MA308: Statistical Calculation and Software

Assignment 1 (Sep 06– Oct 08, 2021)

1.1 [12 points] Obtain the dataset “iris” in R , then

(a) [4 points] Turn the column variable “Species” into a factor, name the three levels as “a”, “b”, “c” respectively, obtain the total number of records for each type of irises.

(b) [4 points] Obtain the boxplot of “Sepal.Length” versus “Species”, i.e. three boxplots of “Sepal.Length” for all types of “Species”, make the boxplot color “darkgreen” with title “Sepal length distribution for each type of irises”.

(c) [4 points] Select all the irises with sepal length larger than 5.5, obtain the boxplot of “Petal.Length” and “Petal.Width” versus “Species”, i.e. for each type of “Species”, two boxplots for “Petal.Length” and “Petal.Width” stay side by side, and only include records for irises with sepal length larger than 5.5.

1.2 [32 points] Obtain the dataset “F2000” by importing “F2000.csv” or “F2000.txt”,

(a) [4 points] Calculate the skewness of marketvalue of all the 2000 companies.

(b) [4 points] Try to remove the skewness of marketvalue by taking log, inverse and Box & Cox transformation with index $\lambda = 0.25$.

(c) [5 points] Make histograms for marketvalue and all the transformed marketvalue in (b), gather all the histograms in one graph with 2 by 2 multiple plots, indicating various data sources by assigning different colors and titles to the four histgrams. All the histgram should have 10 classes.

(d) [4 points] Find out all the companies with missing profits, obtain the simple statistics (e.g. five numbers) for the sales of these companies.

(e) [5 points] Each company belongs to one country. Find out how many different countries are there. Count the total number of companies belong to the same country. Calculate the mean and median assets of all the companies belong to the same country. Formulate the results in one dataframe with four variables, named as “country”, “num_of_companies”, “mean_assets”, “median_assets” respectively, recording the name of country, the number of companies, mean and median value of assets belong to the specific country. Export the dataframe to a txt file “countries.txt”.

(f) [5 points] Use two different ways to select all the companies whose sales are larger than 100, keep five variables “rank”, “name”, “sales”, “profits” and “assets”, sort the selected dataset by “sales” in descending order and “assets” in ascending order, obtain the scatter plot of profits versus assets of these companies, name both axes and color the dots in blue.

(g) [5 points] From(d) we know there are companies with missing profits, try to impute the missing profits of those companies via the K-nearest neighbor method (take K=10) by making use of the other three numerical values (“sales”, “assets”, “marketvalue”).

1.3 [56 points] Problems in elementary statistics and probability:

(a) [5 points] Plot the density function of χ^2 distribution with degree of freedom 4 and 6, the range of x-axis spans from 0 to 8. Calculate the area under the density curve for the interval $[1, 7]$ and the interval $[3, \infty)$ of a $\chi^2(5)$ distribution.

(b) [5 points] Suppose that X is an Exponential random variable with rate λ and $\Pr(X = 1) = 0.2$, then obtain the rate λ and $\Pr(X = 6)$.

(c) [6 points] Let X_1, X_2 be two independent standard uniform random variables. Define $Y = X_1 + 2X_2$, derive the pdf of Y and obtain a plot of the pdf of Y.

(d) [4 points] Obtain a plot of density functions for the standard normal and $t(1)$ distribution, compare whose tail is heavy.

(e) [6 points] The Hypergeometric Distribution: a hypergeometric random variable describes the number of successes in a sequence of n draws from a finite population without replacement. Suppose that an urn contains N balls, m of which represent non-defective items (hence success!), and the remaining $N - m$ are defective items. We draw n items without replacement from the urn and note the number of non-defective items. Let X be the random variable denoting the number of non-defective items. Then the probability distribution of the number of non-defective items is given by

$$\Pr(X = x) = \begin{cases} \frac{\binom{m}{x}\binom{N-m}{n-x}}{\binom{N}{n}}, & x = 0, 1, \cdots, \min\{n, m\}, \\ 0, & \text{otherwise.} \end{cases}$$

Derive the general form of mean and variance for the Hypergeometric distribution. Suppose that a folder on a drive of a hard disk contains $N = 45$ files. Of the 45 books, 28 are statistical text and the remaining are science texts. Assume that 8 files are selected at random from the folder and X statisitcal texts are selected, use R to compute the $\Pr(X = k), k = 0, 1, \ldots, 8$ and the mean and variance of X.

(f) [5 points] Suppose X follows the Poisson distribution with intensity rate λ. Using the dpois function, find the maximum value of λ, such that $\Pr(X > 8) \leq 0.4$.

(g) [5 points] *Graphically* prove that Exponential random variable with rate $\lambda = 3$ and Gamma random variable with shape $\alpha = 1$ and rate $\beta = 3$ have the same pdf.

(h) [5 points] What is the number of people whose birthday you need to ask so that the probability of finding a birthday mate is at least 3/4? Write a brief R program to obtain the size as the probability varies from 0 to 1.

(i) [5 points] Suppose that $X_i \sim \text{Bernoulli}(p)$, $i = 1, 2, \cdots$. Let $S_n = \sum_{i=1}^{n} X_i$, $n = 1, 2, \cdots$. Define $Z_n = \frac{S_n - np}{\sqrt{np(1-p)}}$, the de Moivre-Laplace Central Limit Theorem states that as $n \to \infty$,

$$\Pr(a < Z_n \leq b) = \int_a^b \frac{1}{\sqrt{2\pi}} e^{-x^2/2} \, \mathrm{d}\, x$$

Use R to testify the de Moivre-Laplace Central Limit Theorem.

(j) [5 points] Poisson Limit Theorem. Let X_i be independent random variables $X_i \sim \text{Bernoulli}(p_n)$ and let $S_n = \sum_{i=1}^n X_i$. Assume that as $n \to \infty$, $p_n \to 0$ and $np_n \to \lambda$, where λ is a fixed finite constant, then $S_n \to \text{Poisson}(\lambda)$. That is,

$$\Pr(S_n = k) \to \frac{e^{-\lambda}\lambda^k}{k!} \qquad \text{as } n \to \infty.$$

Use R to testify the Poisson Limit Theorem.

(k) [5 points] Let X follow a geometric distribution. Set up an R program to evaluate $\Pr(X > m + n|X > m)$ and $\Pr(X > n)$, for non-negative integers m and n. Does the memoryless property of exponential random variable hold for geometric distribution? Test the memoryless property for a Poisson random variable too.