight<3) if bwt<.
```

b. Obtain a two-way frequency table of low birthweight and mother's smoking status. What is the proportion of low birthweight in each group?

c. Use Stata to carry out an appropriate test of the null hypothesis that the risk of a low birth weight baby does not differ between groups.

d. By hand, calculate the relative risk and its 95% CI of low birth weight comparing smoking mothers with non-smokers. What do you conclude?

e. Use the `csi` command in Stata to check your results in (d). Interpret the results.

```
csi 4 4 10 11
```

The four numbers are the number of cases in the exposed and unexposed groups respectively followed by the number of non-cases in the exposed and unexposed groups respectively. Here "exposed" is smoking and "cases" are the low birthweight babies.