11 Mar Assignment

March 16, 2023

1 Assignment 35

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.

Ans.A t-test and a z-test are both statistical tests used to determine whether a sample mean is significantly different from a population mean. The main difference between the two tests is that a t-test is used when the sample size is small (typically less than 30), and the population standard deviation is unknown, while a z-test is used when the sample size is large (typically greater than 30), and the population standard deviation is known or assumed to be known.

For example, suppose a researcher is interested in determining whether the average height of male students in a particular university is different from the national average height of male students. If the researcher has access to the population standard deviation of male student heights in the country, and the sample size is large (e.g., more than 30 students), the researcher can use a z-test.

On the other hand, if the researcher does not know the population standard deviation, or the sample size is small (e.g., less than 30 students), the researcher would use a t-test.

Another example scenario where a t-test might be used is in a clinical trial to determine whether a new drug is effective in treating a particular condition. In this case, the sample size may be small (e.g., fewer than 30 patients in each treatment group), and the population standard deviation of the response variable (e.g., the level of pain experienced by patients) is typically unknown.

Overall, the choice between a t-test and a z-test depends on the characteristics of the data and the research question being addressed.

Q2: Differentiate between one-tailed and two-tailed tests.

Ans.In statistics, a hypothesis test is used to determine whether there is enough evidence to reject or fail to reject a null hypothesis. The null hypothesis is a statement that there is no significant difference between two groups or variables. The alternative

hypothesis, on the other hand, is the opposite of the null hypothesis, indicating that there is a significant difference between the two groups or variables.

1. One-tailed test: A one-tailed test, also known as a directional test, is a statistical test in which the alternative hypothesis specifies the direction of the difference between the two groups or variables. For example, if we want to test whether the mean height of men is greater than that of women, we would use a one-tailed test with the alternative hypothesis: "The mean height of men is greater than the mean height of women." In a one-tailed test, the critical region is located entirely on one side of the sampling distribution, and the probability of making a type I error is only in that one tail. The level of significance is usually set at 0.05, and the critical value is determined based on the direction of the alternative hypothesis.

2. Two-tailed test: A two-tailed test, also known as a nondirectional test, is a statistical test in which the alternative hypothesis does not specify the direction of the difference between the two groups or variables. For example, if we want to test whether the mean weight of a certain population is different from a specific value, we would use a two-tailed test with the alternative hypothesis: "The mean weight of the population is different from the specific value." In a two-tailed test, the critical region is located on both sides of the sampling distribution, and the probability of making a type I error is divided equally between the two tails. The level of significance is usually set at 0.05, and the critical value is determined based on the level of significance and the degrees of freedom. In summary, the main difference between one-tailed and two-tailed tests is that in a one-tailed test, the alternative hypothesis specifies the direction of the difference, while in a two-tailed test, the alternative hypothesis does not specify the direction of the difference.

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

Ans.In hypothesis testing, there are two types of errors that can occur: Type I error and Type II error.

1. Type I error: Type I error, also known as a false positive, occurs when the null hypothesis is rejected even though it is true. In other words, a Type I error occurs when a researcher concludes that there is a significant effect when, in fact, there is no significant effect. The probability of making a Type I error is denoted by the symbol alpha ( ) and is usually set at 0.05 or 0.01. Example scenario: A pharmaceutical company is conducting a clinical trial to test the efficacy of a new drug for a certain disease. The null hypothesis is that the drug has no effect on the disease, and the alternative hypothesis is that the drug is effective in treating the disease. If the researchers conclude that the drug is effective based on the data, but in reality, it has no effect on the disease, this would be a Type I error.

2. Type II error: Type II error, also known as a false negative, occurs when the null hypothesis is not rejected even though it is false. In other words, a Type II error occurs when a researcher fails to conclude that there is a significant effect when, in fact, there is a significant effect. The probability of making a Type II error is denoted by the symbol beta ( ). Example scenario: A car manufacturer is testing the safety of its new car model by conducting crash tests. The null hypothesis is that the new car model is no safer than the old model, and the alternative hypothesis is that the new car model is safer. If the researchers fail to conclude that the new car model is safer based on the data, but in reality, it is safer, this would be a Type II error.

In summary, a Type I error occurs when the null hypothesis is rejected when it is true, and a Type II error occurs when the null hypothesis is not rejected when it is false. Both types of errors can have serious consequences and need to be carefully controlled and minimized in hypothesis testing.

Q4: Explain Bayes's theorem with an example.

Ans. Bayes's theorem is a fundamental concept in probability theory that describes how to update probabilities based on new information or evidence. It is named after the British statistician Thomas Bayes.

The theorem can be stated as follows:

P(A|B) = P(B|A) x P(A) / P(B)

where:

- P(A|B) is the probability of A given B (the posterior probability)
- P(B|A) is the probability of B given A (the likelihood)
- P(A) is the prior probability of A
- P(B) is the prior probability of B #### In other words, Bayes's theorem enables us to calculate the probability of a hypothesis or event based on prior knowledge or assumptions, and new evidence or data.

Example:Suppose we have two boxes, Box A and Box B, each containing two types of fruits: apples and oranges. The proportion of apples and oranges in each box is as follows:

Box A: 30% apples and 70% oranges Box B: 70% apples and 30% oranges

We randomly select one of the boxes and draw a fruit from it. We do not know which box we have selected, but we observe that the fruit is an apple.

What is the probability that we have selected Box A?

We can use Bayes's theorem to calculate the posterior probability of selecting Box A given that we have observed an apple:

P(A|B) = P(B|A) x P(A) / P(B)

where:

- P(A|B) is the posterior probability of selecting Box A given an apple
- P(B|A) is the likelihood of selecting an apple given Box A (i.e., the probability of selecting an apple from Box A, which is 0.3)
- P(A) is the prior probability of selecting Box A (i.e., the probability of selecting Box A before any observation, which is 0.5)
- P(B) is the prior probability of selecting an apple (i.e., the probability of selecting an apple from either Box A or Box B, which is 0.5 x 0.3 + 0.5 x 0.7 = 0.5) #### Therefore, we have:

P(A|B) = 0.3 x 0.5 / 0.5 = 0.3

This means that the probability of selecting Box A given that we have observed an apple is 0.3, or 30%.

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

Ans.A confidence interval is a range of values that is likely to contain the true value of a population parameter (such as the mean or proportion) with a certain level of confidence. It is a measure of the precision or uncertainty of an estimate based on a sample of data.

The confidence interval is calculated by specifying a confidence level (usually 95% or 99%) and using the sample data to estimate the standard error of the sample statistic. The confidence interval is then calculated as the sample statistic plus or minus a margin of error, which is determined by multiplying the standard error by a critical value from the appropriate distribution (usually the t-distribution or z-distribution, depending on the sample size and whether the population standard deviation is known or not).

For example, suppose we want to estimate the average height of students in a particular school with 95% confidence. We take a random sample of 50 students and measure their heights. The sample mean height is 170 cm, and the sample standard deviation is 5 cm. We can use the t-distribution to calculate the critical value for a 95% confidence interval with 49 degrees of freedom (n-1):

t(0.025, 49) = 2.009

The standard error of the sample mean is:

SE = s / sqrt(n) = 5 / sqrt(50) = 0.71

Therefore, the 95% confidence interval for the population mean height is:

170 ± 2.009 x 0.71 = (168.40, 171.60)

This means that we can be 95% confident that the true mean height of all students in the school is between 168.40 cm and 171.60 cm. If we were to repeat this sampling and estimation process many times, 95% of the resulting confidence intervals would contain the true population mean height.

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.

Ans.Bayes' Theorem is a fundamental concept in probability theory that allows us to update our beliefs about the likelihood of an event occurring based on new evidence. It is expressed as follows:

P(A|B) = P(B|A) * P(A) / P(B)

where P(A|B) is the probability of event A occurring given that event B has occurred, P(B|A) is the probability of event B occurring given that event A has occurred, P(A) is the prior probability of event A, and P(B) is the prior probability of event B.

Here's an example problem that demonstrates how to use Bayes' Theorem:

Suppose that a certain disease affects 1 in 1000 people in a population. A new test has been developed that can detect the disease with 99% accuracy. If a person tests positive for the disease, what is the probability that they actually have the disease?

Solution:

Let A be the event that a person has the disease, and B be the event that a person tests positive for the disease.

We know that P(A) = 0.001 (prior probability of the disease) and P(B|A) = 0.99 (probability of a positive test result given that the person has the disease).

We want to find P(A|B), the probability of a person having the disease given that they have tested positive.

Using Bayes' Theorem, we can calculate:

P(A|B) = P(B|A) * P(A) / P(B)

P(B) is the probability of testing positive for the disease, and this can be calculated using the Law of Total Probability:

P(B) = P(B|A) * P(A) + P(B|~A) * P(~A)

where

- ~A represents the complement of A (i.e., the person does not have the disease).
- P(B|~A) is the probability of a false positive test result, which is equal to 1 - P(test negative|~A) = 1 - 0.99 = 0.01.
- P(~A) is the complement of P(A), which is 1 - 0.001 = 0.999. #### So, we can calculate P(B) as follows:

P(B) = P(B|A) * P(A) + P(B|~A) * P(~A) = 0.99 * 0.001 + 0.01 * 0.999 = 0.01098

Now, we can use Bayes' Theorem to calculate P(A|B):

$P(A|B) = P(B|A) * P(A) / P(B) = 0.99 * 0.001 / 0.01098 = 0.0901639344$

Therefore, the probability of a person actually having the disease given a positive test result is approximately 0.090 or 9.0%.

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.

Ans.To calculate the 95% confidence interval, we need to use the following formula:

$CI = \bar{x} \pm z*(/\sqrt{n})$

Where $\bar{x}$ is the sample mean,   is the population standard deviation, n is the sample size, and z is the critical value from the standard normal distribution for a 95% confidence interval, which is approximately 1.96.

Substituting the given values, we have:

$CI = 50 \pm 1.96*(5/\sqrt{n})$

We do not know the sample size, so we cannot calculate the exact confidence interval. However, we can provide a general interpretation of the results.

The 95% confidence interval tells us that if we were to take many samples of the same size from the population and calculate the mean for each sample, 95% of those sample means would fall within the calculated confidence interval. In other words, we can be 95% confident that the true population mean lies within the interval.

For example, if we had a sample size of 100, the 95% confidence interval would be:

$CI = 50 \pm 1.96*(5/\sqrt{100}) = 50 \pm 0.98 = (49.02, 50.98)$

This means that we can be 95% confident that the true population mean lies within the range of 49.02 to 50.98.

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.

Ans.The margin of error (MOE) is a measure of the precision of an estimate in a confidence interval. It represents the range of values above and below the estimate that is likely to contain the true population value with a certain level of confidence.

The formula for MOE is: MOE = z* (/√n)

where z is the critical value from the standard normal distribution corresponding to the desired level of confidence, is the population standard deviation, and n is the sample size.

Sample size affects the margin of error in two ways: As sample size increases, the margin of error decreases. As sample size decreases, the margin of error increases. This is because a larger sample size provides more information about the population and reduces the effect of sampling variability. As a result, the estimate of the population parameter becomes more precise, which leads to a smaller margin of error.

Here's an example to illustrate the impact of sample size on the margin of error: Suppose we want to estimate the average income of households in a certain city. We take a sample of 100 households and find that the sample mean income is 60000 with as standard deviation of 10,000. We want to construct a 95% confidence interval for the true population mean income.

Using the formula for MOE, we have:

MOE = 1.96 * (10,000 / √100) = $1,960

So the 95% confidence interval for the true population mean income is:

600001±960 or (58040,61960)

Now, suppose we increase the sample size to 400 households while keeping the same standard deviation of $10,000. The MOE for the new sample size is:

MOE = 1.96 * (10,000 / √400) = $980

So the 95% confidence interval for the true population mean income based on the larger sample size is:

60000±980 or (59020,60980)

As we can see, the larger sample size results in a smaller margin of error and a narrower confidence interval. This means that we have a more precise estimate of the true population mean income with a higher level of confidence.

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.

Ans.The z-score, also known as the standard score, measures how many standard deviations a data point is away from the population mean. It is calculated using the formula:

z = (x -) /

where x is the data point, is the population mean, and is the population standard deviation.

Substituting the given values, we have:

z = (75 - 70) / 5 z = 1

So the z-score for the data point with a value of 75 is 1.

Interpreting the results, we can say that the data point of 75 is one standard deviation above the population mean of 70. This means that the value of 75 is relatively high compared to the rest of the population data, and it is unlikely to have occurred by chance. The z-score also allows us to compare this data point to other data points in the population, regardless of the units or scales used to measure them.

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.

Ans.To conduct a hypothesis test to determine if the weight loss drug is significantly effective at a 95% confidence level using a t-test, we need to follow the following steps:

1. State the null and alternative hypotheses: The null hypothesis (H0) is that the mean weight loss of the population is equal to 0 pounds. The alternative hypothesis (Ha) is that the mean weight loss of the population is greater than 0 pounds. H0: = 0 Ha: > 0

2. Set the level of significance: We are given that we want to test the hypothesis at a 95% confidence level, which means the level of significance is = 0.05.

3. Calculate the test statistic: Since the population standard deviation is unknown, we need to use a t-test. The test statistic is calculated using the formula: t = (x̄ -) / (s / √n) where x̄ is the sample mean, is the hypothesized population mean (0 in this case), s is the sample standard deviation, and n is the sample size. Substituting the given values, we have: t = (6 - 0) / (2.5 / √50) t = 13.416

4. Calculate the p-value: Using a t-distribution table with degrees of freedom (df) = n - 1 = 49 and a one-tailed test, we find that the critical t-value at a 95% confidence level is 1.677. Since our calculated t-value of 13.416 is greater than the critical t-value, we can reject the null hypothesis in favor of the alternative hypothesis. To calculate the p-value, we use a t-distribution calculator with the calculated t-value of 13.416 and degrees of freedom (df) = 49. The resulting p-value is less than 0.0001, which is much smaller than the level of significance of 0.05.

5. Make a conclusion: Since the p-value is less than the level of significance, we can conclude that there is sufficient evidence to support the claim that the weight loss drug is significantly effective at a 95% confidence level. We can reject the null hypothesis and accept the alternative hypothesis that the mean weight loss of the population is greater than 0 pounds.

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.

Ans.To calculate the confidence interval, we will use the formula:

CI = p ± z* (sqrt(p*(1-p)/n))

where:

- p = the sample proportion (65% or 0.65 in this case)
- z = the critical value for the desired confidence level (we will use 1.96 for 95% confidence level)
- n = the sample size (500 in this case) #### Plugging in the values:

CI = 0.65 ± 1.96 * (sqrt(0.65*(1-0.65)/500)) = 0.65 ± 0.047 = (0.603, 0.697)

Therefore, we can say with 95% confidence that the true proportion of people who are satisfied with their job is between 0.603 and 0.697.

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.

Ans.To test whether there is a significant difference in student performance between the two teaching methods, we will use a two-sample t-test. The null hypothesis (H0) is that there is no difference between the two teaching methods, while the alternative hypothesis (Ha) is that there is a significant difference between the two teaching methods.

H0: A = B Ha: A B

where:

- A = mean score of sample A
- B = mean score of sample B #### The significance level is 0.01, so we need to find the critical t-value for a two-tailed test with 98 degrees of freedom (df = nA + nB - 2 = 20 + 25 - 2 = 43). Using a t-table or a statistical software, the critical t-value is approximately ±2.68.

The test statistic can be calculated as follows:

t = (x̄A - x̄B) / (sqrt(Sp^2/nA + Sp^2/nB))

where:

- x̄A = mean score of sample A (85)
- x̄B = mean score of sample B (82)
- Sp^2 = pooled variance [(nA-1)S^2A + (nB-1)S^2B] / (nA + nB - 2)
- S^2A = variance of sample A (6^2 = 36)
- S^2B = variance of sample B (5^2 = 25)
- nA = sample size of sample A (20)
- nB = sample size of sample B (25) #### Plugging in the values:

Sp^2 = [(20-1)36 + (25-1)25] / (20+25-2) = 30.54 t = (85-82) / (sqrt(30.54/20 + 30.54/25)) = 2.20

The calculated t-value (2.20) is less than the critical t-value (±2.68), so we fail to reject the null hypothesis. Therefore, we do not have sufficient evidence to conclude that there is a significant difference in student performance between the two teaching methods at the 0.01 significance level.

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.

Ans.To calculate the confidence interval, we will use the formula:

CI = $\bar{x}$ ± z* (s/√n)

where:

- $\bar{x}$ = the sample mean (65)
- z = the critical value for the desired confidence level (we will use 1.645 for 90% confidence level)
- s = the sample standard deviation (8)
- n = the sample size (50) #### Plugging in the values:

CI = 65 ± 1.645 * (8/√50) = 65 ± 2.89 = (62.11, 67.89)

Therefore, we can say with 90% confidence that the true population mean is between 62.11 and 67.89.

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.

Ans.To conduct a hypothesis test to determine if caffeine has a significant effect on reaction time, we can use a one-sample t-test. The null hypothesis is that the mean reaction time with caffeine is equal to the population mean reaction time without caffeine, and the alternative hypothesis is that the mean reaction time with caffeine is different from the population mean reaction time without caffeine. Hypotheses:

- Null Hypothesis (H0): The mean reaction time with caffeine is equal to the population mean reaction time without caffeine. μc = μ0
- Alternative Hypothesis (Ha): The mean reaction time with caffeine is different from the population mean reaction time without caffeine. μc μ0 Level of significance:

= 0.1 (90% confidence level)

Sample size and statistics: Sample size (n) = 30 Sample mean ($\bar{x}$) = 0.25 seconds Sample standard deviation (s) = 0.05 seconds Assumptions:

The sample is a random sample from the population of interest. The population follows a normal distribution or the sample size is large enough ($n > 30$). The population standard deviation is unknown. Since the population standard deviation is unknown, we will use the t-distribution to calculate the p-value.

The test statistic can be calculated using the formula: $t = (\bar{x} - \mu0) / (s / \sqrt{n})$

where $\bar{x}$ is the sample mean, $\mu0$ is the hypothesized population mean, s is the sample standard deviation, and n is the sample size.

Substituting the values, we get:

$t = (0.25 - \mu0) / (0.05 / \sqrt{30})$

The critical values for a two-tailed test with = 0.1 and degrees of freedom (df) = n - 1 = 29 are ± 1.699.

The rejection region is $t < -1.699$ or $t > 1.699$.

Calculating the test statistic: $t = (0.25 - \mu0) / (0.05 / \sqrt{30})$

Assuming the null hypothesis is true, the hypothesized population mean is equal to the sample mean with caffeine, i.e., $\mu0 = 0.25$.

$t = (0.25 - 0.25) / (0.05 / \sqrt{30}) = 0$

The calculated t-value is 0, which is not in the rejection region.

Calculating the p-value: The p-value is the probability of observing a test statistic as extreme as the calculated t-value, assuming the null hypothesis is true.

For a two-tailed test, the p-value is the probability of observing a t-value greater than the calculated t-value or less than the negative of the calculated t-value.

Using a t-distribution table or a calculator, the p-value for a t-value of 0 with df = 29 is 1.

Since the p-value is greater than the level of significance ($p >$), we fail to reject the null hypothesis.

Conclusion: At a 90% confidence level, there is insufficient evidence to conclude that caffeine has a significant effect on reaction time. The study does not provide evidence that the mean reaction time with caffeine is different from the population mean reaction time without caffeine.