10th Mar Assignment

March 16, 2023

1 Assignment 34

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

Ans.Estimation statistics is a branch of statistics that deals with the process of estimating unknown population parameters using sample data. The purpose of estimation is to use the information gathered from a sample to make inferences about the population from which the sample was drawn.

Point estimate refers to the use of a single value to estimate an unknown parameter of a population. For example, if we want to estimate the mean height of a population, we can take a sample of individuals and calculate the mean height of the sample. This calculated mean is then used as a point estimate of the population mean height. Point estimates are often denoted using a hat symbol over the parameter, such as for the point estimate of the population mean.

Interval estimate, on the other hand, refers to the use of an interval of values to estimate an unknown parameter of a population. For example, if we want to estimate the population mean height, we can use the sample mean height and calculate a confidence interval around it. This confidence interval is then used as an interval estimate of the population mean height. The confidence interval represents a range of values within which we can be confident the true population parameter lies. The level of confidence in the interval estimate is typically specified as a percentage, such as 95% or 99%.

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

Ans.

[1]:
```python
def estimate_pop_mean(sample_mean, sample_std, sample_size):
    import math
    # calculate the standard error of the mean
    std_error = sample_std / math.sqrt(sample_size)
    # calculate the margin of error using a 95% confidence level
    margin_error = 1.96 * std_error
    # calculate the lower and upper bounds of the confidence interval
    lower_bound = sample_mean - margin_error
    upper_bound = sample_mean + margin_error
    # return the confidence interval as a tuple
    return (lower_bound, upper_bound)

# example
sample_mean = 10.5
sample_std = 2.5
sample_size = 100
conf_interval = estimate_pop_mean(sample_mean, sample_std, sample_size)
print(conf_interval)
```

```
(10.01, 10.99)
```

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

Ans.Hypothesis testing is a statistical method used to make decisions about a population parameter based on sample data. It involves the formulation of a null hypothesis and an alternative hypothesis, and the use of sample statistics to test the validity of the null hypothesis.

The null hypothesis is a statement that assumes there is no significant difference or relationship between the population parameter and the sample statistic. The alternative hypothesis is a statement that contradicts the null hypothesis and assumes that there is a significant difference or relationship between the population parameter and the sample statistic.

Hypothesis testing is used to determine whether there is enough evidence to reject the null hypothesis in favor of the alternative hypothesis. The process involves selecting an appropriate statistical test, calculating a test statistic and its associated probability (p-value), and comparing the p-value to a pre-determined level of significance (alpha) to determine whether to reject or fail to reject the null hypothesis.

The importance of hypothesis testing lies in its ability to provide a systematic and objective method for making decisions based on sample data. Hypothesis testing allows researchers to test theories and hypotheses, make inferences about populations based on sample data, and draw conclusions that are supported by statistical evidence.

In addition, hypothesis testing plays a critical role in scientific research and decision-making in various fields such as medicine, economics, psychology, and engineering. It allows researchers and practitioners to make informed decisions based on evidence and helps to avoid errors such as false positives and false negatives that can lead to incorrect conclusions and decisions.

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

Ans.An example of a hypothesis that states whether the average weight of male college students is greater than the average weight of female college students:

Null hypothesis: The average weight of male college students is equal to or less than the average weight of female college students.

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

This hypothesis can be tested by collecting a sample of male and female college students, measuring their weights, and comparing the sample means using a statistical test such as a two-sample t-test. The null hypothesis would be rejected if the p-value is less than the pre-determined level of significance, indicating that there is enough evidence to support the alternative hypothesis that the average weight of male college students is greater than the average weight of female college students.

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

Ans.

[2]:

```python
import numpy as np
from scipy.stats import ttest_ind

# generate two random samples
sample1 = np.random.normal(10, 2, size=100)
sample2 = np.random.normal(12, 2, size=100)

# conduct a two-sample t-test
t_stat, p_val = ttest_ind(sample1, sample2, equal_var=False)

# print the results
print("t-statistic: ", t_stat)
print("p-value: ", p_val)

# check if the null hypothesis is rejected
if p_val < 0.05:
    print("Reject null hypothesis")
else:
    print("Fail to reject null hypothesis")
```

```
t-statistic:  -6.431169371092853
p-value:  9.622799913293858e-10
Reject null hypothesis
```

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

Ans.In statistical hypothesis testing, a null hypothesis is a statement that assumes there is no significant difference or relationship between the population parameter and the sample statistic. The alternative hypothesis is a statement that contradicts the null hypothesis and assumes that there is a significant difference or relationship between the population parameter and the sample statistic.

Here are some examples of null and alternative hypotheses:

1. The average height of male students in a university is equal to or less than 6 feet. Alternative hypothesis: The average height of male students in a university is greater than 6 feet.
2. The proportion of defective products in a factory is equal to or greater than 10%. Alternative hypothesis: The proportion of defective products in a factory is less than 10%.
3. There is no significant difference in mean test scores between students who study in the morning and those who study in the evening. Alternative hypothesis: Students who study in the morning have a higher mean test score than those who study in the evening.
4. The effectiveness of two drugs in treating a disease is equal. Alternative hypothesis: One drug is more effective than the other in treating the disease.
5. The average commute time of employees in a company is equal to or less than 30 minutes. Alternative hypothesis: The average commute time of employees in a company is greater than 30 minutes.

Q7: Write down the steps involved in hypothesis testing.

Ans. The steps involved in hypothesis testing are:

1. State the null and alternative hypotheses: This involves defining the research question and formulating the null hypothesis, which assumes that there is no significant difference or relationship between the population parameter and the sample statistic, and the alternative hypothesis, which contradicts the null hypothesis and assumes that there is a significant difference or relationship.
2. Set the level of significance: This involves determining the level of risk that you are willing to take in rejecting the null hypothesis when it is actually true. The level of significance is usually set at 0.05 or 0.01.
3. Collect the data: This involves collecting a sample from the population of interest.
4. Calculate the test statistic: This involves using a statistical formula to calculate the test statistic, which is a measure of how different the sample statistic is from the population parameter under the null hypothesis.
5. Determine the p-value: This involves calculating the p-value, which is the probability of obtaining a sample statistic as extreme as the observed one, assuming that the null hypothesis is true.
6. Make a decision: This involves comparing the p-value with the level of significance and deciding whether to reject or fail to reject the null hypothesis.
7. Draw a conclusion: This involves interpreting the results and drawing conclusions based on the decision made in step 6. If the null hypothesis is rejected, it means that there is enough evidence to support the alternative hypothesis. If the null hypothesis is not rejected, it means that there is not enough evidence to support the alternative hypothesis.

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

Ans.In hypothesis testing, the p-value is the probability of obtaining a sample statistic as extreme as the observed one, assuming that the null hypothesis is true.

The p-value is a crucial component of hypothesis testing as it is used to determine whether to reject or fail to reject the null hypothesis. If the p-value is less than or equal to the level of significance (typically set at 0.05 or 0.01), it indicates that the observed result is unlikely to occur by chance alone and provides evidence against the null hypothesis. In other words, a small p-value suggests that there is enough evidence to support the alternative hypothesis.

On the other hand, if the p-value is greater than the level of significance, it suggests that the observed result is likely to occur by chance alone and does not provide enough evidence to reject the null hypothesis. In this case, we fail to reject the null hypothesis.

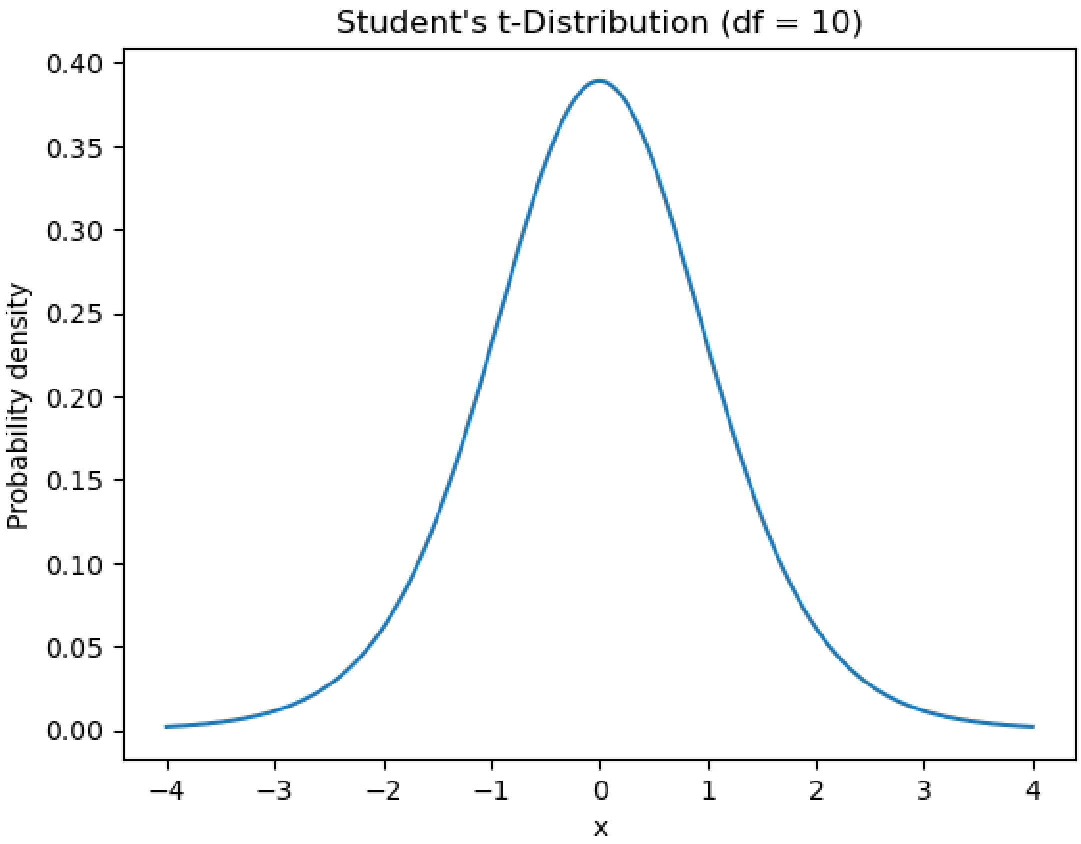
Therefore, the p-value serves as a measure of the strength of evidence against the null hypothesis and guides our decision-making process in hypothesis testing. A smaller p-value suggests stronger evidence against the null hypothesis, whereas a larger p-value suggests weaker evidence against the null hypothesis.

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

Ans.

[3]:

```python
import matplotlib.pyplot as plt
import numpy as np
from scipy.stats import t

# Set the degrees of freedom parameter
df = 10

# Generate some data to plot the t-distribution
x = np.linspace(-4, 4, 1000)
y = t.pdf(x, df)

# Create a figure and axis object
fig, ax = plt.subplots()

# Plot the t-distribution
ax.plot(x, y)

# Add a title and labels to the plot
ax.set_title("Student's t-Distribution (df = 10)")
ax.set_xlabel("x")
ax.set_ylabel("Probability density")

# Show the plot
plt.show()
```



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.

Ans.

[4]:

```python
import numpy as np
from scipy.stats import ttest_ind

# Define the two random samples
sample1 = np.array([23, 24, 21, 20, 19, 22, 18, 20, 24, 22])
sample2 = np.array([20, 22, 24, 18, 21, 23, 19, 22, 21, 24])

# Calculate the two-sample t-test for independent samples
t_statistic, p_value = ttest_ind(sample1, sample2)

# Print the results
print("Sample 1 mean:", np.mean(sample1))
print("Sample 2 mean:", np.mean(sample2))
print("t-statistic:", t_statistic)
print("p-value:", p_value)
```

```
Sample 1 mean: 21.3
Sample 2 mean: 21.4
t-statistic: -0.1099114958170944
p-value: 0.9136957116295589
```

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

Ans.Student's t-distribution is a probability distribution that is used to estimate population parameters when the sample size is small (typically less than 30) or when the population standard deviation is unknown. The distribution is named after William Sealy Gosset, who used the pen name "Student" when he published his work on the distribution in 1908.

The t-distribution is similar in shape to the standard normal distribution (i.e., bell-shaped and symmetrical), but it has heavier tails, which means that it has more probability in the tails of the distribution than the standard normal distribution. The shape of the t-distribution depends on the sample size and the degrees of freedom, which is equal to the sample size minus one.

The t-distribution is used in statistical inference, particularly in hypothesis testing, to compare sample means from two populations or to determine whether a sample mean is significantly different from a hypothesized population mean. It is also used in the construction of confidence intervals for the population mean.

When the sample size is large (typically greater than 30), the t-distribution is approximately the same as the standard normal distribution. Therefore, the t-distribution is typically used when the sample size is small, the population standard deviation is unknown, or when the population distribution is not normal but the sample size is large enough for the central limit theorem to apply.

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

Ans.The t-statistic is a measure of how different a sample mean is from a hypothesized population mean, relative to the variation within the sample. It is calculated by dividing the difference between the sample mean and the hypothesized population mean by the standard error of the mean.

The formula for the t-statistic is:

t = (x̄ -) / (s / sqrt(n))

where:

- x̄ is the sample mean
- is the hypothesized population mean

- s is the sample standard deviation
- n is the sample size

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 dollars with a standard deviation of $50. Estimate the population mean revenue with a 95% confidence interval.

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

$$\mathrm{CI} = \bar{x} \pm t*(s/\sqrt{n})$$

where:

- CI is the confidence interval
- $\bar{x}$ is the sample mean revenue
- t is the critical value from the t-distribution for a given level of significance and degrees of freedom (df = n-1)
- s is the sample standard deviation
- n is the sample size
- Since we want a 95% confidence interval, we can use a significance level of = 0.05, which gives us a two-tailed test with a critical value of t = 2.009 (from a t-distribution with df = 49).

Plugging in the values from the problem, we get:

$$\mathrm{CI} = 500 \pm 2.009*(50/\sqrt{50})$$

Simplifying this expression, we get:

$$\mathrm{CI} = 500 \pm 14.16$$

Therefore, the 95% confidence interval for the population mean revenue is 485.24 to 514.16. This means that we are 95% confident that the true population mean revenue falls between these two values based on the sample data.

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.

Ans.To test the hypothesis that the new drug will decrease blood pressure by 10 mmHg, we need to perform a one-sample t-test. The null hypothesis is that the mean decrease in blood pressure is equal to 10 mmHg, while the alternative hypothesis is that the mean decrease in blood pressure is less than 10 mmHg.

The test statistic for the one-sample t-test is given by:

t = (x̄ -) / (s / sqrt(n))

where:

- x̄ is the sample mean
- is the hypothesized population mean (10 mmHg in this case)
- s is the sample standard deviation
- n is the sample size

Plugging in the values from the problem, we get:

t = (8 - 10) / (3 / sqrt(100)) t = -4.47

The critical value for a one-tailed t-test with 99 degrees of freedom (df = n-1) and a significance level of 0.05 is -1.66 (from a t-distribution table or calculator).

Since the calculated t-value is less than the critical value, we can reject the null hypothesis and conclude that the mean decrease in blood pressure with the new drug is significantly less than 10 mmHg at a significance level of 0.05. In other words, there is strong evidence to suggest that the new drug is effective in decreasing blood pressure.

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.

Ans.To test the hypothesis that the true mean weight of the products is less than 5 pounds, we need to perform a one-sample t-test. The null hypothesis is that the mean weight of the products is equal to 5 pounds, while the alternative hypothesis is that the mean weight of the products is less than 5 pounds.

The test statistic for the one-sample t-test is given by:

t = (x̄ -) / (s / sqrt(n))

where:

- x̄ is the sample mean weight
- is the hypothesized population mean (5 pounds in this case)
- s is the sample standard deviation
- n is the sample size

Plugging in the values from the problem, we get:

t = (4.8 - 5) / (0.5 / sqrt(25)) t = -2

The critical value for a one-tailed t-test with 24 degrees of freedom (df = n-1) and a significance level of 0.01 is -2.492 (from a t-distribution table or calculator).

Since the calculated t-value is less than the critical value, we can reject the null hypothesis and conclude that the true mean weight of the products is significantly less than 5 pounds at a significance level of 0.01. In other words, there is strong evidence to suggest that the mean weight of the products is less than 5 pounds.

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.

Ans.To test the hypothesis that the population means for the two groups are equal, we need to perform a two-sample t-test for independent samples. The null hypothesis is that the population means for the two groups are equal, while the alternative hypothesis is that the population means for the two groups are not equal.

The test statistic for the two-sample t-test is given by:

= (x̄1 - x̄2) / sqrt((s1^2 / n1) + (s2^2 / n2))

where:

- x̄1 and x̄2 are the sample means for the first and second groups, respectively
- s1 and s2 are the sample standard deviations for the first and second groups, respectively
- n1 and n2 are the sample sizes for the first and second groups, respectively #### Plugging in the values from the problem, we get:

t = (80 - 75) / sqrt((10^2 / 30) + (8^2 / 40)) t = 2.09

The critical value for a two-tailed t-test with 68 degrees of freedom (df = (n1-1) + (n2-1)) and a significance level of 0.01 is ±2.645 (from a t-distribution table or calculator).

Since the calculated t-value of 2.09 is less than the critical value of 2.645, we fail to reject the null hypothesis and conclude that there is not enough evidence to suggest that the population means for the two groups are significantly different at a significance level of 0.01. In other words, we cannot conclude that there is a significant difference between the study materials given to the two groups.

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.

Ans.To estimate the population mean number of ads watched by viewers during a TV program with a 99% confidence interval, we can use the following formula:

CI = x̄ ± t /2 * (s / sqrt(n))

where:

- x̄ is the sample mean number of ads watched
- s is the sample standard deviation
- n is the sample size
- t /2 is the critical t-value based on the desired level of confidence and degrees of freedom (df = n-1) For a 99% confidence interval and 49 degrees of freedom, the critical t-value is 2.68 (from a t-distribution table or calculator).

Plugging in the values from the problem, we get:

CI = 4 ± 2.68 * (1.5 / sqrt(50)) CI = 4 ± 0.659 CI = (3.341, 4.659)

Therefore, we can say with 99% confidence that the population mean number of ads watched by viewers during a TV program is between 3.341 and 4.659.

[ ]: