# M358K - Homework 2

posted on: September 26th, 2017

due: October 10th, 2017

## Instructions and grading scheme

**Submission instructions.** Please submit three files on Canvas with the following names:

- **homework2-writeup**: this is a word file containing your answers. All figures, tables etc must be inside this file! For how to save R figures, see instructions on Canvas.
- **homework2-code-final.R**: this is a text file containing the R codes you used to compute the numbers, tables, and figures CONTAINED IN your report ONLY. This file IS graded, see below.
- **homework2-code-draft.R**: this is a text file containing the R codes you used to explore the data. This file should give an idea on how you explored the dataset, and how you arrived at code-final.R. It can tables/figures which are EXCLUDED from your report, comments, etc. For instance, the R history of your various sessions is enough.

Acceptable word files are: anything that can be opened by LibreOffice, OpenOffice, Microsoft Word, or equivalent software. Example file extensions are .doc, .docx, .odt.

**Grading scheme.** For the write up: on each question you can earn 0/1/2 points.

- 2 = correct answer

- 1 = partially correct answer, or correct answer but with muddy or missing justification
- 0 = incorrect answer, unreasonable answer

For code-final.R: you can earn 0/5 points.

- 5 = code-final.R runs on the given dataset, gives all the tables and figures included in your report.
- 0 = no code file OR code file does not run at all OR code file does not produce the reports' tables/figures OR code file raises a plagiarism red flag, etc

For code-draft.R: you can earn 0/5 points.

- 5 = The grader is sufficiently convinced that you explored this dataset on your own.
- 0 = no code file OR code file raises a plagiarism red flag OR code file is irrelvant, does not give an indication on how you arrived at code-final.R

**Bonus points for presentation**:
Write-up: +2 points for nice report(grammatically correct sentences, no rambling discussions, discussions exceed expectations, extra analysis of the dataset)
R code: +2 points for neat layout. (code adequately commented, clearly laid out, easy to understand).

**Bonus questions**
These are extra challenging questions, and are additional opportunities to score points. On each bonus question you can get 0/4 points.

# Questions

To each of the 'is ... statistically sigifnicant' question below, set up a hypothesis test. Write down

- $H_0$ and $H_A$
- What test you use in R: name of test, one-sided or two-sided
- Your chosen significance level
- Report the $p$-value you obtained
- Give a conclusion

## Titanic dataset

Consider the Titanic dataset.

1. (Descriptive question) What are the survival rates of children? Of adult women? Of adults?
2. Do children have have significantly better survival rate than adults?
3. Are the survival rate of children significantly different from that of adult women?

## hsb2 dataset

Consider the hsb2 dataset. This dataset is included in the `openintro` package. You can load this dataset in R with the following commands:

```
install.packages("openintro")
library(openintro)
data(hsb2)
```

Now there will be a dataframe named `hsb2` in your R workspace. You can do usual commands, such as

```
head(hsb2)
```

For information on the variables, type in R

```
?hsb2
```

This dataset has been analyzed in class in the lecture on permutation tests. See the lecture R codes.

1. Describe in graphs and numbers the distribution of math scores between male and female

2. Is there a significant difference in the median score between these two groups? Use a permutation test to find out.

3. Is there a significant difference between male and female in the proportion of those who math score is 65 or more? Use a test of your choice.

4. Does your analysis disprove or support the claim that "top math students tend to be male"?

## murders dataset

The dataset `murders` contains the victim name, age, and location of every murder recorded in the Greater London area by the Metropolitan Police from January 1, 2006 to September 7, 2011.

This dataset is included in the `OIdata` package. You can load this dataset in R with the following commands:

```
install.packages("OIdata")
library(OIdata)
data(murders)
```

Now there will be a dataframe named `murders` in your R workspace. You can do usual commands, such as

```
head(murders)
```

For information on the variables, type in R

```
?murders
```

This documentation has helpful examples, including a code on how you can visualize the murder on the London map.

Our goal is to answer the question: do all boroughs have the same murder rate, or are there some "bad neighborhoods"?

1. Produce a map to visualize the murders by borough.

2. Produce a table that counts the number of murders by borough.

3. Is the count itself meaningful? What other statistics are we missing to compute the murder rate? Go online to find them and compute the murder rate by borough.

4. Is there a significant difference in the rate of murders between boroughs? (be clear on what test you are using).

Bonus question Answer the previous question by performing own permutation test. You will need to write a piece of R code to randomly reassign the location of the murders in the dataset. Be VERY clear on what test statistic you are using.

Bonus question Obtain a similar dataset for Austin, and perform the same analysis.