Political Analysis: Lab 2
Hilary Term 2015

Descriptive statistics

In this lab session we will see how we can explore our data using descriptive statistics. The lab is clearly divided in two parts. In the first one, you will be getting to know the `R` language and its commands. In the second part, there will be several questions and you will use real data to answer them. You will need to use the knowledge acquired in the first part to work on the second. The exercises could take more time that you have in the lab, so we suggest to finish them at home if needed. To get started clean your `R` working space:

```
rm(x)          # if you want to remove object x
rm(list=ls()) # if you want to remove all objects
```

Playing with fake data

Let's start by creating a variable that we will call `gdp_pc`:

```
gdp_pc <- c(12000, 25000, 5000, 1500, 16000, 15000, 30000, 12000, 40000,
            10000, 56000, 2200)
```

A first look at the distribution

Explore your data with the following command:

```
summary(gdp_pc)
```

Questions:

- Is the mean larger than the median?
- What are the minimum and maximum values?
- What do `1st Qu` and `3rd Qu` stay for?

Just visually look at the data: what does the following histogram show? (If you need to learn more about the histogram function use the command `?hist`.)

```
hist(gdp_pc)
hist(gdp_pc, main="GDP per capita") # add title
```

Mean

Let's go back to the sample mean formula:

$$\text{mean of x} = \bar{x} = \frac{1}{n}(x_1 + x_2 + ... + x_n) = \frac{1}{n}\sum_{i=1}^{n} x_i$$

You can calculate the sample mean of `gdp_pc` by using the following components:

```
# what is N or length of the vector?
length(gdp_pc)

# what is the "sigma" or the sum of the values?
sum(gdp_pc)

# go through the formula and you can compute the sample mean using
# the N and the sum
(1 / length(gdp_pc)) * sum(gdp_pc)

# compare the result with the R function
mean(gdp_pc)
```

Now try to generate a dummy variable (0 or 1) for cases that are above or below the mean `gdp_pc`. Part of this function could be useful later:

```
# option 1:
dummy <- gdp_pc > mean(gdp_pc)

# option 2
dummy <- ifelse(gdp_pc > mean(gdp_pc), 1, 0)

# checking the result
cbind(gdp_pc, mean(gdp_pc), dummy)
```

Missing values

Set the last observation in `gdp_pc` as a missing value and work through the following lines:

```
gdp_na <- c(12000, 25000, 5000, 1500, 16000, 15000, 30000, 12000, 40000,
            10000, 56000, NA)

# describe data, notice the new NA
```

```r
summary(gdp_na)

# get the sum now
sum(gdp_na)

# it doesn't work, check why
?sum

# you need to add the argument "na.rm=TRUE")
sum(gdp_na, na.rm=TRUE)

# add "na.rm=TRUE" also when calculating the mean
mean(gdp_na, na.rm=TRUE)

# median
median(gdp_na, na.rm=TRUE)

# quantiles
quantile(gdp_na, na.rm=TRUE)

# and percentiles
quantile(gdp_na, c(.10, .50, .90, .99), na.rm=TRUE)
```

Variance and standard deviation

Let's move on to variance and standard deviation:

$$\text{variance of x} = s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

```r
# variance by hand from formula
(1 / (length(gdp_pc) - 1))*sum((gdp_pc - mean(gdp_pc))^2)

# with R function
var(gdp_pc)
```

Standard deviation (s.d.) formula:

$$\text{s.d. of x} = \sqrt{s^2} = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

```r
# standard deviation by hand from formula
sqrt((1 / (length(gdp_pc) - 1)) * sum((gdp_pc - mean(gdp_pc))^2))

# with R function
sd(gdp_pc)
```

Boxplot

Summarising data in one graph, using the boxplot:

```r
boxplot(gdp_pc)
```

Questions:

- What is the tick black line?
- What does the box represent?
- What are the two whiskers?
- And the point(s)?

Mean and median

Some further analysis: some thoughts on the relation between mean and median values.

```r
# create another vector on gdp per capita
# (it could be  a different year)
gdp2_pc <- c(11000, 5000, 5000, 1500, 16000, 15000, 30000, 8000,
             52000, 7000, 60000, 900)

# plot the two different years
hist(gdp_pc)
hist(gdp2_pc)

# what do we get from the ratios of mean and median?
# In what case is this ratio 1?
mean(gdp_pc) / median(gdp_pc)
mean(gdp2_pc) / median(gdp2_pc)

# which standard deviation would you expect to be higher?
sd(gdp_pc)
sd(gdp2_pc)
```

Working with Real Data

Now we load our data:

```
Gov <- read.csv("http://andy.egge.rs/data/L.csv")
```

In the previous part of the lab you have learnt how to get descriptive statistics and explore data, now we will start using unemployment data between 1981 and 2009. First, let's get an histogram and discuss the data:

```
hist(Gov$unemployment_1981_2009)
summary(Gov$unemployment_1981_2009)
```

Discuss the unemployment variable: for instance, are there missing values? Maximum and mean? Which countries have the highest umemployment rates? Which the lowest?

```
# calculate the mean
mean(Gov$unemployment_1981_2009)
mean(Gov$unemployment_1981_2009, na.rm=TRUE)

# show the columns "country" and "unemployment_1981_2009"
Gov[,c("country", "unemployment_1981_2009")]

# show the same thing, but sorted by unemployment rate:
Gov[order(Gov$unemployment_1981_2009), c("country", "unemployment_1981_2009")]
```

Investigating Economic Inequality

We will use for this part of the lab a proxy for inequality that is called GINI index. It is an index that goes from 0 to 100, where 0 is a country were the wealth is equally owned by all citizens, whereas 100 is a country where only one citizen owns all the national wealth. Start having a look using a couple of graphs (*hint*: `histogram` and `boxplot`):

- Plot a histogram and add one line each where you think the sample is divided in two and where the mean is. (*Hint*: check the `abline` function.)
- What is the average level of inequality in your sample in 2000?
- What country is the most equal and which one is the most unequal?
- Nature decided to put you in a country of the top 10th percentile for fairness (measured as Gini Index)? In which countries you could be?

- Then, since you got a job with a NGO, you have to move to the lowest quantile for fairness (again GINI Index). In which countries could we find you?
- What if you were in the country that split the sample 50/50, would you be in a county that is more unequal than average?
- Are there any missing cases? What countries do not have statistics for inequality in 2000?
- Which percentile does your country belong to? What other countries belong to nearby percentiles?

Investigating Women Representation in Parliament

In this section we compare data on women representation in parliaments for two different years: 1990 and 2010. Try to answer the following questions:

- On average what is the percentage of women in parliament in 1990? And what about 2010?
- Compare the two means. What about the median values? What do they indicate?
- Compare their standard deviations.
- What do the two different boxplots show for 1990 and 2010?
- What are the best 5 performers and worst 5 performers in Women Parliamentary Representation in 1990? And 2010?
- What countries perform higher than the mean in 1990? And in 2010?
- In 1990, 80% of the countries had at least what percentage of women in their parliament?
- In 2010, 80% of the countries had at least what percentage of women in their parliament? To what percentile is this comparable of 1990?
- Plot histograms for women representation for the years 1990 and 2010 comparing the distribution of this variable if a country is above or below the mean of corruption. (*Hint*: Do 4 plots using the logic of creating a dummy variable.)