Lab 3

Statistical Inference, Spring 98

In this Lab, we focus on the foundations for statistical inference. We discuss CLT, CIs, simulations and hypothesis testing.

a) Central Limit Theorem

The Poisson distribution is a discrete probability distribution that expresses the probability of a given number of events occurring in a fixed interval of time or space if these events occur with a known constant rate and independently of the time since the last event.

1- Take three natural phenomena examples that are Poisson distributed.

2- Plot Poisson distribution with lambda 1, 2, 3, …, 10 in one frame. Discuss the skewness of Poisson distribution in relation with lambda.

We can create an assumptive population with the size of 100,000 that is Poisson distributed (lambda 4) using the following command:

```
> population <- rpois(100000, 4)
```

3- Plot histogram of the population distribution. Find the mean of the distribution and show it on the plot with a vertical line.

"Sample" function takes a random sample of the specified size from the elements of a vector.

4- Sample 15 groups of data with the size of 100 from the population. Plot all the sample histograms with their mean lines in a 5*3-grid canvas. Discuss the similarity of the sample distributions with the population. You can use QQ-Plot for this purpose.

5- Sample 500 groups of data with the size of 50 from the population. Create a vector named "samples_mean" which consists of the mean of the samples. Now, plot histogram of "samples_mean". What distribution is the sampling distribution similar to? Using QQ-Plot, test the similarity of the sampling distribution with both the population distribution and the Normal distribution. Explain the results with CLT.

6- Does it make any difference for the sampling distribution if the population distribution was different? Discuss.

7- Find the mean and the standard error of the "samples_mean". Then calculate $\hat{\mu}$ and $\hat{\sigma}$ with the help of those values. Compare the estimated values of the mean and standard deviation of the population with the true statistics of the population.

b) Confidence Intervals

Galton dataset ("mosaicData" library) contains results of a survey on the height of children and their parents. Consider the dataset as a population. Also, in the following questions, assume that the standard deviation of the population is unknown.

8- Create a variable named "heights" which contains the "height" column of this dataset.

9- Create 10000 samples of size 50. Form the 97% confidence interval for each of the samples. Find the true mean of the population and check if it is included in the CI of each sample. What is the percentage of the samples that contain true population mean in their CIs? Explain the results.

10- Repeat the above exercise with 10000 samples of size 10 with 90% CIs. Explain if the conditions of CI formation are satisfied and compare the results with the previous part.

11- Write a function named "calculate_ci" which accepts a sampled data (vector) and confidence level (two tails) as the arguments, and calculates the lower and upper bound of CI and printout the

result in an appropriate format. Also, it should check the conditions of CI formation (the ones that could be validated with the available information). If the conditions are not satisfied, the function should print a warning message and explain the reason.

12- Take a sample of size 60 from "heights" population. Draw a graph which shows the relation between confidence level and the length of the interval. Set the range of the x axis from 50% to 100%. Interpret the graph.

c) Simple Simulations

In a simulation, you set the ground rules of a random process and then the computer uses random numbers to generate an outcome that adheres to those rules. As a simple example, you can simulate flipping a fair coin for one time with the following commands:

```
> outcomes <- c("head","tail")
> sample(outcomes, 1, replace = TRUE)
```

13- Simulate rolling a fair dice with 6 faces for 15000 times. Plot the distribution of the outcomes. Show the percentage of each outcome category with the help of the "table" command.

14- Mr. Murphy eats jam and butter toast every morning for breakfast. Mr. Murphy believes that if he throws the toast, it will definitely land on its buttered and jammed side. If we know the weight of the bread is 20 grams and there is 4 grams of butter and 8 grams of jam rubbed on it; Simulate the bread falling for 1000 times. Find the distribution of outcomes and report the frequencies.

15- Two fair 6 faces dices are rolled. What is the probability that sum of the outcomes of two dices is greater than 8? Simulate the problem for 100000 times. Compare the simulation result with the probability that is calculated theoretically.

d) Hypothesis Testing

Weather dataset ("mosaicData" library) contains weather information for some cities in 2016-17.

16- Scientists say that Beijing is a city with an average humidity of 50%. Does the data confirm this proposition? Form the hypothesis test framework, check the conditions, specify the null and alternative clauses and find the p-value. Make a decision with alpha=0.05 and interpret the results.

17- Apart from the results of running the hypothesis test in the previous exercise, and with the knowledge that the average humidity of the Beijing city is about 53% in this dataset, could you really reject the scientists proposition about average humidity in Beijing? If not, what is the problem with your inference?

Shrimp dataset ("MASS" library) contains the amount (percentage of the declared total weight) of shrimp in shrimp cocktail measured by different laboratories sampled from all over the world.

18- A restaurant recipe claims that the amount of shrimp in shrimp cocktail is much less than 31% in all over the world. Does the shrimp data confirm this claim? Run the hypothesis test framework, check the conditions, calculate p-value and explain the results.

19- Write a function named "two_tail_z_dist_mean_hyp_test" which accepts a sampled data (as a vector), null hypothesis value of mean and alpha value as its arguments. It should then printout the null and alternative hypotheses, calculate and print the p-value, and print the decision made. It should also check the conditions of the test and print a warning message with if some of them are not satisfied.

20- Run this function on problems stated in exercises 16 and 18.