# Homework 2

*Alla Tambovtseva*

*Deadline: 20 December, 23:59*

## Submission

Part 1: answers should be submitted via Google forms:

https://goo.gl/forms/ltkXvUCvw39hsrqi1.

Part 2: the pdf-file with answers, code and graphs should be uploaded via Dropbox:

https://www.dropbox.com/request/nccLG4CQXUI2eLszx5Ch.

The pdf-file should not be necessarily created in RStudio, it can be a Word/Open Office file converted to pdf. You can create a text file, write anwers, copy your R code & outputs and add graphs (export pictures via *Export* in *Plots* tab in the right bottom corner in RStudio) and save as pdf.

An alternative way, more convenient, but more demanding: create an Rmd-file with texts and code chuncks, then knit it to Word via *Knit* button in RStudio and export to pdf. Recall our first lecture or look here and here.

## Part 1

Read pp. 28-42 from *OpenIntro Statistics* (you can find book on our course page) and answer the following questions. To submit your answers, fill in the Google form.

1. Formulate the difference between a simple mean and a weighted mean.
2. Provide definitions of unimodal, bimodal or trimodal distributions. Make up your own examples of variables that can have a) bimodal; b) trimodal distributions. Try to provide examples from the social sciences or at least realistic ones.
3. Suppose that the standard deviation of the variable $X$ is known. How to calculate the variance of $X$? What values the variance of a variable can take? In what cases the variance can be equal to 0?
4. List possible sources of outliers in data (why outliers can occur?). Note: you can think of your own examples, you do not have to base your answer only on the book.
5. Which estimates are called robust? Which of the statistics discussed (mean, median, quartiles, variance) are robust?
6. Why it might be necessary to transform data given?

### Part 2

Choose *one* of the variants listed below and do the tasks (all variants are of the equal level).

**Variant 1: Correlates of War**

1. Open the codebook for the data set *Non-state Wars*. Define the scales of the following variables (exclude values for coding missing data while making conclusions on a scale):

- WhereFought
- StartYear
- SideADeaths
- Initiator.

2. Upload a data set in RStudio. As NA's are coded in a special way (-8 and -9), use the following command:

```
wars <- read.csv("http://math-info.hse.ru/f/2018-19/pep/hw/Non-StateWarData_v4.0.csv",
                 na.strings = c(-8, -9))
```

   Provide the information on data structure: number of observations, number of variables, types of variables. Provide your R code for finding this information as well.

3. Does this data frame contain rows with missing values (NA)? If does, report the number of rows with missing values. Plot any graph for detecting patterns of missing values and comment on patterns.

4. Report the descriptive statistics for all variables in the data set. Provide your R code used for it. Choose one of the variables and interpret the statistics given, e.g. explain what the `Median` means, what are `1st Qu.` and `3rd Qu.`, and so on. Write in simple words (imagine that you should describe the results to your grandmother or any grandmother that does not have PhD in Statistics).

5. Choose the variable `WhereFought` and plot a bar chart for it. Change the color of the bar chart and add the title. Provide your R code used for it. Provide the interpretation for the bar chart, e.g. explain where most combats occured, state the region with the smallest number of wars, etc.

6. Choose the variable `SideBDeaths` and plot a histogram for it. Change the color and add the title. Provide your R code used for it. Provide the interpretation for the histogram: describe the shape of the distribution (whether it is skewed or not, how it is skewed, what values prevail, does it resemble normal distribution, etc).

7. Choose the variable `SideADeaths` and plot a box plot for it. Change the color and add the title. Provide your R code used for it. Are there outliers? If yes, calculate the number of outliers. It is not reliable to calculate them from the graph (it shows only unique values), you can get the vector of outliers using the following code:

```
boxplot(x)$out # replace x with your variable
```

8. Select rows that correspond to wars that started after 1901. Save them to the data frame `wars_20`. Provide your R code used for it.

9. Select rows that correspond to wars started after 1800 in which side A won (choose the appropriate value of `Outcome` consulting the codebook from above). Save them to the data frame `wars_A`. Provide your R code used for it.

10. Select columns `WarName`, `WarType`, and `WhereFought` and save them to the data frame `small_w`. Provide your R code used for it.

**Variant 2: State Failure Problem Set**

1. Open the codebook for the data set *Adverse Regime Change*. Define the scales of the following variables (exclude values for coding missing data while making conclusions on a scale):

- YEAR
- PTYPE
- MAGAVE
- SCODE.

2. Upload a data set in RStudio. As NA's are coded in a special way (9), use the following command:

```
reg <- read.csv("http://math-info.hse.ru/f/2018-19/pep/hw/ARChange2017.csv",
            na.strings = 9, dec = ",")
```

   Provide the information on data structure: number of observations, number of variables, types of variables. Provide your R code for finding this information as well.

3. Does this data frame contain rows with missing values (NA)? If does, report the number of rows with missing values. Plot any graph for detecting patterns of missing values and comment on patterns.

4. Report the descriptive statistics for all variables in the data set. Provide your R code used for it. Choose one of the variables and interpret the statistics given, e.g. explain what the `Median` means, what are `1st Qu.` and `3rd Qu.`, and so on. Write in simple words (imagine that you should describe the results to your grandmother or any grandmother that does not have PhD in Statistics).

5. Choose the variable `MOBEGIN` and plot a bar chart for it. Change the color of the bar chart and add the title. Provide your R code used for it. Provide the interpretation for the bar chart, e.g. explain in what months most cases occured, state the month with the smallest number of events, etc.

6. Choose the variable `MAGAVE` and plot a histogram for it. Change the color and add the title. Provide your R code used for it. Provide the interpretation for the histogram: describe the shape of the distribution (whether it is skewed or not, how it is skewed, what values prevail, does it resemble normal distribution, etc).

7. Choose the variable `MAGAVE` and plot a box plot for it. Change the color and add the title. Provide your R code used for it. Are there outliers? If yes, calculate the number of outliers. It is not reliable to calculate them from the graph (it shows only unique values), you can get the vector of outliers using the following code:

```
boxplot(x)$out # replace x with your variable
```

8. Select rows that correspond to events occured in Laos. Save them to the data frame `df_laos`. Provide your R code used for it.

9. Select rows that correspond to events occured after 1980 that have MAGFAIL no less than 2. Save them to the data frame `dat`. Provide your R code used for it.

10. Select columns `COUNTRY`, `YEAR`, and `MAGAVE` and save them to the data frame `small`. Provide your R code used for it.

**Variant 3: Minorities at Risk**

1. Open the codebook for the data set *Minorities at Risk*. Define the scales of the following variables (exclude values for coding missing data while making conclusions on a scale):

- GPOP
- LANG
- GC119
- SEPX

2. Upload a data set in RStudio. As NA's are coded in a special way (-99), use the following command:

```
mar <- read.csv("http://math-info.hse.ru/f/2018-19/pep/hw/mar.csv", na.strings = -99)
```

   Provide the information on data structure: number of observations, number of variables, types of variables. Provide your R code for finding this information as well.

3. Does this data frame contain rows with missing values (NA)? If does, report the number of rows with missing values. Plot any graph for detecting patterns of missing values and comment on patterns.

4. Report the descriptive statistics for all variables in the data set. Provide your R code used for it. Choose one of the variables and interpret the statistics given, e.g. explain what the `Median` means, what are `1st Qu.` and `3rd Qu.`, and so on. Write in simple words (imagine that you should describe the results to your grandmother or any grandmother that does not have PhD in Statistics).

5. Choose the variable `VMAR_Region` and plot a bar chart for it. Change the color of the bar chart and add the title. Provide your R code used for it. Provide the interpretation for the bar chart, e.g. explain in what regions most groups are concentrated, state the region with the smallest number of cases, etc.

6. Choose the variable `GPRO` and plot a histogram for it. Change the color and add the title. Provide your R code used for it. Provide the interpretation for the histogram: describe the shape of the distribution (whether it is skewed or not, how it is skewed, what values prevail, does it resemble normal distribution, etc).

7. Choose the variable `GPRO` and plot a box plot for it. Change the color and add the title. Provide your R code used for it. Are there outliers? If yes, calculate the number of outliers. It is not reliable to calculate them from the graph (it shows only unique values), you can get the vector of outliers using the following code:

```
boxplot(x)$out # replace x with your variable
```

8. Select rows that correspond to groups in Eastern Europe and the former Soviet Union (choose the appropriate region consulting the codebook mentioned above). Save them to the data frame `eu_sov`. Provide your R code used for it.

9. Select rows that correspond to groups that speak multiple languages, at least one different from plurality group (see `LANG` variable) and that are not politically discriminated (see `POLDISC`). Save them to the data frame `pdisc`. Provide your R code used for it.

10. Select columns `Group`, `Year`, and `LANG` and save them to the data frame `mar_small`. Provide your R code used for it.

**Variant 4: Comparative Political Data Set**

1. Open the codebook for the data set *Comparative Political Data Set*. Define the scales of the following variables (exclude values for coding missing data while making conclusions on a scale):

- eu
- gov_right1
- gov_type
- rae_ele.

2. Upload a data set in RStudio.

```
df <- read.csv("http://math-info.hse.ru/f/2018-19/pep/hw/CPDS.csv", dec = ",")
```

   Provide the information on data structure: number of observations, number of variables, types of variables. Provide your R code for finding this information as well.

3. Does this data frame contain rows with missing values (NA)? If does, report the number of rows with missing values. Plot any graph for detecting patterns of missing values and comment on patterns.

4. Report the descriptive statistics for all variables in the data set. Provide your R code used for it. Choose one of the variables and interpret the statistics given, e.g. explain what the `Median` means, what are `1st Qu.` and `3rd Qu.`, and so on. Write in simple words (imagine that you should describe the results to your grandmother or any grandmother that does not have PhD in Statistics).

5. Choose the variable `poco` and plot a bar chart for it. Change the color of the bar chart and add the title. Provide your R code used for it. Provide the interpretation for the bar chart, e.g. explain which type prevails, etc.

6. Choose the variable `dis_gall` and plot a histogram for it. Change the color and add the title. Provide your R code used for it. Provide the interpretation for the histogram: describe the shape of the distribution (whether it is skewed or not, how it is skewed, what values prevail, does it resemble normal distribution, etc).

7. Choose the variable `womenpar` and plot a box plot for it. Change the color and add the title. Provide your R code used for it. Are there outliers? If yes, calculate the number of outliers. It is not reliable to calculate them from the graph (it shows only unique values), you can get the vector of outliers using the following code:

```
boxplot(x)$out # replace x with your variable
```

8. Select rows that correspond to countries with the voter turnout (`vturn`) in elections greater than 60%. Save them to the data frame `elect`. Provide your R code used for it.

9. Select rows that correspond to EU members with the index of electoral fractionalization less than 0.3. Save them to the data frame `eu03`. Provide your R code used for it.

10. Select columns `year`, `country`, and `gov_party` and save them to the data frame `df_small`. Provide your R code used for it.

**Variant 5: Press Freedom**

1. Open the methodology for the data set *Press Freedom*. Define the scales of the following variables (exclude values for coding missing data while making conclusions on a scale):

- LEGAL ENVIRONMENT
- POLITICAL ENVIRONMENT
- PRESS FREEDOM SCORE
- STATUS (Free, Partly Free, Not Free)

2. Upload a data set (2017) in RStudio.

```
mf <- read.csv("http://math-info.hse.ru/f/2018-19/pep/hw/Mfree.csv",
               na.strings = "-")
```

Provide the information on data structure: number of observations, number of variables, types of variables. Provide your R code for finding this information as well.

3. Does this data frame contain rows with missing values (NA)? If does, report the number of rows with missing values. Plot any graph for detecting patterns of missing values and comment on patterns.

4. Report the descriptive statistics for all variables in the data set. Provide your R code used for it. Choose one of the variables and interpret the statistics given, e.g. explain what the `Median` means, what are `1st Qu.` and `3rd Qu.`, and so on. Write in simple words (imagine that you should describe the results to your grandmother or any grandmother that does not have PhD in Statistics).

5. Choose the variable `Status` and plot a bar chart for it. Change the color of the bar chart and add the title. Provide your R code used for it. Provide the interpretation for the bar chart, e.g. explain which type prevails, etc.

6. Choose the variable `TotalScore` and plot a histogram for it. Change the color and add the title. Provide your R code used for it. Provide the interpretation for the histogram: describe the shape of the distribution (whether it is skewed or not, how it is skewed, what values prevail, does it resemble normal distribution, etc).

7. Choose the variable `Economic` and plot a box plot for it. Change the color and add the title. Provide your R code used for it. Are there outliers? If yes, calculate the number of outliers. It is not reliable to calculate them from the graph (it shows only unique values), you can get the vector of outliers using the following code:

```
boxplot(x)$out # replace x with your variable
```

8. Select rows that correspond to Partly Free countries and save them to the data frame `pfree`. Provide your R code used for it.

9. Select rows that correspond to Free countries with Total Score no less than 20. Save them to the data frame `free20`. Provide your R code used for it.

10. Select columns `Country`, `Status`, and `TotalScore` and save them to the data frame `df_small`. Provide your R code used for it.