Name: Mirko Glavan
GitHub Username: mglavan07
Purdue Username: bglavan
Instructor: Qiu
Section: 2

Problem 1

1.

Given the probability density function:

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

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 $X_1, X_2, \ldots X_{100}$.

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?

Gaussian (Normal)

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

0

Explain how you know the shape and mean.

Because the samples are independent and identically distributed, thy each will have the same PDF. Therefore, the mean will not move. The shape of the plot will appear gaussian, or normally distributed according to the Central Limit Theorem. This is because the new distrubution is an average on n=100 i.i.d. samples.

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?

Gaussian (Normal)

What is the mean?

1

What is the variance?

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:

   Ho: The true population mean for the age of vehicles is 5 years (μ = 5 years)

   2. Alternative Hypothesis:

   Ha: The true population mean for the age of vehicles is not 5 years (μ $\neq$ 5 years)

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

   A 1-sample z-test with 2 tails should be used to test the population mean, μ.
   This is because Ha is in $\neq$ form and n = 1024 >> 30, which,
   by Central Limit Theorem, approximates the sample distribution as normal

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)	0.0631

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?

Since p > α (0.0631 > 0.05) we fail to reject the null hypothesis and conclude that the true population mean, μ, for car age is not different than 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?

Since $p < \alpha$ $(0.0631 < 0.10)$ we reject the null hypothesis and conclude that the true population mean, μ_1 for car age is different than 5 years.

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 (discrete)

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:

H_0: There is no mean difference in population means for vehicle age between the two counties $(\mu_0 - \mu_1 = 0)$

   2. Alternative Hypothesis:

H_a: There is a difference in population means for vehicle age between the two counties $(\mu_0 - \mu_1 \neq 0)$

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

The hypothesis test should use a 2-sample, 2-tailed z-test for difference in population means. This is because H_a is in $\neq$ form, there are 2 samples and n_0 and $n_1 >> 30$ (512, 868)

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)	1.2695 x 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)?

Based on the results, $p << \alpha$ $(1.2695 \times 10^{-11} << 0.1, 0.05)$ therefore we have enough evidence to reject the null hypothesis and conclude that the two means (μ) for ages of cars are different in counties without emission programs compared to counties with them.

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?)

We will conduct a 1-sample 2-sided t-interval for population mean. This was selected because n = 20 < 30, therefore CLT may not apply.

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

Sample mean	140.4500
Standard error	1.3426
Standard score (t or z value)	1.7291
90% confidence interval	(138.1285, 142.7715)

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)	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.

Now that the population std. deviation is known (σ), we can use a z interval. Even though the sample is the same, a known σ justifies normal distribution.

******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**