Name: Jinhyung Park
GitHub Username: Park1447
Purdue Username: park1447
Instructor: Yi Ding
Section: 001

Problem 1

1.

Given the probability density function:

$$f_X(x) = \begin{cases} \frac{3}{4}(1-x^2) \; when \; x \in [-1,1] \\ 0 \qquad otherwise \end{cases}$$

If you happen to plot this distribution, then by looking at the plot, determine by inspection, what is the mean of the distribution.

Mean of the distribution:

0

From this distribution we grab 100 i.i.d samples X_1, X_2,… X_{100}.
The sample mean random variable is defined as

$$\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$$

What is the shape of the probability density function of the sample mean?

Normal distribution (Gaussian)

What is the mean of the probability density function of the sample mean?

0

Explain how you know the shape and mean.

According to central limit theorem, the sufficiently large sample size like n > 30, the sample mean would approach to a normal distribution. In the end, it leads to Gaussian shape. Also, the mean of probability density function would be zero since it is symmetric about x = 0. In addition, the sample mean should be zero since it follows the normal distribution.

Part B

Let X be a normally distributed random variable with mean = 0 and variance = 1. What is the distribution of the random variable Z = 3X+1 (Give your answer with the shape, mean and variance).

What is the shape of the distribution?

Normal distribution

$\mathbb{E}[X] = 0 \quad Var[X] = 1$

What is the mean?

1

$\mathbb{E}[3X+1] = 3 \cdot 0 + 1 = 1$

What is the variance?

9

$Var[3X+1] = 9 \cdot 1 = 9$

Problem 2

You are given data in *city_vehicle_survey.txt* representing the average age of vehicles across various counties. The transportation department claims the average age of vehicles is 5 years. You are tasked with testing this claim.

1. Formulate null and alternative hypotheses for a statistical test that seeks to challenge this belief. What are the null and alternative hypotheses?

   1. Null Hypothesis:

      The average age of vehicles is 5 years

   2. Alternative Hypothesis:

      The average age of vehicles is not 5 years

   3. What type of test should be used and why?

      In this case, we should use z-test since we have large sample size.

2. Carry out this statistical test using the *city_vehicle_survey.txt* sample. Report the sample size, the sample mean, the standard error, the standard score (z or t, depending on what was used), and the p-value.

   ******ROUND ALL DECIMAL VALUES TO 4 DECIMAL PLACES****

Sample size	1024
Sample mean	5.1141
Standard error	0.0614
Standard score	1.8585
p – value (if less than 0.01 use scientific notation)	0.06310

Are the results statistically significant at a level of 0.05?

Yes [✓ **No**]

What (if anything) can we conclude about the hypothesis at the confidence level of 0.05?

It means the data fails to reject the null hypothesis since it is larger than 0.05. Consequently, the sample mean could be 5 years or it does not provide enough evidence to reject the mean is not 5 years.

Are the results statistically significant at a level of 0.10?

[✓ **Yes**] **No**

What (if anything) can we conclude about the hypothesis at the confidence level of 0.10?

Rejecting the null hypothesis is reasonable since the p-value is lower than 0.10 level.

3. What is the largest standard error for which the test will be significant at a level of 0.05? What is the corresponding minimum sample size? (You may assume that the population variance and mean does not change.)

   ********ROUND ALL DECIMAL VALUES TO 4 DECIMAL PLACES******

Largest standard error	0.0582
Corresponding minimum sample size	1139

4. Suppose the transportation department believes the mean vehicle age is the same in counties with and without emission control programs. Two datasets, *vehicle_data_1.txt* (with emission programs) and *vehicle_data_2.txt* (without emission programs), are used to test this assumption.

   1. Null Hypothesis:

      No difference of means in two countries

   2. Alternative Hypothesis:

      There is difference of means in two countries

   3. What type of test should be used and why?

      Z-test should be used since the sample size is large enough

5. Carry out this statistical test using the *vehicle_data_1.txt* population and *vehicle_data_2.txt* population samples. Report the sample sizes, the sample means, the standard error, the z-score, and the p-value. Are the results significant at levels 0.05 or 0.10? What (if anything) can we conclude about the hypothesis at the two different confidence levels?
   ********ROUND ALL DECIMAL VALUES TO 4 DECIMAL PLACES**** ******

Sample size of *vehicle_data_1* (Emission)	512
Sample size of *vehicle_data_2* (Without Emission)	868
Sample mean of *vehicle_data_1* (Emission)	5.5415
Sample mean of *vehicle_data_2* (Without Emission)	6.2482
Standard error	0.1044
Standard score	-6.7721
p – value (if less than 0.01 use scientific notation)	1.269 * 10^(-11)

1. Are the results statistically significant at a level of 0.05?

[✓ Yes] **No**

2. Are the results statistically significant at a level of 0.10?

[✓ Yes] **No**

3. What (if anything) can we conclude (i.e., what is the interpretation of the result)?

The result shows there is significant difference between two sample means since the p-value is extremely smaller than 0.05 or 0.10, suggesting to reject the null hypothesis.

Consequently, we can say that the emission and non-emission groups do not have same means.

Problem 3

1. Use the sample to construct a 90% confidence interval for the average sodium of snacks. Report whether you will use a z-test or t-test and report the sample mean, the standard error, the standard statistic (t or z value), and the interval. (Think, which distribution should you use here if very few data points are available?)

Since the sample size is small, we should use t-test

********ROUND ALL DECIMAL VALUES TO 4 DECIMAL PLACES******

Sample mean	140.4500
Standard error	1.3426
Standard score (t or z value)	t = 1.7291
90% confidence interval	(138.1285, 142.7714)

2. Repeat Q1 for a 95% confidence interval.

********ROUND ALL DECIMAL VALUES TO 4 DECIMAL PLACES******

Standard error	1.3426
Standard score (t or z value)	t = 2.0930
95% confidence interval	(137.6400, 143.2600)

Is your interval wider or narrower compared to using the 90% confidence interval in Q1?

[✓ **Wider**] **Narrower**

3. Repeat Q2 if you are told that the population standard deviation is 5.
   Will you use a t-test or z-test (Hint: Think which distribution should you use here now that you have the true population standard deviation)? Justify your answer.

Since we are given population standard deviation, we should use z-test

********ROUND ALL DECIMAL VALUES TO 4 DECIMAL PLACES******

Standard error	1.1180
Standard score (t or z value)	1.9600
95% confidence interval	(138.2587, 142.6413)

Is your interval wider or narrower than the interval computed in Q2?

Wider [✓ **Narrower**]