Elements of Statistics and Econometrics - 2021

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 data we work with is the Ames Housing data set available for download at Kaggle:

   https://www.kaggle.com/c/house-prices-advanced-regression-techniques/overview

   or included into the `AmesHousing` package in R. The full data set contains almost 3000 observations on 81 features.

   (a) For the variable `SalePrice` 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. $\hat{F}^{-1}(0.2)$ and $\hat{F}^{-1}(0.8)$

   ii. $\hat{F}(100000)$ and $1 - \hat{F}(50000)$

   (c) Plot the histogram of `SalePrice` 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) Check the computational details on the plotted Box-Plot as implemented in your software. What is defined as an outlier? Where the smoothing comes from if you plot the violin plot?

   (f) There are methods which help us make the distribution of a sample more symmetric. Consider the natural logarithm of the sale price: ln(`SalePrice`). 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? How would you interpret now the location measures from the economic perspective?

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

   (a) We suspect that the sale price and other variables are related. Compute its correlation coefficients of Pearson with several interval-scaled features (`LotArea, YearBuilt, TotalBsmtSF, GrLivArea, GarageArea`) and plot them as a heatmap (correlation map). Discuss the strength of the correlations.

   (b) Plot the corresponding scatter plots. 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 `SalePrice`= 286000?

   (c) Consider the correlation and the scatter plot between `SalePrice` and `SecondFlrSF`. Why it might be problematic to take these variables just as they are? What would you suggest to overcome this problem?

   (d) Consider the three subsamples: houses with `YearBuilt` before 1970, between 1970 and 1990, younger than 1990. Plot for the three subsamples overlapping histograms/ecdf's of the `SalePrice` and discuss the results. What can we learn from the corresponding location and dispersion measures?

3. Consider another grouping of the data. The groups are build according to the following splitting criteria: `SalePrice` below 150K, between 150K and 250K and above 250K. The second variable that we consider is `HouseStyle`.

   (a) Aggregate the data to a 3×8 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 the grouped sale price and the style of the house. What can be concluded from its value?

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) 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.

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 $V \sim N(0,1)$. 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. 45% of the students in a particular program are male. 30% of all students are vegetarian. The probability, that a female student is vegetarian is 50%.

   M : a randomly chosen student is a man
   W : a randomly chosen student is a woman
   V : a randomly chosen student is vegetarian

   (a) What is the probability that a randomly chosen student is a female AND is vegetarian?

   (b) What is the probability that a randomly chosen student is a male AND is vegetarian?

   (c) What is the probability that a randomly chosen student is a male, if we know that he is a vegetarian.

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 $X \sim N(5,9)$. Compute the following probabilities:

   (a) $P(10 < X < 15)$

   (b) $P(10 \leq X < 15)$

   (c) $P(10 < X)$

   (d) $P(X < 15)$

   (e) $P(X = 15)$