Titanic Data Analysis

I. INTRODUCTION

The Titanic's narrative is the most common story in the last centuries. The most passengers were killed after the ship collided with an iceberg and sank to the bottom of the ocean. It is also well-known that women and children were given higher priority in the lifeboats and priority was also given to first and second-class passengers over third-class passengers. Those unlucky persons who did not have access to lifeboats, either drowned or froze to death in the icy water. Although the Titanic story is tragic, it provided a great opportunity to conduct statistical analysis and learn the numerical story of the Titanic. Based on this, the statistical question was clear: What was the statistical effect of lifeboat selection on the survival rate of the passengers, based on gender, age, and passenger class?
Moreover, the answer to that question helps the largest shipbuilders to take into consideration the necessary numbers of lifeboats that must be on the ship in order to save the largest number of people. Comprehensively, this type of thinking formed the concept of "a natural progression toward using the data to improve estimates, forecasts, decisions, and ultimately, efficiency".

II. GENERAL RULES

1. ZERO will be assigned to copied reports.
2. You should work in a team consisting of 5 students.
3. The report will be evaluated from 7 and bonus marks will be given to outstanding reports.
4. A report must be prepared and include (introduction about your project, your R-code, results, comments on results, and problems you have faced and how you overcame it
5. Softcopy of the report must be sent to probabilityfcds@gmail.com before the final discussions, moreover, a hardcopy from report must be brought while the discussion.
6. You must respond to all project questions (parts).
7. The report must be written using 10 font size and Times New Roman style.
8. Team members could be from different departments.
9. No excuses for delay.
10. The final discussion and deadline for the project will be on December 20, 2022.

III. QUESTIONS

- Download the necessary data via the URL below
  https://drive.google.com/file/d/1vNNtMwlHdIyQTT5ty12aOOlapp-5vM37/view?usp=share_link

- Import the dataset (called "train.csv") into RStudio and store it in a variable called "titanic."

Q1) Provide an accurate data summary (minimum, maximum, first quantile, third quantile, median, and mean) utilizing RStudio's built-in summary function.

Q2) In statistics, there are two main categories of variables: quantitative and qualitative. Can you explain the difference between these two types of variables? Then, categorize the titanic data set's variables into these two categories.

Q3) In the dataset, there are 891 observations and 12 variables; some of these variables have missing data (have a value of NA); answer the following concerns.

i) Can you count the missing data in the Age variable?
ii) Is it normal for a data set to contain missing data?
iii) Can you provide solutions to estimate this missing data?
iv) Indicate whether or not it will affect our statistics?

Q4) Remove these missing data using the following command >>> titanic=na.omit(titanic)

Q5) Draw the histogram for the dataset's Age variable.

Q6) By inspection (from the histogram), which distributions may be employed to provide a good fit for the Age and Fare histograms?

Q7) Assuming that the 891 observations in the dataset represent the population, and that the distribution of ages is normal, determine the true mean and standard deviation for the age variable?

Q8) Take a random sample of size 50 from Age. Using this sample, what is your point estimate of the population mean and standard deviation?

Q9) Since you have access to the population, simulate the sampling distribution for $\overline{Age}$ by taking 50 samples from the population of size 50 and computing 50 sample means. Store these means in a vector called sample_means50. Plot the data, then describe the shape of this sampling distribution. Based on this sampling distribution, what would you guess the mean age of the population to be?

Q10) Simulate the sampling distribution for $\overline{Age}$ by taking 100 samples from the population of size 50 and computing 100 sample means. Store these means in a vector called sample_means100. Plot the data, then describe the shape of this sampling distribution. Based on this sampling distribution, what would you guess the mean age of the population to be?

Q11) Simulate the sampling distribution for $\overline{Age}$ by taking 1000 samples from the population of size 50 and computing 1000 sample means. Store these means in a vector called sample_means1000. Plot the data, then describe the shape of this sampling distribution. Based on this sampling distribution, what would you guess the mean age of the population to be?

Q12) What happens to the sampling distribution when the number of samples increases?

Q13) Since you have access to the population, simulate the sampling distribution for $\overline{Age}$ by taking 1500 samples from the population of size 20 and computing 1500 sample means. Store these means in a vector called sample_means_s20. Plot the data, then describe the shape of this sampling distribution. Based on this sampling distribution, what would you guess the mean age of the population to be?

Q14) Simulate the sampling distribution for $\overline{Age}$ by taking 1500 samples from the population of size 100 and computing 1500 sample means. Store these means in a vector called sample_means_s100. Plot the data, then describe the shape of this sampling distribution. Based on this sampling distribution, what would you guess the mean age of the population to be?

Q15) Simulate the sampling distribution for $\overline{Age}$ by taking 1500 samples from the population of size 200 and computing 1500 sample means. Store these means in a vector called sample_means_s200. Plot the data, then describe the shape of this sampling distribution. Based on this sampling distribution, what would you guess the mean age of the population to be?

Q16) What happens to the sampling distribution when the size of each sample increases? check the compatibility of the results with the central limit theorem.

Q17) Since you have access to the population, simulate the sampling distribution for the sample variances by taking 1500 samples from the population of size 2 and computing 1500 sample variance. Store these sample variances in a vector called sample_U1500. Plot the data, then describe the shape of this sampling distribution of variances. Based on this sampling distribution, does it follow normal distribution or not?

Q18) Simulate the sampling distribution for the sample variances by taking 1500 samples from the population of size 50 and computing 1500 sample variance. Store these sample variances in a vector called sample_U1500. Plot the data, then describe the shape of this sampling distribution of variances. Based on this sampling distribution

i) what happens to the shape of the sampling distribution
ii) using this sampling distribution to estimate the population variance?

Q19) Take a random sample of size 50 from Age. Using this sample to estimate the mean Age of the population using MLE and MME, then determine the bias of those estimators?

Q20) Take a random sample of size 200 from Age. Using this sample to estimate the mean Age of the population with MLE and MME, does the bias increase or decrease as the sample size increases? Can you identify which estimators are the most effective?

Q21) Separate the age column of the titanic dataset into two groups based on gender (male and female) with variable name age_male_(anyone in the team ID) and age_female_(anyone in the team ID), then simulate the sampling distribution for $\overline{age\_male} - \overline{age\_female}$ by taking 15000 samples from the male and female populations of size 50 and computing 15000 sample difference means. Store these means in a vector called samplediff_means15000. Plot the data, then describe the shape of this sampling distribution. Based on this sampling distribution, do you believe there is a significant difference in the average ages of men and women in titanic?

Q22) Separate the Survived column of the titanic dataset into two groups based on gender (male and female) with variable name Survived_male and Survived_female, then simulate the sampling distribution for Survived_male − Survived_female by taking 15000 samples from the male and female populations of size 50 and computing 15000 sample difference between the number of survivals in males and females. Store these differences in a vector called samplediff_Survived 15000. Plot the data, then describe the shape of this sampling distribution. Based on this sampling distribution, do you think there is bias in the rescue process between males and females?

Q23) Take a random sample of size 10 from Age. Using this sample, Find 95% confidence interval for the mean of the ages?

Q24) Take a random sample of size 50 from Age. Using this sample, Find 95% confidence interval for the mean of the ages?

Q25) Take a random sample of size 200 from Age. Using this sample, multiply these sample ages by 5 to determine what happens to the E(age) and var (age) after multiplication?

Q26) Take a random sample of size 200 from Age. Using this sample, add these sample ages to 5 to determine what happens to the E(age) and var (age) after addition?

Q27) Online task: Read about the kernel distribution and write a short summary of its applications?

Q28) Online task: Read about KS test and write a short summary of its applications?