Elements of Statistics and Econometrics - 2025/26

Descriptive Statistics and Probability Theory

Problem 1: Descriptive Statistics and Probability Theory: Real Data

1. In this assignment we will deal with 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 (at least a sentence for every 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) using the default option in R/Python. What is the interpretation of the area of a rectangle in the histogram? Link the parts of the Box-Plot to the location measures above.

   (d) 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.

   (e) 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?

(f) 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 to each other. 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 the results for Pearson.

   (c) What is the rank of the observation totcomp= 6737? Explain in your own words the conceptual difference between the two correlation coefficient and link it to linear/monotone dependence.

   (d) 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? What can we infer about the co-movement of the variables or their changes in opposite directions?

Problem 2: Descriptive Statistics and Probability Theory: Simulated Data

1. 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) Draw a sample of size 100 from $N(10, 1^2)$. To obtain a realistic heterogenous sample add to this 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 properties of the whole data set. Vary m from 10 to 200 (say, by step 20). (The resulting sample is said to stem from a mixture normal distribution and arises in practice if we put together observations from two homogenous cohorts). Calculate mean, median, and sample variance as a function of m. Discuss the impact of heterogeneity on the these measures.

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

2. 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 $V \sim N(0,1)$ be independent. 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.

Problem 3: Descriptive Statistics and Probability Theory: Probabilities and Distributions

1. All students belong to one of two groups: well prepared for the exam (group A) and not so well prepared for the exam (group B). A "well prepared" student passes the exam with probability of 80% and a "not so well prepared"-student with probability of 30%. We assume that only 10% of all students belong to group B.

   A : a randomly chosen student belongs to group A
   B : a randomly chosen student belongs to group B
   E : a randomly chosen student passes the exam

   (a) What is the probability that a randomly chosen student passes the exam AND belongs to group A.

   (b) What is the probability that a randomly chosen student passes the exam?

   (c) What is the probability that a student who passed the exam, belongs to group B?

2. The dean and the head of the program invite all students to a barbecue party at the end of the year. Five vegetarians are among the guests. The pitmaster is an expert in juicy steaks and still needs some practice with grilled vegetables. For this reason a portion of vegetables is burned to ashes with probability of 60%.

   (a) What is the probability that non of the five vegetable servings is charred?

   (b) What is the probability that exactly two of the five vegetable servings are charred?

   (c) What is the probability that at least three out of five serving are charred?

3. Let X be the weekly return of stock A. We assume that it follows a normal distribution with $\mu = 0.03$ (3%) and $\sigma^2 = 0.02$. We invest 40% of our capital in 10 such stocks and the remaining 60% in a risk-free asset with return r_f=0.01.

   (a) Calculate the probability that the return of a single stock is

      i. is larger then 0.01;
      ii. is larger or equal 0.01;
      iii. is less then 0.01;
      iv. lies between 0.01 and 0.03.

   (b) Determine the distribution of the portfolio return by stating the distribution and computing its parameters. (Use here the properties of the normal distribution, e.g. linear transformation)

   (c) How a correlation among the assets can influence the variance of the investment? Which formula supports your idea?

4. Download (for example, from Yahoo Finance) the weekly prices and calculate the returns of Tesla Inc for the last two years (use simple returns $(P_t - P_{t-1})/P_{t-1}$ or continuous returns $\ln(P_t/P_{t-1})$). You are interested in the probability that the Tesla return falls a) below -2%; b) below -4%.

   (a) Calculate this probability using the tools and methods of the descriptive statistics (quantiles, empirical CDF, etc.).

   (b) Calculate this probability assuming that the returns follow a normal distribution with the parameters given by the sample mean and sample variance.

   (c) Compare the probabilities. Discuss pros and cons of both methods.

5. Assume that the price of a given asset can either decrease of increase in every period of time. This setting is called a binomial model. The probability of an upward movement equals 55% and the starting price $S(0) = 10$. The upward and downward returns equal 0.02 and -0.01 respectively. Consider holding the asset for several periods.

   (a) Calculate the probability that the stock price increases

      i. ... five periods in a row.

      ii. ... three periods in a row.

      iii. ... in three out of five periods.

      iv. ... in 10 out of 15 periods.

   (b) Calculate the expected return of this investment after two periods.

6. Compute the following probabilities for the distributions $N(0,1)$, t_2, $Exp(1)$, χ^2_4:

   (a) $P(1 < X < 1.5)$

   (b) $P(1 \leq X < 1.5)$

   (c) $P(1 < X)$

   (d) $P(X < 1.5)$

   (e) $P(X = 1.5)$