Statistics Advance Assignment - 3

January 17, 2024

0.1 Q1: What is Estimation Statistics? Explain point estimate and interval estimate.

Estimation statistics is a branch of inferential statistics that deals with the estimation of population parameters based on sample statistics. It involves using sample data to make inferences about the population.

There are two main types of estimation: point estimation and interval estimation.

Point estimation involves using a single value, called the point estimate, to estimate the population parameter of interest. The point estimate is usually calculated from the sample data and is used as a best guess for the true value of the parameter. Common point estimates include the sample mean, sample proportion, and sample variance.

Interval estimation involves using a range of values, called the confidence interval, to estimate the population parameter of interest. A confidence interval is constructed from the sample data and is used to provide a range of values within which the true value of the population parameter is likely to fall. The confidence interval is usually expressed as a range of values with an associated level of confidence. For example, a 95% confidence interval for the population mean would be a range of values within which we are 95% confident that the true value of the population mean falls.

In general, interval estimation provides more information about the population parameter of interest than point estimation, as it provides a range of plausible values for the parameter rather than just a single estimate. However, interval estimation is generally more complex and computationally intensive than point estimation.

0.2 Q2. Write a Python function to estimate the population mean using a sample mean and standard deviation.

[1]:
```
def estimate_population_mean(sample_mean, sample_std, sample_size):
    """
    Calculates an estimate of the population mean given the sample mean,␣
 ↪standard deviation, and sample size.
    """
    standard_error = sample_std / (sample_size ** 0.5)
    z_score = 1.96 # z-score for 95% confidence interval
    margin_of_error = z_score * standard_error
    lower_bound = sample_mean - margin_of_error
    upper_bound = sample_mean + margin_of_error
    return (lower_bound, upper_bound)
```

0.3 Q3: What is Hypothesis testing? Why is it used? State the importance of Hypothesis testing.

Hypothesis testing is a statistical method used to determine if there is enough evidence to reject or fail to reject a proposed explanation or hypothesis about a population based on a sample of data. It involves setting up a null hypothesis, which assumes that there is no significant difference between the observed data and the expected population parameter, and an alternative hypothesis, which assumes that there is a significant difference between the observed data and the expected population parameter. The null hypothesis is tested using a statistical test, and the results are used to determine whether the null hypothesis can be rejected in favor of the alternative hypothesis.

Hypothesis testing is used in many areas of research and business to make decisions based on data. For example, a pharmaceutical company might use hypothesis testing to determine whether a new drug is more effective than an existing drug, or a marketing company might use hypothesis testing to determine whether a new advertising campaign is more effective than an old one.

The importance of hypothesis testing lies in its ability to provide a structured and objective way to evaluate evidence and draw conclusions based on data. By using hypothesis testing, researchers and analysts can avoid making decisions based on anecdotal evidence or personal biases and can make more informed decisions based on rigorous statistical analysis. Hypothesis testing also provides a framework for making decisions based on uncertainty, which is essential in many areas of research and business where decisions must be made based on incomplete or imperfect information.

0.4 Q4. Create a hypothesis that states whether the average weight of male college students is greater than the average weight of female college students.

Hypothesis: The average weight of male college students is greater than the average weight of female college students.

Null Hypothesis (H0): The average weight of male college students is not greater than the average weight of female college students. Alternative Hypothesis (HA): The average weight of male college students is greater than the average weight of female college students.

We can test this hypothesis by collecting a random sample of male and female college students and calculating the mean weight for each group. We can then use a statistical test, such as a t-test, to determine whether the difference in mean weight between the two groups is statistically significant. If the p-value is less than the significance level (e.g. 0.05), we can reject the null hypothesis and conclude that the average weight of male college students is indeed greater than the average weight of female college students. If the p-value is greater than the significance level, we fail to reject the null hypothesis and conclude that there is not enough evidence to support the claim that the average weight of male college students is greater than the average weight of female college students.

0.5 Q5. Write a Python script to conduct a hypothesis test on the difference between two population means, given a sample from each population.

[2]:
```
import numpy as np
from scipy.stats import ttest_ind

# Generate two random samples from two normal distributions
```

```
np.random.seed(123)
sample1 = np.random.normal(loc=10, scale=2, size=50)
sample2 = np.random.normal(loc=12, scale=2, size=50)

# Calculate the sample means and standard deviations
mean1, mean2 = np.mean(sample1), np.mean(sample2)
std1, std2 = np.std(sample1, ddof=1), np.std(sample2, ddof=1)

# Calculate the pooled standard deviation
pooled_std = np.sqrt(((len(sample1)-1)*(std1**2) + (len(sample2)-1)*(std2**2)) /
  ↪ (len(sample1) + len(sample2) - 2))

# Calculate the t-statistic and p-value
t_stat, p_val = ttest_ind(sample1, sample2)

# Print the results
print("Sample 1 Mean: {:.2f}".format(mean1))
print("Sample 2 Mean: {:.2f}".format(mean2))
print("Pooled Standard Deviation: {:.2f}".format(pooled_std))
print("T-Statistic: {:.2f}".format(t_stat))
print("P-Value: {:.4f}".format(p_val))
```

```
Sample 1 Mean: 10.03
Sample 2 Mean: 12.08
Pooled Standard Deviation: 2.28
T-Statistic: -4.51
P-Value: 0.0000
```

0.6 Q6: What is a null and alternative hypothesis? Give some examples.

In hypothesis testing, a null hypothesis (H0) is a statement that assumes there is no difference or no relationship between the variables under investigation. An alternative hypothesis (Ha) is a statement that contradicts the null hypothesis, suggesting that there is a difference or a relationship between the variables.

Here are some examples:

Null hypothesis: The average weight loss of participants in a weight loss program is not different from zero. Alternative hypothesis: The average weight loss of participants in a weight loss program is greater than zero.

Null hypothesis: The mean IQ scores of two groups of students are the same. Alternative hypothesis: The mean IQ scores of the two groups of students are different.

Null hypothesis: The rate of defective products produced by a machine is within the acceptable range. Alternative hypothesis: The rate of defective products produced by a machine is outside the acceptable range.

Null hypothesis: There is no correlation between exercise and blood pressure. Alternative hypothesis: There is a correlation between exercise and blood pressure.

0.7 Q7: Write down the steps involved in hypothesis testing.

Hypothesis testing involves the following steps:

State the null hypothesis (H0) and alternative hypothesis (Ha): The null hypothesis assumes that there is no significant difference between the two groups being compared, while the alternative hypothesis assumes that there is a significant difference.

Determine the level of significance ( ): This represents the maximum probability of rejecting the null hypothesis when it is actually true. Typically, the level of significance is set at 0.05 or 0.01.

Select an appropriate test statistic: The choice of test statistic depends on the nature of the hypothesis being tested, the type of data being analyzed, and the sample size.

Determine the critical region: This is the range of values of the test statistic that would lead to rejection of the null hypothesis at the chosen level of significance. The critical region is determined based on the assumptions of the test and the distribution of the test statistic.

Compute the test statistic: The test statistic is computed using the sample data.

Compare the test statistic to the critical region: If the test statistic falls within the critical region, the null hypothesis is rejected in favor of the alternative hypothesis.

Draw conclusions: Based on the results of the hypothesis test, conclusions can be drawn about the difference or relationship between the groups being compared.

0.8 Q8. Define p-value and explain its significance in hypothesis testing.

In statistical hypothesis testing, the p-value is the probability of observing a test statistic at least as extreme as the one computed from the sample, assuming that the null hypothesis is true. In simpler terms, the p-value is the probability of obtaining a result as extreme as, or more extreme than, the observed result, assuming that the null hypothesis is true.

If the p-value is less than or equal to a chosen significance level (usually 0.05), then we reject the null hypothesis in favor of the alternative hypothesis. This means that the observed result is statistically significant and not just due to chance. On the other hand, if the p-value is greater than the significance level, we fail to reject the null hypothesis. This means that the observed result is not statistically significant, and we do not have enough evidence to reject the null hypothesis.

In summary, the p-value is an important concept in hypothesis testing as it helps us determine the statistical significance of our results and whether they are likely due to chance or not.

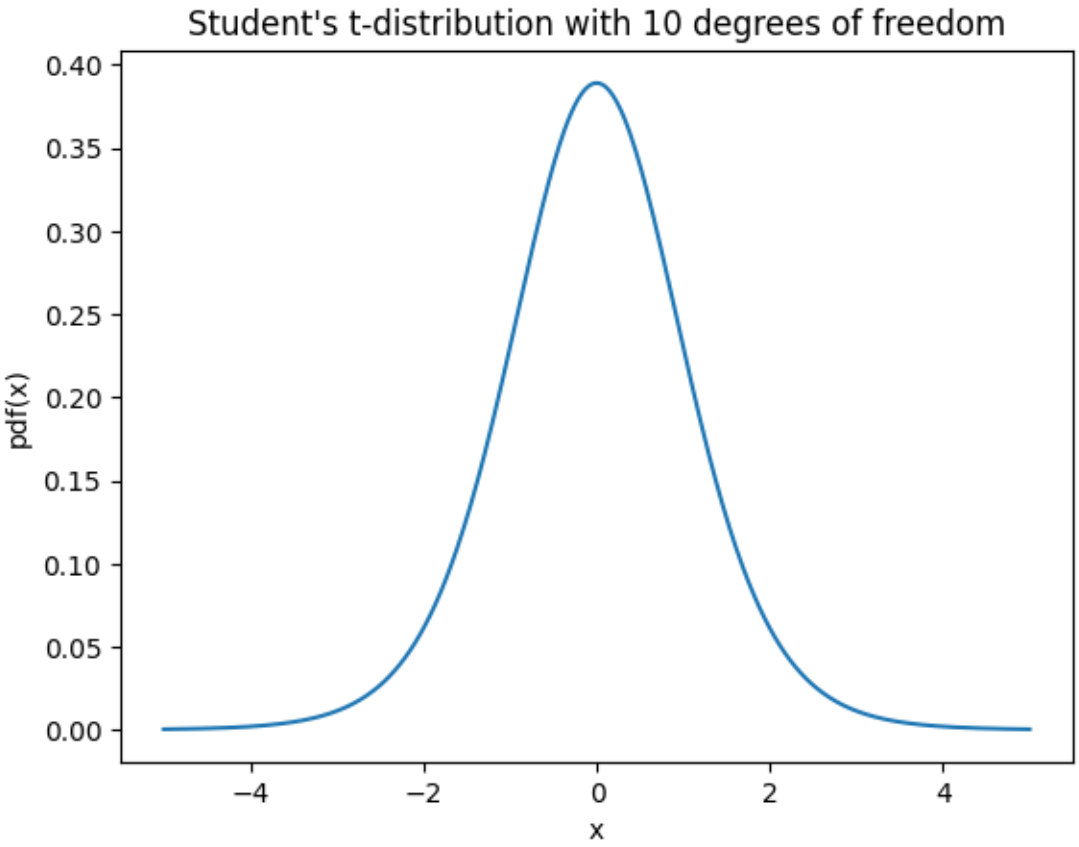
0.9 Q9. Generate a Student's t-distribution plot using Python's matplotlib library, with the degrees of freedom parameter set to 10.

[3]:
```
import numpy as np
import matplotlib.pyplot as plt
from scipy.stats import t

df = 10  # degrees of freedom

x = np.linspace(-5, 5, 500)
y = t.pdf(x, df)
```

```
plt.plot(x, y)
plt.title("Student's t-distribution with 10 degrees of freedom")
plt.xlabel('x')
plt.ylabel('pdf(x)')
plt.show()
```



0.10 Q10. Write a Python program to calculate the two-sample t-test for independent samples, given two random samples of equal size and a null hypothesis that the population means are equal.

[4]:
```
import numpy as np
from scipy.stats import ttest_ind

# Generate two random samples of equal size
sample1 = np.random.normal(loc=10, scale=2, size=100)
sample2 = np.random.normal(loc=12, scale=2, size=100)

# Calculate the t-test
```

```
t_stat, p_value = ttest_ind(sample1, sample2)

# Print the results
print("t-statistic:", t_stat)
print("p-value:", p_value)
```

```
t-statistic: -6.706602317971851
p-value: 2.0365144704730256e-10
```

0.11 Q11: What is Student's t distribution? When to use the t-Distribution.

Student's t-distribution is a probability distribution that is used to estimate the population parameters when the sample size is small and the population standard deviation is unknown. It is used in situations where the sample size is less than 30 and the population standard deviation is not known.

The t-distribution is similar to the standard normal distribution, but it has heavier tails, meaning that it is more spread out and has more variability. The t-distribution is characterized by a parameter known as degrees of freedom, which is equal to the sample size minus one.

The t-distribution is used in hypothesis testing, specifically when the sample size is small and the population standard deviation is not known. It is also used to calculate confidence intervals for population parameters, such as the mean or the difference between two means.

0.12 Q12: What is t-statistic? State the formula for t-statistic.

The t-statistic is a measure used to determine if the means of two groups are significantly different from each other when the sample size is small or the population standard deviation is unknown. It is the ratio of the difference between the means of the two groups and the standard error of the difference.

The formula for t-statistic is:

$t = (\bar{x}1 - \bar{x}2) / (s\sqrt{(1/n1 + 1/n2)})$

where $\bar{x}1$ and $\bar{x}2$ are the sample means of the two groups, s is the pooled standard deviation of the two groups, n1 and n2 are the sample sizes of the two groups.

0.13 Q13. A coffee shop owner wants to estimate the average daily revenue for their shop. They take a random sample of 50 days and find the sample mean revenue to be $500 with a standard deviation of $50. Estimate the population mean revenue with a 95% confidence interval.

To estimate the population mean revenue with a 95% confidence interval, we can use the following formula:

Confidence Interval = sample mean ± margin of error where, sample mean = $500 margin of error = t-value * standard error t-value = the critical value from the t-distribution table for the given confidence level and degrees of freedom (df = n-1) standard error = standard deviation / sqrt(n)

We can assume a 95% confidence level and the degrees of freedom (df) is 49 since the sample size is 50.

Using the t-distribution table, we find the t-value for 95% confidence level and df = 49 to be 2.009.

Now, we can calculate the margin of error as follows: standard error = 50 / sqrt(50) = 7.071 margin of error = 2.009 * 7.071 = 14.196

Therefore, the 95% confidence interval for the population mean revenue is: $500 \pm 14.196 = (485.804, 514.196)$

So, we can say with 95% confidence that the true population mean revenue lies between $485.804 and $514.196.

0.14 Q14. A researcher hypothesizes that a new drug will decrease blood pressure by 10 mmHg. They conduct a clinical trial with 100 patients and find that the sample mean decrease in blood pressure is 8 mmHg with a standard deviation of 3 mmHg. Test the hypothesis with a significance level of 0.05.

To test the hypothesis, we can use a one-sample t-test with the null hypothesis stating that the true mean decrease in blood pressure is equal to 10 mmHg, and the alternative hypothesis stating that it is less than 10 mmHg. The test statistic is calculated as follows:

t = (sample mean - hypothesized mean) / (standard error) where the standard error is calculated as the sample standard deviation divided by the square root of the sample size.

So, in this case:

hypothesized mean = 10 mmHg sample mean = 8 mmHg sample standard deviation = 3 mmHg sample size = 100

The standard error is therefore:

standard error = 3 / sqrt(100) = 0.3

And the test statistic is:

t = (8 - 10) / 0.3 = -6.67

We can then compare this to the critical value from a t-distribution with 99 degrees of freedom (since we have 100 observations and are estimating one parameter), and a significance level of 0.05. Using a one-tailed test, the critical value is -1.66.

Since our test statistic is less than the critical value, we can reject the null hypothesis and conclude that there is evidence to support the claim that the new drug decreases blood pressure by less than 10 mmHg.

0.15 Q15. An electronics company produces a certain type of product with a mean weight of 5 pounds and a standard deviation of 0.5 pounds. A random sample of 25 products is taken, and the sample mean weight is found to be 4.8 pounds. Test the hypothesis that the true mean weight of the products is less than 5 pounds with a significance level of 0.01.

Given:

Mean weight of the product = 5 pounds Standard deviation = 0.5 pounds Sample size (n) = 25 Sample mean weight = 4.8 pounds Significance level (alpha) = 0.01 We need to test the hypothesis that the true mean weight of the products is less than 5 pounds.

Null hypothesis: The true mean weight of the products is greater than or equal to 5 pounds. Alternate hypothesis: The true mean weight of the products is less than 5 pounds.

We can use a one-tailed t-test to test this hypothesis. The test statistic is calculated as follows:

t = (sample_mean - population_mean) / (standard_error / sqrt(n))

where standard_error = standard_deviation / sqrt(n)

The critical t-value is obtained from the t-distribution table with n-1 degrees of freedom and alpha level of 0.01.

Using the above formulas, we get:

standard_error = 0.5 / sqrt(25) = 0.1

t = (4.8 - 5) / (0.1) = -2

The critical t-value with 24 degrees of freedom and alpha level of 0.01 is -2.492.

Since the calculated t-value (-2) is less than the critical t-value (-2.492), we reject the null hypothesis and conclude that the true mean weight of the products is less than 5 pounds at a significance level of 0.01.

1 Q16. Two groups of students are given different study materials to prepare for a test. The first group (n1 =30) has a mean score of 80 with a standard deviation of 10, and the second group (n2 = 40) has a mean score of 75 with a standard deviation of 8.Test the hypothesis that the population means for the two groups are equal with a significance level of 0.01.

To test the hypothesis that the population means for the two groups are equal, we can use a two-sample t-test for independent samples. The null and alternative hypotheses are:

Null hypothesis: 1 = 2 Alternative hypothesis: 1 2

We will use a significance level of 0.01.

First, we can calculate the pooled standard deviation:

Sp = sqrt(((n1 - 1) * s1^2 + (n2 - 1) * s2^2) / (n1 + n2 - 2))

where s1 and s2 are the sample standard deviations.

Then, we can calculate the t-statistic:

t = (x1 - x2) / (Sp * sqrt(1/n1 + 1/n2))

where x1 and x2 are the sample means.

Using a t-table or a t-distribution calculator with (n1 + n2 - 2) degrees of freedom and a significance level of 0.01, we find the critical values to be -2.583 and 2.583.

If the calculated t-statistic falls outside this range, we reject the null hypothesis and conclude that the population means are different. Otherwise, we fail to reject the null hypothesis.

Plugging in the values, we get:

Sp = sqrt(((30 - 1) * 10^2 + (40 - 1) * 8^2) / (30 + 40 - 2)) = 8.671

t = (80 - 75) / (8.671 * sqrt(1/30 + 1/40)) = 2.67

Since 2.67 > 2.583, we reject the null hypothesis and conclude that the population means are different.

1.1 Q17. A marketing company wants to estimate the average number of ads watched by viewers during a TV program. They take a random sample of 50 viewers and find that the sample mean is 4 with a standard deviation of 1.5. Estimate the population mean with a 99% confidence interval.

To estimate the population mean with a 99% confidence interval, we can use the following formula:

Confidence interval = sample mean ± margin of error

Where margin of error = critical value x standard error

The critical value can be obtained from the t-distribution table with n-1 degrees of freedom and a 99% confidence level. For a sample size of 50 and a 99% confidence level, the critical value is 2.678.

The standard error can be calculated as the standard deviation of the sample divided by the square root of the sample size. In this case, the standard error is:

standard error = 1.5 / sqrt(50) = 0.212

Therefore, the margin of error is:

margin of error = 2.678 x 0.212 = 0.568

Finally, the confidence interval is:

Confidence interval = 4 ± 0.568 = [3.432, 4.568]

So, we can say with 99% confidence that the population mean of the number of ads watched by viewers during a TV program is between 3.432 and 4.568.