Exam
Statistical Models in Science, June 2013

Massimiliano Tamborrino
mt@math.ku.dk

Rules for the exam

These are the rules about the exam. Please, read them carefully and make sure you do what it is written!

- The exam has to be re-hand in on Absalon within **Thursday the 13th June at 11.59am** (24hours).
- Submissions after the deadline would be evaluated as -3.
- Specify your **name** and **exam number** in the solution to the exam.
- The exam is an individual exam. You are **not allowed to talk to each other about the exam**, communicate/share procedures, solutions, results. It is also not allowed to write joint notes or solutions. Violation of these rules will be considered cheating and will be reported to the program director (so please avoid this uncomfortable situation).
- The statistical analysis of data has to be done with R.
- You have to return a single pdf-file, where you report both the solution of the exam, the R-code and R-output.
- When you write your answer, be careful that it is clear what you have done and why. There is no need to write long essays: rather short and precise than long and weaving. A conclusion without argumentation is not sufficient and would be considered incomplete/wrong.

- Relevant R-codes and R-outputs must be attached. You can of course make references to the book, notes, slides or exercises that have been made in the course.
- Once again: make sure you write down the calculations you make (if you do by hands), and/or the R-code and R-output (if you use R). If you use R, relevant estimates, confidence and predicted intervals, p-values, etc. should be taken from the R-output. Comments like "the estimates are those obtained using summary(mod) in R"are not sufficient.
- You may write the solution either in Danish or in English.
- If you have any doubt or question about the exam, write a post in the Discussion Board, such that everybody may benefit from that.

Exercise 1

Data consist of rainfall from 52 clouds. During the experiment, 26 clouds were affected with silver iodide, while the other 26 were control clouds which were not affected. The rainfall is measured in *acre-feet*, which is the amount of water required to cover an area on 1 *acre* (4047 m^2) at a height of 1 foot (0.305 m).

The data (available in the file rain.txt on Absalon → Exam) have two variables:

- **Treatment** with values **control** and **silverIodide**
- **Rain** indicating rainfall

When the dataset is loaded and attached, use the following commands to define the two variables containing rainfall for the two groups of clouds, as well as their natural logarithm values:

```
control <- rain [treatment == "control"]
silverIodide <- rain [treatment == "silverIodide"]
logControl <- log(control)
logSilverIodide <- log(silverIodide)
```

In the first part of the assignment, use only data from the 26 control clouds.

Q1. Consider the following two statistical models:

   A. The values in `control` are independent and normally distributed with mean μ and standard deviation σ.

   B. The values in `logControl` are independent and normally distributed with mean μ and standard deviation σ.

   Decide which model – A or B – is the most appropriate for the data.

Q2. Specify the estimates of the unknown parameters μ and σ, and the standard error of μ. Determine also the 95% confidence interval for the mean value μ. Report μ and its confidence interval in terms of the original scale.

Q3. The rainfall of a new cloud is 200 acree-feet. Is that an unusual value?

In the next part of the exercise, you should use data from all the 52 clouds.

Q4. Write a statistical model that makes it possible to compare the rainfall from clouds with and without the influence of silver iodide.

Q5. Perform a hypothesis test for the hypothesis that the logarithm of rainfall for the silver iodide group is higher than the control one.

Q6. Give an estimate and a 95% confidence interval for the difference between the expected values of the logarithm of rainfall for the two groups of clouds. Compare the results with Q5.

Exercise 2

I am claiming that there is a dependence between the average work load per week of a student taking the StatNat course and the percentage of correct answer in his/her exam. The data (available in the file percentage.txt on Absalon → Exam) have two variables:

- **Hour** indicating the average work load per week of a student.
- **Perc** indicating the percentage of correct answers in the exam.

Q7 Write a statistical model, specifying also the model assumptions, that makes it possible to examine if my claim is true.

Q8 Specify the estimates of the unknown parameters of the model and their standard errors. Determine also the 95% confidence intervals for the unknown parameters.

Q9 Set up and test a hypothesis claiming that the intercept is 0. In case you cannot reject the null hypothesis:

  - Write down a new proper statistical model (with all its assumptions)
  - Specify the estimate of the unknown parameter and the standard error.

From now on, you will work ONLY with the model which you believe to be the most reliable, i.e. either the original model or the new one you defined in Q9 (If you cannot answer to Q9, then use the old model).

Q10. Set up and test a hypothesis test that the percentage of correct answer depends on the average work load per week. What would you conclude about my claim?

Q11. Model validation: check if the model assumptions are satisfied.

Exercise 3

Let $X_1, \ldots, X_n$ be independent and identically distributed random variables with $X_i \sim N(\mu, \sigma^2)$. Define

$$Y = \left(\frac{3}{n^2} \sum_{i=1}^{n} X_i \right) - \frac{2\mu}{n}.$$

Q12. Calculate $\mathbb{E}[Y]$ and $\mathrm{Var}(Y)$. What is the distribution of Y? (Note that Q13-15 can be solved even if you cannot answer Q12)

Q13. Set $n = 5$. Make a R-function which does the following:

- Simulate values of $X_1, \ldots, X_5$ from $X_i \sim N(\mu, \sigma^2)$ with $\mu = 10$ and $\sigma^2 = 5$.
- Compute the value of Y and make it as the output of the function.

Q14. Make a loop (Danish: løkke), where at each iteration you use the function defined in Q13. You should end up with 10000 simulated values of Y. Compare the mean and variance of these 10000 simulated values of Y with the true mean and variance of Y that you calculated in Q12. Comment on the result.

Q15. Make a histogram of the 10000 simulated values of Y obtained in Q14 and comment on the figure.