Political Analysis: Lab 3
Hilary Term 2015

1 Comparing distributions

In this lab session we will see how we can assess the strength of association between two continuous or categorical variables. We will start by comparing the corruption perception index of old and new democracies. We adopt a rough-and-ready definition of old democracies as those countries with uninterrupted democratic rule since 1945 (the end of WWII). Thus, we need to generate a new variable which will take the value 1 for *New democracies* and the value of 0 for *Old democracies*.

```
L <- read.csv("http://andy.egge.rs/data/L.csv")
L$newdem = L$country %in% c('ARG', 'BAH', 'BAR', 'BOT', 'CR', 'GRE',
                            'IND', 'JAM','KOR', 'MAL', 'MAU', 'POR',
                            'SPA','TRI','URU')
```

Let's start by looking at the distribution of `corruption_perception_index_2010`. We will start by looking at all countries in the dataset. The code for the density plot was introduced in the previous lab session and is given also below:

```
density.corruption <- density(L$corruption_perception_index_2010,
                              na.rm=TRUE)
plot(density.corruption)
```

Since we want to compare old and new democracies, we need to replicate the graph looking separately at each set of countries:

```
density.corruption.old <- density(L$corruption_perception_index_2010[L$newdem==F],
                                  na.rm=TRUE)
density.corruption.new <- density(L$corruption_perception_index_2010[L$newdem==T],
                                  na.rm=TRUE)
plot(density.corruption.old)
plot(density.corruption.new)
```

You can easily combine them by using:

```
plot(density.corruption.old, col= "blue", xlab="")
lines(density.corruption.new, col="red")
```

2 Correlation

Let's focus now on the possible roots of corruption. One factor that might increase corruption is the presence of interest groups. Interest groups exercise pressure to governments trying to extract rent. We thus expect that the more interest groups there are in a given country, the higher the level of perceived corruption will be. Let's examine if this is the actually the case.

We start by visualizing the relationship between interest groups and corruption. We use the perception index for the corruption variable and the index of interest group pluaralism as measured in 2010 (`index_of_interest_group_pluralism_1945_2010`). We use a scatterplot to explore visually this relationship:

```
plot(L$index_of_interest_group_pluralism_1945_2010,
     L$corruption_perception_index_2010)
```

or by using the `scatterplot` function from the `car` library, which we need to load first:

```
library(car)

## Loading required package:  MASS
## Loading required package:  nnet

scatterplot(L$index_of_interest_group_pluralism_1945_2010,
            L$corruption_perception_index_2010,
            smoother=F, box=F, grid=F, lwd=3, col=1, reg.line=F,
            xlab="Interest Group Pluralism",
            ylab="Corruption")
```

Questions:

- Can you discern any pattern between the two variables?
- How would you like to summarise this relationship?
- Would fitting a line help?

Let's try to fit a line into this scatterplot, as a way to summarise the overall pattern in the relationship between the two variables. This line will only help if it actually provides a reliable summary of the actual relationship between corruption and interest group pluralism:

```
scatterplot(L$index_of_interest_group_pluralism_1945_2010,
            L$corruption_perception_index_2010,
            smoother=F, box=F, grid=F, lwd=3, col=1,
            id.n = length(L$country), label=L$country,
            xlab="Interest Group Pluralism", ylab="Corruption")
```

Questions:

- Does the line help? Do you think it serves to disentangle the main trend between the two variables?
- Do you find exceptions, i.e. cases that do not adhere to the pattern denoted by the line? What does that tell you about the relationship between the two variables?

Although visualisation is very important in helping us gauge the pattern of association between two variables, we still want to use a way to summarise the strength of this association. We can use the Pearson Correlation coefficient for this purpose. Remember that the Pearson correlation coefficient can only give us some idea about the strength of the linear association between these variables. It cannot capture non-linear patterns and it cannot disentangle whether this association also implies causation. Here is the code:[1]

```
cor.test(L$corruption_perception_index_2010,
         L$index_of_interest_group_pluralism_1945_2010)
```

Questions:

- What is the correlation coefficient estimate? What does it tell you about the strength of the association between these two variables?
- Is this association statistically significant? How would you know?

Optional Exercise for Scatterplots

An interesting extension would be to examine whether the correlation is different for new and old democracies. Let's return to the graph. We have seen that new democracies differ from old democracies in their perceived levels of corruption. We have also seen that the number of interest groups is negatively associated with corruption. But does this hold for both new and old democracies? Are interest groups equally prone to generate corruption in both old and new democracies? One would expect that in old democracies there is an institutional infrastructure that would make more difficult for interest groups to extract rent. If this is the case we should see that the correlation between the number of interest groups and corruption is lower in old compared to new democracies. Let's see if this is the case:

```
scatterplot(L$corruption_perception_index_2010 ~
            L$index_of_interest_group_pluralism_1945_2010 | L$newdem,
            smoother=F, box=F, grid=F, lwd=3,
            xlab="Interest Group Pluralism", ylab="Corruption")
```

Questions:

[1] The command `cor` only gives you the correlation coefficient, but it could also be used as an alternative.

- Do your conclusions differ between new and established democracies?
- Looking at established democracies, do you see any country that does not adhere to the general trend? Can you guess which country that is?

Finally, we have thus far assumed that a linear fit is suitable to summarize the relationship between interest groups and corruption. It's now time to relax this assumption. Let's try to be more flexible, allowing nonlinearities to emerge in the graph. We do that by fitting a loess curve, which traces the local mean of corruption across all levels of interest group pluralism. The added value from this exercise is that you can spot gross deviations from linearity, which would render the linear approximation a poor fit of the data:

```
scatterplot(L$index_of_interest_group_pluralism_1945_2010,
            L$corruption_perception_index_2010,
            smoother=loessLine,  box=F, grid=F, lwd=3, col=1, reg.line=F,
            id.n=length(L$country), label=L$country,
            xlab="Interest Group Pluralism", ylab="Corruption")
```

Questions:

- Describe the pattern summarised by the `loess` curve. Do your conclusions differ from the ones based on the linear fit used above?
- Does the `loess` line challenge the linearity assumption that we have been making thus far?

Contingency Tables

Although scatterplots are useful to understand relationships between two continuous variables, they are not helpful when analysing categorical data. Let's start by creating a categorical variable. The `car` library has a very good function for recoding the data. The following code creates a binary variable (`dummycorruption`) for corruption with 1 denoting high corruption and 0 low corruption. We have used the median value as a way to split corruption into two groups:

```
dummycorruption <- recode(L$corruption_perception_index_2010,
                          "0:7.1=0; 7.1:10=1")
```

Do the same for interest groups and create a `dummyinterest` with 1 denoting high number and 0 denoting low numbers in interest group plurality. How would you know if the relationship depicted above holds? A contingency table can be of help. The following code creates a 2×2 table and examines the cell proportions.

```r
corruption.interest <- table(dummyinterest, dummycorruption)
corruption.interest
prop.table(corruption.interest)
```

Is the association statistically significant? The test to determine the significance of the association is called the Test of Independence. The key idea behind independence tests is Observed and Expected Values. Whereas Observed values are more or less self-explanatory, the Expected Values relate to the scores we would get under the Null hypothesis. The expected value for each cell in a two-way table equals

$$\frac{\text{Row Total} \times \text{Column Total}}{n},$$

where n is the total number of observations included in the table. To test for independence we examine the χ^2 statistic which is given by the following formula:

$$\chi^2 = \sum_{i=1}^{n} \frac{(\text{Observed}_i - \text{Expected}_i)^2}{\text{Expected}_i}$$

where i indexes the cells in the table. Thus, χ^2 is the sum of the difference between expected (under the null of no association) and observed frequencies in each cell. A χ^2 statistic helps determine if the associations are due to pure luck or if there is a systematic pattern in our sample. The test can be ran using the following code:

```r
chisq.test(corruption.interest)
```

Is the association between the two categorical variables statistically significant?

Optional Exercise for Contingency Tables

By using the above lines of code, determine the relationship between new and old democracies and corruption or/and interest groups. When you are done try to plot the contingency table using the `mosaic` function of the `vcd` library.

Following the logic of the code you have learned today it is relatively straightforward to change the labels and add titles.