statistics-advance-4

August 13, 2023

[ ]:

```
# Q1: What is the difference between a t-test and a z-test? Provide an example scenario where you would use each type of test
```

Use Case and Sample Size:

t-test: The t-test is used when the sample size is small (typically less than 30) and the population standard deviation is unknown.

z-test: The z-test is used when the sample size is relatively large (typically larger than 30) and the population standard deviation is known or when the sample size is large enough for the Central Limit Theorem to apply. It relies on the standard normal (z) distribution.

Calculation of Critical Values:

t-test: The critical values for t-tests depend on the degrees of freedom and the desired significance level. They can vary with different sample sizes.

z-test: The critical values for z-tests are fixed and determined by the chosen significance level (alpha) because the standard normal distribution's properties are constant.

Example Scenarios:

t-test: Suppose you want to compare the average scores of two groups of students who received different teaching methods. You have a small sample size for each group (e.g., 20 students in each group), and you don't know the population standard deviation. In this case, you would use a two-sample t-test for independent samples.

z-test: Imagine you are conducting a survey to estimate the proportion of people in a city who support a particular policy. You have a large sample size (e.g., 500 respondents), and you know the population standard deviation. Since you have a large sample size and a known population standard deviation, you can use a z-test to test your hypothesis about the population proportion.

[ ]:

```
# Q2: Differentiate between one-tailed and two-tailed tests
```

One-Tailed Test:

A one-tailed test, also known as a one-sided test, is a type of hypothesis test that focuses on a specific direction of effect or difference. It is used to determine whether the observed sample data provides evidence in favor of a particular alternative hypothesis in only one direction. The critical region for a one-tailed test is located on one side of the distribution curve (either the right or the left), depending on the research question.

Two-Tailed Test:

A two-tailed test, also known as a two-sided test, is a type of hypothesis test that considers the possibility of an effect or difference in either direction from the null hypothesis. It is used to determine whether the observed sample data provides evidence of any significant difference or effect, regardless of the direction. The critical region for a two-tailed test is divided between both tails of the distribution curve.

In a two-tailed test, the null hypothesis is rejected if the sample statistic falls into either of the two critical regions (both tails). This indicates that the observed data provides evidence of a significant effect or difference, regardless of whether it is positive or negative.

[ ]:

```
# Q3: Explain the concept of Type 1 and Type 2 errors in hypothesis testing.␣
 ↪Provide an example scenario for each type of error
```

Type 1 Error (False Positive):

Definition: A Type 1 error occurs when the null hypothesis is incorrectly rejected when it is actually true.

Example Scenario: Suppose a medical researcher is testing a new drug to see if it's effective at reducing blood pressure. The null hypothesis states that the drug has no effect on blood pressure. If the researcher incorrectly rejects the null hypothesis and concludes that the drug is effective when it's actually not, that's a Type 1 error. Patients might be prescribed a drug that doesn't have the intended effect.

Type 2 Error (False Negative):

Definition: A Type 2 error occurs when the null hypothesis is incorrectly failed to be rejected when it is actually false.

Example Scenario: Continuing the medical example, if the new drug is actually effective at reducing blood pressure, but the researcher fails to reject the null hypothesis and concludes the drug has no effect, that's a Type 2 error. Patients might miss out on a beneficial treatment.

[ ]:

```
# Q4:  Explain Bayes's theorem with an example
```

Bayes' theorem describes the probability of occurrence of an event related to any condition. It is also considered for the case of conditional probability. Bayes theorem is also known as the formula for the probability of "causes". For example: if we have to calculate the probability of taking a blue ball from the second bag out of three different bags of balls, where each bag contains three different colour balls viz. red, blue, black. In this case, the probability of occurrence of an event is calculated depending on other conditions is known as conditional probability.

Suppose you have two bowls of cookies. Bowl A contains 30 chocolate cookies and 10 vanilla cookies, while Bowl B contains 20 of each. You randomly choose one of the bowls, and then you randomly select a cookie from that bowl. If you pick a chocolate cookie, what's the probability that you picked it from Bowl A?

Let: A be the event that you picked a cookie from Bowl A. B be the event that you picked a chocolate cookie.

[1]:

```
# Given probabilities
p_bowl_a = 1/2  # Probability of picking Bowl A
```

```
p_bowl_b = 1/2  # Probability of picking Bowl B

p_chocolate_given_bowl_a = 30 / (30 + 10)  # Probability of picking a chocolate␣
  ↪cookie given Bowl A
p_chocolate_given_bowl_b = 20 / (20 + 20)  # Probability of picking a chocolate␣
  ↪cookie given Bowl B

# Calculate the probability of picking a chocolate cookie
p_chocolate = (p_chocolate_given_bowl_a * p_bowl_a) + (p_chocolate_given_bowl_b␣
  ↪* p_bowl_b)

# Calculate the probability of picking from Bowl A given a chocolate cookie␣
  ↪using Bayes's Theorem
p_bowl_a_given_chocolate = (p_chocolate_given_bowl_a * p_bowl_a) / p_chocolate

# Print the result
print("Probability of picking from Bowl A given a chocolate cookie:",␣
  ↪p_bowl_a_given_chocolate)
```

```
Probability of picking from Bowl A given a chocolate cookie: 0.6
```

[ ]:

```
# Q5: What is a confidence interval? How to calculate the confidence interval,␣
  ↪explain with an example.
```

A confidence interval is a range of values that is likely to contain the true population parameter, such as a population mean or proportion, based on a sample from the population.

Confidence Interval Formulas If n 30 Confidence Interval = $\bar{x} \pm z\ /2(\ /\sqrt{}n)$ If n<30 Confidence Interval = $\bar{x} \pm t\ /2(S/\sqrt{}n)$

Where,

n = Number of terms $\bar{x}$ = Sample Mean = Standard Deviation z /2 = Value corresponding to 2 in z table t /2 = Value corresponding to 2 in t table = (1 – Confidence Level /100)

Step 1: Determine the Confidence Level Step 2: Find the Critical Value Step 3: Calculate the Standard Error Step 4: Calculate the Margin of Error Step 5: Calculate the Confidence Interval

[2]:

```
import scipy.stats as stats
import math

# Sample data
sample_mean = 3.5
sample_std = 1.2
sample_size = 50
confidence_level = 0.95

# Calculate the critical value (z-value) for the given confidence level
critical_value = stats.norm.ppf(1 - (1 - confidence_level) / 2)
```

```
# Calculate the standard error
standard_error = sample_std / math.sqrt(sample_size)

# Calculate the margin of error
margin_of_error = critical_value * standard_error

# Calculate the confidence interval
lower_bound = sample_mean - margin_of_error
upper_bound = sample_mean + margin_of_error

# Print the results
print("Critical Value:", critical_value)
print("Standard Error:", standard_error)
print("Margin of Error:", margin_of_error)
print("Confidence Interval:", (lower_bound, upper_bound))
```

```
Critical Value: 1.959963984540054
Standard Error: 0.1697056274847714
Margin of Error: 0.33261691784392267
Confidence Interval: (3.1673830821560776, 3.8326169178439224)
```

[ ]:

```
# Q6. Use Bayes' Theorem to calculate the probability of an event occurring␣
 ↪given prior knowledge of the event's probability and new evidence.

# Provide a sample problem and solution.
```

Sample Problem: Email Filtering

Suppose you are working on an email filtering system to identify whether an email is spam or not. You have historical data that suggests that 20% of all emails are spam. You also know that your email filter correctly identifies 95% of the spam emails as spam (sensitivity), and it incorrectly marks 3% of the legitimate emails as spam (false positive rate).

Given this information, what is the probability that an email flagged as spam by your filter is actually spam?

[8]:

```
# Given probabilities
P_S = 0.20  # Probability that an email is spam
P_F_given_S = 0.95  # Probability that the filter flags spam as spam␣
 ↪(sensitivity)
P_F_given_not_S = 0.03  # Probability that the filter flags non-spam as spam␣
 ↪(false positive rate)

# Calculate the complement probabilities
P_not_S = 1 - P_S  # Probability that an email is not spam
P_not_F_given_S = 1 - P_F_given_S  # Probability that the filter does not flag␣
 ↪spam as spam
```

```
P_not_F_given_not_S = 1 - P_F_given_not_S  # Probability that the filter does␣
  ↪not flag non-spam as spam

# Calculate the probability that the filter flags an email as spam (total␣
  ↪probability)
P_F = (P_F_given_S * P_S) + (P_F_given_not_S * P_not_S)

# Calculate the probability that an email is actually spam given it's flagged␣
  ↪as spam by the filter
P_S_given_F = (P_F_given_S * P_S) / P_F

# Print the result
print("Probability that an email flagged as spam by the filter is actually spam:
  ↪", P_S_given_F)
```

```
Probability that an email flagged as spam by the filter is actually spam:
0.8878504672897196
```

[ ]:

```
# Q7. Calculate the 95% confidence interval for a sample of data with a mean of␣
  ↪50 and a standard deviation of 5.

# Interpret the results
```

To calculate the 95% confidence interval for a sample of data, you can use the following formula:

Confidence Interval = Mean ± (Critical Value) * (Standard Deviation / √(Sample Size))

For a 95% confidence level, the critical value is approximately 1.96.

Confidence Interval = 50 ± 1.96 * (5 / √30)

Confidence Interval = 50 ± 1.772 Now calculating the upper and lower bounds of the confidence interval:

Upper Bound = 50 + 1.772   51.772 Lower Bound = 50 - 1.772   48.228

95% Confidence: This means that if we were to collect many samples and calculate a 95% confidence interval for each sample, approximately 95% of those intervals would contain the true population mean.

Lower Bound (48.228): We are 95% confident that the true population mean is at least 48.228. In other words, if we were to repeatedly sample and calculate the mean, we would expect that 95% of these calculated means would be greater than or equal to 48.228.

Upper Bound (51.772): Similarly, we are 95% confident that the true population mean is at most 51.772. In other words, if we were to repeatedly sample and calculate the mean, we would expect that 95% of these calculated means would be less than or equal to 51.772.

[ ]:

```
# Q8. What is the margin of error in a confidence interval?
# How does sample size affect the margin of error?
# Provide an example of a scenario where a larger sample size would result in a␣
↪smaller margin of error
```

The margin of error (MOE) in a confidence interval is a measure of the range within which we expect the true population parameter to lie.

MOE = Z*√n/

n= sample size

= Population Standard Deviation

z = z score

Impact of Sample Size on Margin of Error:

Sample size has an inverse relationship with the margin of error. As the sample size increases, the margin of error decreases. In other words, a larger sample size leads to a more precise estimate and a narrower confidence interval. This happens because larger samples provide more information about the population, leading to a better representation of its characteristics and less variability in the estimated statistic.

Example Scenario:

Suppose you want to estimate the average height of students in a university with a 95% confidence level. You take two different sample sizes: 50 students and 500 students. Let's say the population standard deviation of heights is known to be 4 inches.

For the 50-student sample:

Z for a 95% confidence level 1.96 (approximately) (population standard deviation) = 4 inches n (sample size) = 50

MOE 1.11

For the 500-student sample:

Z for a 95% confidence level 1.96 (same as before) (population standard deviation) = 4 inches n (sample size) = 500

MOE 0.57

In this example, the larger sample size of 500 results in a smaller margin of error (0.57 inches) compared to the smaller sample size of 50 (1.11 inches). This illustrates how a larger sample size leads to a more precise estimate with a narrower confidence interval.

```
[ ]: # Q9. Calculate the z-score for a data point with a value of 75, a population␣
     ↪mean of 70, and a population standard deviation of 5. Interpret the results.
```

The formula for calculating a z-score is z = (x-)/ , where x is the raw score, is the population mean, and is the population standard deviation.

Given the values: Data point value (x) = 75 Population mean ( ) = 70 Population standard deviation ( ) = 5

The calculated z-score is 1.

The z-score is a measure of how many standard deviations a particular data point is away from the mean of the population. In this case, a z-score of 1 indicates that the data point with a value of 75 is 1 standard deviation above the population mean of 70.

[ ]:

```python
# Q10. In a study of the effectiveness of a new weight loss drug, a sample of 50 participants lost an average of 6 pounds with a standard deviation of 2.5 pounds.
# Conduct a hypothesis test to determine if the drug is significantly effective at a 95% confidence level using a t-test.
```

[7]:

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

# Given data
sample_mean = 6
sample_std = 2.5
sample_size = 50
population_mean_null = 0  # Assumed population mean under null hypothesis
confidence_level = 0.95

# Calculate the standard error (sample standard deviation divided by the square root of the sample size)
standard_error = sample_std / np.sqrt(sample_size)

# Calculate the t-statistic
t_statistic = (sample_mean - population_mean_null) / standard_error

# Degrees of freedom
degrees_of_freedom = sample_size - 1

# Calculate the critical t-value
alpha = 1 - confidence_level
critical_t_value = stats.t.ppf(1 - alpha / 2, degrees_of_freedom)

# Compare the t-statistic with the critical t-value
if np.abs(t_statistic) > critical_t_value:
    result = "Reject the null hypothesis"
else:
    result = "Fail to reject the null hypothesis"

# Print the results
print("T-statistic:", t_statistic)
print("Critical t-value:", critical_t_value)
print("Result:", result)
```

```
T-statistic: 16.970562748477143
Critical t-value: 2.009575234489209
Result: Reject the null hypothesis
```

```
[ ]: # Q11. In a survey of 500 people, 65% reported being satisfied with their␣
      ↪current job.
     # Calculate the 95% confidence interval for the true proportion of people who␣
      ↪are satisfied with their job.
```

```
[6]: import numpy as np
     from scipy.stats import norm

     # Given data
     sample_proportion = 0.65
     confidence_level = 0.95
     sample_size = 500

     # Calculate the z-score for the desired confidence level
     z_score = norm.ppf(1 - (1 - confidence_level) / 2)

     # Calculate the standard error
     standard_error = np.sqrt((sample_proportion * (1 - sample_proportion)) /␣
      ↪sample_size)

     # Calculate the margin of error
     margin_of_error = z_score * standard_error

     # Calculate the confidence interval
     confidence_interval = (sample_proportion - margin_of_error, sample_proportion +␣
      ↪margin_of_error)

     # Print the results
     print("Confidence Interval:", confidence_interval)
```

```
Confidence Interval: (0.6081925393809212, 0.6918074606190788)
```

```
[ ]: # Q12. A researcher is testing the effectiveness of two different teaching␣
      ↪methods on student performance.
     # Sample A has a mean score of 85 with a standard deviation of 6, while sample␣
      ↪B has a mean score of 82 with a standard deviation of 5.
     # Conduct a hypothesis test to determine if the two teaching methods have a␣
      ↪significant difference in student performance using a t-test with a␣
      ↪significance level of 0.01
```

```
[5]: import numpy as np
     from scipy import stats

     # Given data
     sample_a_mean = 85
     sample_a_std = 6
     sample_a_size = 30  # Assumed sample size
```

```
sample_b_mean = 82
sample_b_std = 5
sample_b_size = 30  # Assumed sample size
significance_level = 0.01

# Calculate the pooled standard deviation
pooled_std = np.sqrt(((sample_a_std ** 2) + (sample_b_std ** 2)) / 2)

# Calculate the standard error of the difference between means
standard_error = pooled_std * np.sqrt(1/sample_a_size + 1/sample_b_size)

# Calculate the t-statistic
t_statistic = (sample_a_mean - sample_b_mean) / standard_error

# Degrees of freedom
degrees_of_freedom = sample_a_size + sample_b_size - 2

# Calculate the critical t-value
critical_t_value = stats.t.ppf(1 - significance_level / 2, degrees_of_freedom)

# Compare the t-statistic with the critical t-value
if np.abs(t_statistic) > critical_t_value:
    result = "Reject the null hypothesis"
else:
    result = "Fail to reject the null hypothesis"

# Print the results
print("T-statistic:", t_statistic)
print("Critical t-value:", critical_t_value)
print("Result:", result)
```

```
T-statistic: 2.10386061995483
Critical t-value: 2.6632869538098674
Result: Fail to reject the null hypothesis
```

[ ]:

```
# Q13. A population has a mean of 60 and a standard deviation of 8.
# A sample of 50 observations has a mean  of 65. Calculate the 90% confidence␣
 ↪interval for the true population mean.
```

[4]:

```
import numpy as np
from scipy import stats

# Given data
population_mean = 60
population_std = 8
sample_mean = 65
sample_size = 50
```

```
confidence_level = 0.90

# Calculate the standard error (sample standard deviation divided by the square root of the sample size)
standard_error = population_std / np.sqrt(sample_size)

# Calculate the critical value for the given confidence level
alpha = 1 - confidence_level
critical_value = stats.norm.ppf(1 - alpha / 2)

# Calculate the margin of error
margin_of_error = critical_value * standard_error

# Calculate the confidence interval
confidence_interval = (sample_mean - margin_of_error, sample_mean + margin_of_error)

# Print the results
print("Confidence Interval:", confidence_interval)
```

```
Confidence Interval: (63.13906055411732, 66.86093944588268)
```

```
[ ]: # Q14. In a study of the effects of caffeine on reaction time, a sample of 30 participants had an average reaction time of 0.25 seconds with a standard deviation of 0.05 seconds.

# Conduct a hypothesis test to determine if the caffeine has a significant effect on reaction time at a 90% confidence level using a t-test.
```

Hypotheses:

Null Hypothesis (H0): Caffeine does not have a significant effect on reaction time. Alternative Hypothesis (Ha): Caffeine has a significant effect on reaction time.

```
[3]: import numpy as np
from scipy import stats

# Given data
sample_mean = 0.25
sample_std = 0.05
sample_size = 30
population_mean_null = 0.23  # Assumed population mean under null hypothesis

# Degrees of freedom
degrees_of_freedom = sample_size - 1

# Calculate the t-statistic
```

```python
t_statistic = (sample_mean - population_mean_null) / (sample_std / np.
 ↪sqrt(sample_size))

# Calculate the critical t-value
confidence_level = 0.90
alpha = 1 - confidence_level
critical_t_value = stats.t.ppf(1 - alpha / 2, degrees_of_freedom)

# Compare the t-statistic with the critical t-value
if np.abs(t_statistic) > critical_t_value:
    result = "Reject the null hypothesis"
else:
    result = "Fail to reject the null hypothesis"

# Print the results
print("T-statistic:", t_statistic)
print("Critical t-value:", critical_t_value)
print("Result:", result)
```

```
T-statistic: 2.1908902300206634
Critical t-value: 1.6991270265334972
Result: Reject the null hypothesis
```