Elements of Statistics and Econometrics

Pre-reading: Descriptive Statistics and Probability Theory

1. Please, read carefully the slides on Chapter 1: Descriptive Statistics and Chapter 2: Basics of Probability Theory.
2. I will quickly go through some parts of these two chapters during the first set of lectures, but we will not have time to discuss them in detail.
3. For the rest of the course, those who have a good background in Math and Statistics I would recommend to have a look into the relevant chapters of a very concise book by Larry Wassermann *All of Statistics*.
4. For those who have a less quantitative background please check the relevant topics in the book by Grinstead and Snell *Introduction to Probability*. For the Econometrics part I will give the relevant list of books later.
5. You are free to choose any software you prefer (R, Python, Matlab, ...). I am sure that most of you are experts in programming, so using correct tools and drawing correct conclusions from the results is more important in this course than the implementation itself. Thus the solutions containing only code will not be taken into account!
6. Feel free to contact me per Email yarema.okhrin@wiwi.uni-augsburg.de or by Skype, if you have any questions or problem.
7. I look forward to our course!

Problem 1: Descriptive Statistics and Probability Theory: Real Data on CEO Compensation

1. In the pre-reading you/we have discussed tools and methods for visualizing data and computing some simple characteristic measures. Our aim here is to apply all the basic techniques and to draw correct conclusions. The file `ceo.xls` contains data on the CEO compensations and some additional variables listed below.

```
salary  = 1999 salary + bonuses in 1000 US$
totcomp = 1999 CEO total compensation
tenure  = # of years as CEO (=0 if less than 6 months)
age     = age of CEO
sales   = total 1998 sales revenue of firm i
profits = 1998 profits for firm i
assets  = total assets of firm i in 1998
```

Our aim is to evaluate the data set with basic tools.

(a) For the variable `totcomp` compute the common location measures: mean, 5%-trimmed mean, median, upper and lower quartiles, the upper and lower 5%-quantiles. Give an economic interpretation for every location measure.

(b) Plot the empirical cumulative distribution function. Compute and explain in economic terms the following quantities

i. $\hat{F}^{-1}(0.1)$ and $\hat{F}^{-1}(0.9)$

ii. $\hat{F}(2000)$ and $1 - \hat{F}(4000)$

(c) Plot the histogram of `totcomp` and the Box-plot (or `violin`-plot). What can be concluded about the distribution of the data? Are the location measures computed above still appropriate? Compute and discuss an appropriate measure of symmetry.

(d) Check which method is used in your software to compute the optimal bandwidth (or the number of bars) in the histogram. Describe it shortly here. Make plots of too detailed and too rough histograms. What can we learn from these figures?

(e) There are methods which help us make the distribution of the sample more symmetric. Consider the natural logarithm of the total compensation: ln(totcomp). Plot the histogram (and Box-plot) and compare it with the figures for the original data. Compute the mean and the median. What can be concluded from the new values? Provide economic interpretation.

2. Next we try to make a more detailed analysis of the data (without logarithm).

(a) We suspect that the total compensation of the CEO and other variables are related. Compute the correlation coefficients of Pearson and plot them as a heatmap (correlation map). Discuss the strength of the correlations.

(b) Plot the scatter plots (`pairs` in R). Conclude if the linear correlation coefficients are appropriate here. Compute now the Spearman's correlations and make a heatmap. Compare the results with Pearson. What is the rank of the observation totcomp= 6000?

(c) Consider the two subsamples: CEOs younger than 50 and older than 50. Plot for both subsamples overlapping histograms/ecdf's and discuss the results. What can we learn from the corresponding location and dispersion (!) measures?

3. Consider another grouping of the data. Define the groups:

$$\begin{cases} S_1 & \text{if salary } < 3000 \\ S_2 & \text{if salary } \geq 3000 \text{ but } < 5000 \\ S_3 & \text{if salary } \geq 5000 \end{cases} \qquad \begin{cases} A_1 & \text{if age } < 50 \\ A_2 & \text{if age } \geq 50 \end{cases}$$

(a) Aggregate the data to a 2×3 contigency table with absolute and with relative frequencies.

(b) Give interpretation for the values of n_{12}, h_{12}, $n_{1\cdot}$ and $h_{1\cdot}$.

(c) Compute an appropriate dependence measure for S_i and A_j. What can be concluded from its value?

Problem 2: Descriptive Statistics and Probability Theory: Simulated Data

It is obvious that the descriptive tools are very sensitive to contamination or outliers in the data. The objective of this problem is to assess the sensitivity of these measures/tools to outliers or very heterogenous data.

1. Simulate (with a fixed seed) a sample of size $n = 100$ from the normal distribution with $\mu_1 = 10$ and $\sigma_1^2 = 9$.

   (a) Plot the histogram and compare it to the density of $N(10, 9)$.

   (b) Now draw a sample y_i of size $n = 100$ from t_5. Transform it as follows: $10 + 3\sqrt{3/5} \cdot y_i$. Plot the histogram and compare the density of $N(10, 9)$. What can be concluded and why this example might be relevant for empirical studies?

2. In practice the data is always very heterogenous. To reflect it we contaminate the data by adding an outlier or a subsample with different characteristics.

   (a) To obtain a realistic heterogenous sample add to the original normal data a new sample of size m simulated from $N(20, 2^2)$, i.e. $\mu_2 = 20$ and $\sigma_2^2 = 4$. The size m will obviously influence the above measures. Vary m from 10 to 200. (The resulting sample is said to stem from a mixture normal distribution).

   (b) Plot Box-plots (or `violin` plots) and histograms for each subsample individually and for the sample for a few different values of m.

   (c) Make animated or interactive graphics (with `manipulate, plotly, ggplot`, etc.) to visualize the impact of m on the histogram and location measures (added as vertical lines in the graph) of the data.

3. Next step is to simulate two dependent data sets. We simulate two samples with a given value of the correlation coefficient.

   (a) Let $U \sim N(0, 1)$ and $V \sim N(0, 1)$. Let $U^* = U$ and $V^* = \rho U + \sqrt{1 - \rho^2} V$. Prove that $Corr(U^*, V^*) = \rho$ and the variances of both variables U^* and V^* equal one.

   (b) Use the above idea to simulate two samples of size $n = 100$ from a normal distribution with different values of ρ. Compute the correlation coefficients of Pearson and of Spearman. Compare the correlation to the original parameter ρ (for example, plot Pearson vs. ρ and Spearman vs. ρ).

   (c) Make a nonlinear but monotone transformation of V^*, say exp for simplicity. Check the impact of this transformation on the correlation coefficients of Spearman and Pearson. Think about an appropriate visualization of the findings.