Name Surname: Yunus Kardaş

Student Number: 2014400120

CMPE 343: Introduction to Probability and Statistics for Computer Engineers

Fall 2020 Final Exam

Start: February 11, at 12:30 End: February 13, at 23:59

Important Notes

- Please type your answers. Handwritten exams will NOT be accepted.
- Include your name and student number on the first page.
- You do not need to type the questions again.
- Only .pdf files will be accepted.
- Show your steps clearly to get full points.
- DO NOT copy and paste anything from internet or another friend.
- Do your exam ON YOUR OWN. This is NOT a group activity.
- If you get benefit from a book or online resource, please cite the references.
- Academic dishonesty will not be tolerated because of any excuse.
- Late submissions will NOT be accepted.

Questions

1. (20 pts) Suppose there is a survey about interest in sports cars. The results show that 33% of men prefer to use sports car. There is a claim that the proportion of men, who prefer sports cars, is higher than the proportion of women, who prefer sports cars. To investigate the claim, we took a random sample of 86 men, where 38 of them prefer sports car. We also randomly sampled 94 women and 25 of them said they prefer sports car. What is your conclusion at $\alpha = 0.01$ level of significance?

   Solution:

$$H_0 : p_1 = p_2$$

$$H_1 : p_1 > p_2$$

$$\alpha = 0.01$$

$$CriticalRegion z > 2.326$$

$$p_1 = x_1/n_1 = 38/86 = 0.442$$
$$p_2 = x_2/n_2 = 25/94 = 0.266$$
$$p = (x_1 + x_2)/(n_1 + n_2) = 63/180 = 0.35$$

$$z = (p_1 - p_2)/(p(1-p)(1/n_1 + 1/n_2)) = 0.176/0.071 = 2.479$$
$$P = P(Z > 2.479) = 0.0066$$

Thus, we reject null hypothesis and agree that the proportion of this proposal.

2. (20 pts) Time complexity of training deep models is high. GPUs are commonly used to accelerate the training time. A student recently bought a GPU card. The GPU company claims a 40 minute improvement on average for training a certain benchmark model with a benchmark dataset. The student records the training time of 40 tests before and after the addition of the GPU to her computer using the same benchmark model and dataset. The before and after training times are saved in `before.npy` and `after.npy` files.

   (a) How should she test the company's claim? Explain.

   Solution:

   Because of she doesn't get any information about population and she makes dummy test, also she gets two different sample, she should use "Paired t-Test" to test company's claim

   (b) Test the company's claim at $\alpha = 0.01$. Comment on the test result.

   Solution:

$$t = (d_{mean} - d_0)/(s_d/\sqrt{n})$$
$$d_{mean} = 40.05$$
$$s_d = 1.65$$
$$d_0 = 40$$
$$n = 40$$
$$\alpha = 0.01$$

$$=> t = 0.2022$$

$$P(|T| > 0.2022) = 0.8408$$

$$0.8408 >> 0.01$$

Thus, company's claim fail to reject.

3. (20 pts) In this question, you are given two random samples from two populations. The data is provided as `sample1.npy` and `sample2.npy` in the exam folder. Find 95% confidence interval for the ratio of the population variances σ_1^2/σ_2^2.

Solution:

by Two Samples: Estimating the Ratio of Two Variances;

$$1f_{\alpha/2(v_1,v_2)}$$

$$(s_1^2/s_2^2) * (1/f_{\alpha/2(v1,v2)}) < \sigma_1^2/\sigma_2^2 < (s_1^2/s_2^2) * (f_{\alpha/2(v1,v2)})$$

$$\alpha = 0.05$$

$$v_1 = v_2 = 22$$

$$f_{\alpha/2(v1,v2} = 2.3579$$

$$s_1 = 10.2921 => s_1^2 = 105.9271$$

$$s_2 = 8.7985 => s_2^2 = 77.4138$$

$$(105.9271/77.4138) * (1/2.3579) < \sigma_1^2/\sigma_2^2 < (105.9271/77.4138) * 2.3579$$

$$0.5803 < \sigma_1^2/\sigma_2^2 < 3.2264$$

4. (20 pts) A random sample of measurements of the diameter of a cylinder component is given in `measurements.npy` file in the exam folder. The machine that produces those components recently made 5 more components. Engineers would like to know whether there are any anomalies among the 5 components. Diameters of the 5 components are given in the table below.

Diameters of the latest components
25.64
18.26
45.23
30.12
33.45

(a) Can you suggest an approach to assist the engineers in finding the anomalies? Explain your approach.

Solution:

Engineers had first data. After that they get 5 more data and they should find anomalies in the 5 more data. Thus, they should calculate future data by using first data and they test it with 5 more data that given late. In this situation, they should make future observation and for future observation, they can use prediction interval because prediction interval is more relevant than confident interval. Therefore, the prediction interval can be suggested to assist the engineer in finding anomalies.

(b) By using your approach, find out which of the 5 cylinder components seem problematic. Use $\alpha = 0.05$.

Solution:

by Prediction Interval;

$$x_{mean} - t_{\alpha/2}.s\sqrt{1 + 1/n} < x_0 < x_{mean} + t_{\alpha/2}.s\sqrt{1 + 1/n}$$

$$x_{mean} = 30.6102$$

$$s = 3.0518$$

$$\alpha = 0.05$$

$$v = 39$$

$$n = 40$$

$$=> 24.133 < x_0 < 37.087$$

Thus, 18.26 and 45.23 are anomalies because they are not between the intervals.

5. (20 pts) Given a normal random variable X with mean 15 and variance 16, suppose you take a random sample of size n.

   (a) Find what sample size is necessary in order that

$$P\left(14.51 \leq \bar{X} \leq 15.49\right) = 0.95$$

   Solution:

$$z = (X_{mean} - \mu)/(\sigma/\sqrt{n})$$
$$(X_{mean} = 15.49$$
$$\mu = 15$$
$$\sigma = 4$$
$$z = 1.96$$
$$=> (15.49 - 15)/(4/\sqrt{n}) = 1.96$$
$$=> \sqrt{n} = 16$$
$$=> n = 256$$

   (b) Take a random sample of the size you found in (a) and plot the distribution. Include a screen shot of your script. You can use `numpy.random` and `matplotlib`.

   Solution:

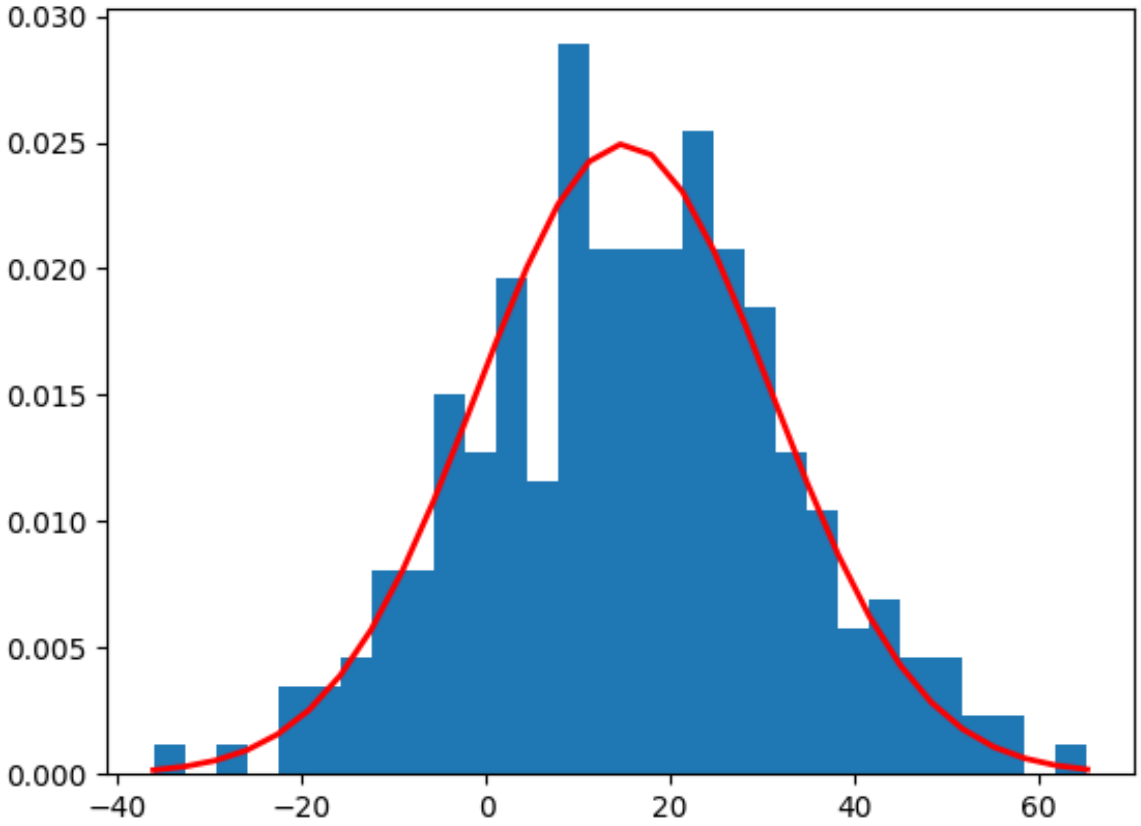

Figure 1: A boat.