HUL315 Assignment-3

Amaiya Singhal 2021CS50598

Problem 1

a) What is meant by Type I and Type II errors in the context of statistical hypothesis testing?

Solution. In hypothesis testing we can make 2 kinds of error namely **Type I** and **Type II** errors associated with the decisions we make based on the statistical tests.
Type I Error: Rejecting the null hypothesis when it was true.
Type II Error: Failing to reject the null hypothesis when it was false.

> **Example.** Consider you decide to get tested for COVID-19. The null hypothesis in this scenario is that you **do not** have COVID. If the test returns positive but you do not have COVID then this is an example of Type I error. And if the test returns negative but you do actually have COVID, then this is an example of Type II error.

b) The recommended daily dietary allowance for zinc among males older than 50 years is 15 mg/day. The article "Nutrient Intakes and Dietary Patterns of Older Americans: A National Study" reports the following summary data on intake for a sample of males aged 65-74 years: $n = 115$, $\bar{x} = 11.3$, and $s = 6.43$. Does this data indicate that average daily zinc intake in the population of all males aged 65-74 falls below the recommended allowance?

Solution. We first define the Null Hypothesis and the Alternate Hypothesis.

- Let μ = the average daily Zinc intake among males older than 50 years.
- **Null Hypothesis (H_0)**: $\mu = 15$ mg/day
- **Alternate Hypothesis (H_1)**: $\mu < 15$ mg/day

Given data is

- Sample Size (n) = 115
- Sample Mean ($\bar{x}$) = 11.3
- Sample Standard Deviation (s) = 6.43

Now we calculate the value of the t statistic using the formula:

$$t = \frac{\bar{x} - \mu}{s/\sqrt{n}}$$

$$t = \frac{11.3 - 15}{6.43/\sqrt{115}}$$

$$t = -6.17$$

We know that under the null hypotheses, the random variable

$$T = \sqrt{n}(\bar{Y} - \mu_0)/S$$

follows the t distribution with degrees of freedom $= n - 1 = 114$ i.e. $T \sim t_{n-1}$

Taking the significance level (α) = 5% or 0.05, we now calculate the p-value of the test using a `python` script.

> **p-value:** Given a value of t, the largest significance level at which we could carry out the test and still fail to reject the null hypothesis.

The **p-value** for $t = -6.171$ comes out to be around 5.338×10^{-9}.

Since the p-value is significantly smaller than our chosen significance level of 0.05, we can reject the null hypothesis, H_0. Hence, we conclude that the average daily zinc in the population of all males ages 65-74 falls below the recommended allowance.

Problem 2

Have you ever been frustrated because you could not get a container of some sort to release the last bit of its contents? The article "Shake, Rattle and Squeeze: How Much Is Left In That Container?" *(Consumer Reports, May 2009: 8)* reported on an investigation of this issue for various consumer products. Suppose five 6 oz tubes of toothpaste of a particular brand are randomly selected and squeezed until no more toothpaste will come out. Then each tube is cut open and the amount remaining is weighted, resulting in the following data (consistent with what the cited article reported): 0.53, 0.65, 0.46, 0.50, 0.37. Does it appear that the true average amount left is less than 10% of the advertised net contents?

a) Check the validity of any assumptions necessary for testing the appropriate hypothesis.

Solution. We first define the Null Hypothesis and the Alternate Hypothesis.

- Let μ = the average amount left in the toothpase tubes
- **Null Hypothesis (H_0):** $\mu = 0.6$ oz
- **Alternate Hypothesis (H_1):** $\mu < 0.6$ oz

The necessary assumptions for performing hypothesis testing on the above hypothesis are:

1. Random Sample

- A random sample ensures that the chosen sample is representative of the population of interest.
- Since the tubes were randomly selected, this assumption holds in the current scenario.

2. Independent Observations

- This assumption is necessary to ensure that measurement of one tube does not influence another.
- The assumption holds because each tube was squeezed until no more toothpaste came out and there is no reason for the squeezing of one tube to influence another.

3. Normal Distribution

- The assumption that the sample is drawn from a normal distribution is necessary for statistical tests such as the t-test which we use to test the hypothesis.
- Since the sample is very small (5 tubes), we cannot use the Central Limit Theorem to claim that the sample mean would follow a normal distribution. Hence, we perform statistical tests to check this condition.

The Shapiro Wilk Test: A statistical test used to assess whether a given sample comes from a normally distributed population. The null hypothesis assumes that the sample comes from a normally distributed population.
The test statistic, denoted as W, is calculated based on the ordered sample values. The closer W is to 1, the more likely it is that the sample comes from a normal distribution. A significant result (small p-value) indicates evidence against normality, suggesting that the data may not follow a normal distribution.

Although there are various methods for normality testing but for small sample size ($n < 50$), Shapiro–Wilk test should be used as it has more power to detect the nonnormality and this is the most popular and widely used method. Refer this paper on Normality testing for more details.

```python
from scipy.stats import shapiro

data = [0.53, 0.65, 0.46, 0.50, 0.37]
statistic, p_value = shapiro(data)
print(f"Shapiro-Wilk test: Statistic={statistic}, p-value={p_value}")
alpha = 0.05
if p_value < alpha:
    print("The data does not appear to come from a normal distribution.")
else:
    print("There is no significant evidence to reject the null hypothesis that the data comes from a normal distribution.")
```

```
Shapiro-Wilk test: Statistic=0.9839217662811279, p-value=0.9544451236724854
There is no significant evidence to reject the null hypothesis that the data comes from a normal distribution.
```

Figure 1: Testing Normality

Since the p-value is very close to 1, we can confidently assume that the sample is drawn from a normal distribution, hence, this assumption holds as well.

b) Carry out a test of the appropriate hypothesis using a significance level of 0.05. Would your conclusion change if a significance level of 0.01 had been used?

Solution. We first calculate the necessary information required for performing the hypothesis testing.

- Sample mean $(\bar{x}) = (0.53 + 0.65 + 0.46 + 0.50 + 0.37)/5 = 0.502$
- Sample standard deviation (s)

$$s = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}} = \sqrt{\left((0.53 - 0.502)^2 + ... + (0.37 - 0.502)^2\right)/4} = 0.102$$

Now we calculate the value of the t statistic using the formula

$$t = \frac{\bar{x} - \mu}{s/\sqrt{n}}$$

$$t = \frac{0.502 - 0.6}{0.102/\sqrt{5}} = -2.15$$

We know that under the null hypotheses the random variable

$$T = \sqrt{n}(\bar{Y} - \mu_0)/S$$

follows the t distribution with degrees of freedom $= n - 1 = 4$ i.e. $T \sim t_{n-1}$. The value of p comes out to be around 0.0494.

- For significance level (α) of 0.05, since $\alpha > p$, hence we reject the hypothesis. The true average amount left is less than 10% of the advertised net contents.
- For significance level (α) of 0.01, since $\alpha < p$, hence we do not reject the hypothesis. The true average amount left is equal to 10% of the advertised net contents.