University of Tehran
School of Electrical and Computer Engineering

Statistical Inference

Instructor: Dr. Hossein Vahabie

Assignment 2

Probability and R

Teaching Assistants:
Hamed Gholami (hamedgholami@ut.ac.ir)
Melika Sadeghi (melikasadeghi16@gmail.com)

Winter 2022

Homework 2

Statistical Inference, Winter 2022

Question 1

Answer the following questions.

a. X and Y are standard normal random variables. Find the correlation between smaller one and the bigger one.
b. Is it possible that a random variable be independent of itself? If so, when?
c. Can there be two random variables say, W and Z such that E(W) = 50 * E(Z) and Z is greater than W with probability at least 0.98?

Question 2

"Why too much evidence can be a bad thing" is an article by Lisa Zyga which says:

a. Under ancient Jewish law, if a suspect on trial was unanimously found guilty by all judges, then the suspect was acquitted. This reasoning sounds counterintuitive, but the legislators of the time had noticed that unanimous agreement often indicates the presence of systemic error in the judicial process.
b. Assume we have n judges deciding a suit. The prior probability of a suspect being guilty is p. each of the judges has to vote whether to acquit or convict the suspect. With probability s a systematic error happens (e.g., the defense would be unable to state their argument clearly enough so judges could understand). If a systematic error happens, then there would be a unanimous vote for convicting the defendant (i.e., all the judges vote to convict).
c. Whether the defendant is guilty or not is independent of a systematic error. Given that the defendant is guilty and a systematic error does not happen, each judge has a probability c of convicting, independent of other judges. Given that defendant is not guilty and a systematic error does not happen, each judge has a probability v of convicting, independently. We know that $v < 0.5 < c$.
   a. Assume for $k < n$, that precisely k of the judges vote to convict. Given this information, find the probability that the defendant is guilty.
   b. Now assume that all the judges vote for convicting. Given this information, find the probability that the defendant is guilty.
   c. Is the answer of last part, viewed as a function of n, an increasing function? Give an intuitive explanation in words. Also, explain whether the observation of the Jewish legislators was right or wrong?

Question 3

We model the emails that arrive in Bob's inbox by a Poisson distribution with rate λ, measured in emails per hour; explain why we are justified to use this model?

Each email is work-related with probability p and personal with probability q = 1 - p. The amount of time it takes to answer a work-related email is a random variable with mean μ_W and variance σ^2_w, the amount of time it takes to answer a personal email has mean μ_p and variance σ^2_p, and the response times for different emails are independent. What is the average amount of time Bob has to spend answering all the emails that arrive in a t-hour interval? What about the variance?

Hint: use the law of iterated expectation and the law of total variance.

Question 4

Consider all of the permutations of a sequence of numbers like $x_1, x_2, \dots, x_n$. A sequence is said to have a local minimum at position i if $x_i < x_{i-1}$ and $x_i < x_{i+1}$ for $2< i < n - 1$; for i = 1 a local minimum means $x_1 < x_2$; similarly, for i = n a local minimum means $x_n < x_{n-1}$. What is the expected number of local minimums in all of these permutations?

Hint: define indicator random variables that would sum up to the desired variable and consider the symmetries, then, apply linearity of expectation.

Question 5

There are 150 people in a queue waiting to attend a theater. Each has a ticket corresponding to his seat (there are exactly 150 seats). Unfortunately, the first person loses his ticket in an accident and randomly sits on a seat (with all seats equally likely). Each of the next attendees would sit on his seat if the seat is empty or would sit at random (with all the remaining empty seats equally likely). What is the probability that the last person would sit on his own seat?

Hint: what are the possibilities for the seats available to the last person?
Hint: watch for symmetries.

Homework 2

Statistical Inference, Winter 2022

Question 6

Mr. John Doe has a family with two children. We ask him whether at least one of his children is a boy that was born in Tuesday and his answer is positive.

a. What is the probability that the other child is a boy?
b. Assume instead of asking Mr. Doe we saw him at confectionary ordering a cake for Tuesday and when we ask him whose cake is it he answered it is for my son. What would be the probability that the other child is a boy?

Question 7®

There exist 5 types of positions in a software company: “UI Developers”, “Back-end Developers”, “management”, “HR” and “HSE”. Each type has 8, 12, 4, 3, 3 employees in order. Answer the following questions about this company:

a. Create two vectors containing the position types and their corresponding population.

b. Plot a bar chart of distributions of employees in different position types. Note that the plot must have a proper title in green color, and the x-label and the y-label should be in blue.

c. Consider each position type has below salaries. Visualize a plot with 5 boxplots that shows salaries of different groups of position types.
UI Developers →75000, 25000, 48000, 42000, 35200, 45000, 23000, 45500

Back-end Developers → 20000, 80000, 36000, 46300, 41000, 43000, 22000, 37000, 39000, 43500, 69000, 5000

Management →80000, 67000, 56000, 82000

HR→45000, 39000, 30000

HSE→12000, 25000, 31500

d. According to part c, find the exact quartile of the salary of each group and calculate IQR. Are there any outliers in any group? What are the exact values? Show the calculation of detecting outliers.

e. Discuss the skewness of distributions in each group salary. Then plot histogram and density plot of each group in 5 plots (both histogram and density plot must be in a single plot).

f. Categorize all employees based on their salary into 5 groups: “very high” (>50000), “high” (>40000), “middle” (>30000), “low” (>20000), and “very low” (<=20000). Plot a pie chart that visualizes the frequency of these five categories. Each category must have a percentage and should have a unique color. Draw a legend for your pie chart.

g. For back-end developers group calculate mean, median, variance, and standard derivations.

Question 8®

Below are the final exam scores of twenty introductory statistics students.

57, 66, 72, 78, 79, 79, 81, 81, 82, 83, 84, 87, 88, 88, 89, 90, 91, 92, 94, 95

a. Create a vector of scores.

b. Calculate median, mode, variance and standard deviation of scores.

c. Are there any outliers in any group? What are the exact values? Show the calculation of detecting outliers.

d. Plot the boxplot.

e. Plot the histogram and the density of scores in a single plot.

i. Based on the plots, discuss the skewness of scores.

ii. Based on the plots, would you expect the mean of this dataset to be smaller or larger than the median? Explain your reasoning.

iii. What is the best measurement of the center for the scores? Why?

Qusetion 9®

Create a numeric vector with 6 elements as follows: 25,30,35,20, 25. These values are the number of students in five different classes: “class1”, “class2”, “class3”, “class4”, “class5”.

a. Add the corresponding classes as the name of the elements to the vector you have created.

b. Plot a bar chart to analyze the distribution of the number of students. The bar chart should have a title with a dark green font and the axes labels should be in red.

c. Create the “top_student” vector with 5 elements 3,5,2,1 and 4 that shows the number of students with a GPA above 19. Plot a scatterplot that displays the relationship between the number of top students and the number of students in the class. Change the plotting character to “*”.

d. Could it be conducted from the scatterplot in that there is a relation between these two variables?

e. GPAs of the students in “class4” are as follows. Find the exact quartiles of the grades by plotting the boxplot. Are there any outliers? What are the exact values?

“19.5, 16.75, 13.5, 16.25, 5, 12.5, 15.5, 15, 11, 16.5, 12.75, 15.5, 12.75, 9.75, 11, 14.5, 16.75, 11.5, 17, 18.25”

f. Discuss the skewness of the distribution of the grades in “class4” by plotting its histogram and density plot. What is the difference between these two plots?

g. We want to categorize the grades into “D” (<13), “C” (<15), “B” (<18) and “A” (>=18). Plot a pie chart that visualizes frequency of these four categories. Your chart should be colorized and the labels should contain each category with its percentage