QTM 220 Homework 1

Due date: 12 September 2024

September 6, 2024

# Introduction

This exercise sheet contains a series of problems designed to test and enhance your understanding of the topics covered in the course. Please ensure that you attempt all problems and provide detailed solutions where necessary. If you have any questions or need clarification, feel free to reach out your TA.

# Exercises

## Exercise 1: Playing with a Gaussian distribution - 25pts

1. Use the command **rnorm(n = 100, mean = 0, sd = 1)** to sample data point from a standard normal distribution. Explain in words the meaning of the parameters **n**, **mean**, and **sd**. **[5pts]**

n = number of observations (there are 100 observations).

mean = the average of the data set (found by dividing the sum of the data set by the number of observations) which is 0 for a standard normal distribution.

sd = the standard distribution of the data set (quantifies the spread of the data set and is always positive).

1. Compute mean, median, standard deviation and median absolute deviation. **[5pts]**

A screenshot of a computer code

Description automatically generated

1. Plot the distribution of data generated in (a) using the command hist (or ggplot). Overlay on the plot the position of the mean and the median. **[5pts]**

A screenshot of a computer program

Description automatically generated

A graph with a line graph

Description automatically generated with medium confidence

1. Describe the distribution plotted, taking into account the position of mean, median and mode(s), and the skewness. **[5pts]**

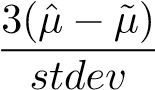
This is most likely a standard normal distribution since the mean and median are very close to each other. That said, the local maxima make it look like this is a bimodal distribution. However, both maxima are close to the center of the graph where the mean and median are located. Finally, there is almost equal skewness on either side of the curve. Therefore, this is most likely a standard normal distribution.

1. Now compare the mean obtained in (b) with the mean you expect from the data generation in (a). Are they the same? Explain why. **[5pts]**

Since I made sure to use set.seed() to make the results reproducible in the beginning, these are the same value. That said if I had not used set.seed(), the code would generate a new sample every time and the mean from (b) would not be the same as the mean in (a).

## Exercise 2: Pearson’s second skewness coefficient - 25pts

The Pearson’s second skewness coefficient is a coefficient based on the median used to measure the skewness of a distribution. Here is its formulation:



where ˆ*µ* is the mean, ˜*µ* is the median, and *stdev* is the standard deviation. Keep in mind the following criteria to help you in the discussion of the next questions: if the coefficient is zero, it means no skewness at all; if the coefficient has a large negative value it means that the distribution is negatively skewed; if the coefficient has a large positive value it means that the distribution is positively skewed.

1. Code a function to compute the Pearson’s second skewness coefficient (you can use R’s built-in commands now) **[10pts]**.

A close-up of a computer code

Description automatically generated

1. Simulate from a *β* distribution using the command rbeta(1000, 0.45, 5) and generate an histogram. Use the function written in (a) to compute the Pearson’s second skewness coefficient **[5pts]**.

A white background with green and blue text

Description automatically generated

A graph of a bar graph

Description automatically generated

1. Now simulate from a *β* distribution using the command rbeta(1000, 5, 0.45) and generate an histogram. Use the function written in (a) to compute the Pearson’s second skewness coefficient **[5pts]**.

A white rectangular object with green numbers

Description automatically generated

A graph of a graph

Description automatically generated

1. Describe and compare the two histograms and the two Pearson’s second skewness coefficients obtained in (b) and (c) **[5pts]**.

The histogram in (b) has positive skewness and a positive Pearson’s second skewness coefficient. The histogram in (c) has negative skewness and a negative Pearson’s second skewness coefficient.

## Exercise 3: A constant case - 15pts

You have a vector *x* of *n* numbers equal to a constant value *a*. What are the values of the mean, median, standard deviation, and median absolute deviation? For each, explain why or write a short proof.

The values of the mean and median are equal to the constant value *a*. This is because the average value of a set that only contains only value *a* will be equal to the unique value in the set. Similarly, placing all numbers in order will give you a median value of *a* because it is the only value present in the set. Furthermore, the standard deviation and median absolute deviation are both equal to 0. This is because the set does not deviate from value *a*, therefore the standard deviation is 0. Finally, subtracting the median *a* from every value in the set results in 0 for every iteration. Therefore, dividing by n results in a median absolute deviation of 0.

**Exercise 4: Multiple choices - 5pts** Select the correct answer.

- Which of the following measures can have more than one value for a set of data? **[1.5pts]**

1. Mean;
2. Mode;
3. Standard deviation;
4. None of the above;

- Which of the following measures can be determined for quantitative data? **[1.5pts]**

1. Median;
2. Mode;
3. Standard deviation;
4. All of these;

- The mean of four numbers is 71.5. If three of the numbers are 58, 76, and 88, what is the value of the fourth number? **[2pts]**

1. 64;
2. 70;
3. 76;

(d) 82;

**Exercise 5: Sampling Distributions and Estimators - 25pts** Use the provided data on Pokémon for the following exercises.

1. Draw a random sample with replacement from the population of Pokemon of size *n* = 100 (make sure to set a seed so your work is replicable). Report your sample mean of HP (hit points).**[5pts]**

A screenshot of a computer code

Description automatically generated

Mean HP = 68.96

1. Create a sampling distribution of the sample mean of size *n* = 100 with 10,000 iterations. Make a histogram of your sampling distribution with a vertical line indicating the population mean HP. **[10pts]**

A screenshot of a computer program

Description automatically generated

(added + theme\_minimal() for the final plot)

A graph of a bar graph

Description automatically generated

1. Find the width of the middle 95% of your sampling distribution. If you created an interval of that width centered on your sample mean, would it cover the population mean HP? **[10pts]**

No. This is because we don’t have the estimand to check that. It’s possible that our sample could have come from the tails of the sampling distribution.

A screenshot of a computer program

Description automatically generated

A graph with a line in the middle

Description automatically generated with medium confidence

## Exercise 6: Visualizing Confidence Intervals - 25pts

Navigate to this [web app](https://seeing-theory.brown.edu/frequentist-inference/index.html) [https://seeing-theory.brown.edu/frequentist-inference/index.html] and choose the confidence interval section.

1. Select the Normal distribution and a very small sample size (*n <* 5). Set your confidence interval coverage to 95% and start sampling. What does it mean when the interval is turquoise? What does it mean when the interval is pink? What is the relationship between the sample (the orange balls) and the interval width? **[5pts]**

When the interval is turquoise, it means that the standard deviation of each sample mean contains the estimated mean. When the interval is pink, it means that the standard deviation of each sample mean does not contain the estimated mean. The more spread apart the orange balls, the larger the interval width. Therefore, the interval width represents the standard deviation of each sample.

1. Change your sample size to something larger (*n >* 25), but leave your distribution and confidence level the same. What changes? What is different about these CIs than your last simulation? What is the same? **[5pts]**

The interval width decreased. Additionally, the sample intervals are much closer to the estimated mean. Most of the sample intervals include the estimated mean in either simulation.

1. Leaving your sample size and distribution the same, change your confidence level to 50%.What is different about these intervals? Change your confidence level to 99%. What changes? **[5pts]**

When changing the CI to 50%, the interval decreases and more of the sample intervals do not include the estimated mean. When changing the CI to 99%, the interval increases and most of the sample intervals include the estimated mean.

1. Change the distribution to the Exponential and repeat (a)-(c). Does a different underlying distribution change anything? Discuss. **[10pts]**

A different underlying distribution does not change anything. As you increase the CI, the interval width increases. Additionally, as you decrease the sample size, the interval width increases. When you increase the sample size the interval width decreases. That said, as you increase the sample size, more of the sample sizes include the estimated mean. This is all the same as the normal distribution.

# Submission Instructions

Please submit your completed exercises by **12 September** through **gradescope** (connected to our course Canvas page). Ensure that your solutions are well-organized, clearly written, and include all necessary calculations and explanations. Questions about submission should be directed to your TA.

# Helpful Resources

To better assist you in the completion of this exercise sheet, we suggest you to review the following material:

* **Lecture 1** - covering the summaries of location, spread, and distribution;
* **Lab 1** - covering the introduction to R, coding of mean, median, standard deviation, and median absolute deviation, plotting of distribution;
* **Lecture 2** - covering sampling distributions and perfectly calibrated confidence intervals (using population)
* **Lab 2** - practicing sampling and sampling distribution coding