# 11 march assignment

**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.**

A t-test and a z-test are both statistical tests used to make inferences about population parameters based on sample data. However, they have different applications and assumptions based on the characteristics of the data and the sample size.

**T-Test:**
A t-test is used when the sample size is relatively small (typically less than 30) and the population standard deviation is unknown. It's particularly useful when working with small samples or when the population standard deviation is not available. The t-test takes into account the variability within the sample and adjusts the results accordingly.

**Example Scenario for a T-Test:**
Imagine you are testing the effectiveness of a new drug to lower blood pressure. You randomly select 20 participants and measure their blood pressure before and after administering the drug. Since the sample size is small and the population standard deviation of blood pressure is not known, you would use a t-test to determine if the mean blood pressure before and after taking the drug is significantly different.

**Z-Test:**
A z-test is used when the sample size is large (typically greater than 30) and the population standard deviation is known or can be reasonably estimated. The z-test assumes a normal distribution of the data and is more appropriate when dealing with larger samples where the Central Limit Theorem comes into play, ensuring that the sample mean is normally distributed.

**Example Scenario for a Z-Test:**
Suppose you want to assess whether the average height of a certain population differs significantly from a known average height of another population. You collect a large sample of 200 individuals from the first population and compute their average height. If you have a reliable estimate of the population standard deviation of height, you could use a z-test to determine if the difference in mean heights is statistically significant.

In summary, choose a t-test when dealing with a small sample size and an unknown population standard deviation, and choose a z-test when working with a large sample size and a known or estimated population standard deviation.

**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 statistical hypothesis test where the alternative hypothesis is focused on a specific direction of effect or difference. In other words, it's used to determine if a parameter is significantly different from a hypothesized value in one particular direction. This can be either a positive direction (greater than) or a negative direction (less than).

For example, if you're testing whether a new treatment increases the average test scores of students, a one-tailed test might involve a hypothesis like:
- Null Hypothesis (H0): The new treatment does not increase the average test scores.
- Alternative Hypothesis (H1): The new treatment increases the average test scores.

Two-Tailed Test:
A two-tailed test, on the other hand, is a type of statistical hypothesis test where the alternative hypothesis is focused on a difference in either direction from the hypothesized value. It's used to determine if a parameter is significantly different from a hypothesized value, regardless of whether the effect is positive or negative.

Continuing with the example of testing a new treatment on test scores, a two-tailed test might involve hypotheses like:
- Null Hypothesis (H0): The new treatment has no effect on the average test scores.
- Alternative Hypothesis (H1): The new treatment does have an effect on the average test scores, whether that effect is an increase or a decrease.

In summary, the key difference between one-tailed and two-tailed tests lies in the direction of the alternative hypothesis. One-tailed tests focus on a specific direction (either greater than or less than), while two-tailed tests consider differences in both directions from the hypothesized value. The choice between one-tailed and two-tailed tests depends on the specific research question and the nature of the hypothesis being tested.

**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 and Type 2 errors are concepts in hypothesis testing that deal with the potential mistakes that can occur when making decisions based on statistical tests.

**Type 1 Error (False Positive):**
A Type 1 error occurs when you reject a null hypothesis that is actually true. In other words, you conclude that there is a significant effect or difference when there isn't one in reality. This is also known as a false positive or alpha error. The probability of making a Type 1 error is denoted by the symbol "α" (alpha) and is typically set as the significance level for the test (e.g., 0.05).

**Example Scenario for Type 1 Error:**
Suppose a pharmaceutical company is testing a new drug to determine if it's effective in reducing cholesterol levels. The null hypothesis (H0) is that the drug has no effect on cholesterol levels. If the researchers perform a statistical test and conclude that the drug is effective (rejecting H0) when it actually isn't, they have committed a Type 1 error. This could lead to the drug being approved and marketed as effective when it's not.

**Type 2 Error (False Negative):**
A Type 2 error occurs when you fail to reject a null hypothesis that is actually false. In other words, you conclude that there is no significant effect or difference when there is one in reality. This is also known as a false negative or beta error. The probability of making a Type 2 error is denoted by the symbol "β" (beta).

**Example Scenario for Type 2 Error:**
Let's consider a quality control scenario. A factory is testing whether a new manufacturing process has significantly reduced the number of defective products. The null hypothesis (H0) is that the new process has no effect on defect rates. If the factory performs a statistical test and fails to reject H0 (concluding that the process hasn't improved defect rates) when the process actually does reduce defects, they have committed a Type 2 error. This could lead to the factory continuing to use an inefficient process.

In summary:
- **Type 1 Error (False Positive):** Rejecting a true null hypothesis.
- **Type 2 Error (False Negative):** Failing to reject a false null hypothesis.

These errors represent the trade-off between the risk of making a false positive decision and the risk of making a false negative decision in hypothesis testing. Adjusting the significance level (α) and sample size can influence the likelihood of these errors occurring, but there's typically a trade-off between the two types of errors – reducing one often increases the other.

**Q4: Explain Bayes's theorem with an example.**

Bayes's theorem is a fundamental concept in probability theory and statistics that allows us to update the probability of a hypothesis based on new evidence or information. It's named after the 18th-century mathematician and theologian Thomas Bayes.

The theorem mathematically relates the conditional probability of an event A given event B to the conditional probability of event B given event A. In simple terms, it helps us calculate the probability of a cause given an observed effect, by taking into account both the prior probability and the likelihood of the evidence.

The formula for Bayes's theorem is as follows:

$$ P(A|B) = \frac{P(B|A) \cdot P(A)}{P(B)} $$

Where:
- $P(A|B)$ is the probability of event A given that event B has occurred (posterior probability).
- $P(B|A)$ is the probability of event B given that event A has occurred (likelihood).
- $P(A)$ is the prior probability of event A.
- $P(B)$ is the probability of event B.

**Example Scenario: Medical Test**
Let's consider an example involving a medical test. Suppose a rare disease affects 1 in 1000 people. The test for this disease is not perfect – it has a false positive rate of 5% and a false negative rate of 2%.

1. Prior Probability: $P(\text{Disease}) = 0.001$ (1 in 1000)
2. False Positive Rate: $P(\text{Positive Result|No Disease}) = 0.05$
3. False Negative Rate: $P(\text{Negative Result|Disease}) = 0.02$

Now, imagine a person tests positive for the disease. We want to find the probability that the person actually has the disease.

Using Bayes's theorem:
- $P(\text{Positive Result|Disease}) = 1 - P(\text{Negative Result|Disease}) = 0.98$
- $P(\text{Positive Result}) = P(\text{Positive Result|Disease}) \cdot P(\text{Disease}) + P(\text{Positive Result|No Disease}) \cdot P(\text{No Disease})$
  $= 0.98 \cdot 0.001 + 0.05 \cdot (1 - 0.001) = 0.05092$

Now, applying Bayes's theorem:
$$ P(\text{Disease|Positive Result}) = \frac{P(\text{Positive Result|Disease}) \cdot P(\text{Disease})}{P(\text{Positive Result})} $$
$$ = \frac{0.98 \cdot 0.001}{0.05092} \approx 0.0192 $$

So, even though the test is positive, the probability of the person having the disease is only around 1.92%. This demonstrates how Bayes's theorem helps us update our beliefs based on new evidence.

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

A confidence interval is a range of values within which a population parameter (such as a mean or proportion) is estimated to lie, based on a sample from that population and a certain level of confidence. It provides a measure of the uncertainty associated with the estimation process.

When you calculate a point estimate (e.g., sample mean) from a sample, it's just a single value. However, this value might not perfectly represent the true population parameter due to random sampling variability. A confidence interval gives you a range of values that is likely to contain the true population parameter with a specified level of confidence.

**Calculating a Confidence Interval:**
A confidence interval is typically expressed as:

\[ \text{Point Estimate} \pm \text{Margin of Error} \]

The "Point Estimate" is the calculated value from your sample data (e.g., sample mean), and the "Margin of Error" represents how much the estimate might vary due to sampling variability.

The formula for calculating the margin of error depends on the type of parameter you're estimating (mean, proportion, etc.), the sample size, and the desired level of confidence.

**Example Scenario: Confidence Interval for Mean Height:**
Suppose you want to estimate the average height of all adults in a certain city. You take a random sample of 100 adults and find that the average height in your sample is 170 cm, with a sample standard deviation of 5 cm. You want to calculate a 95% confidence interval for the true population mean height.

Using a t-distribution (since sample size is moderate), and assuming a normal distribution for heights:

1. Find the critical value corresponding to a 95% confidence level and the degrees of freedom (which is \(n - 1\) for a t-distribution, where \(n\) is the sample size). Let's assume the critical t-value is 1.984.

2. Calculate the standard error of the mean (\(SE\)) using the formula: \(\frac{\text{Sample Standard Deviation}}{\sqrt{n}}\)
   \(SE = \frac{5}{\sqrt{100}} = 0.5\)

3. Calculate the margin of error (\(MOE\)) using the formula: \(t \times SE\), where \(t\) is the critical t-value.
   \(MOE = 1.984 \times 0.5 = 0.992\)

4. The confidence interval is then: \(170 \pm 0.992\), which gives a range of 169.008 to 170.992.

Interpretation: We are 95% confident that the true average height of all adults in the city falls between 169.008 and 170.992 cm.

Remember that the level of confidence indicates the percentage of times this method would produce an interval that contains the true population parameter if you were to repeat the sampling process many times.

**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.**

Certainly! Let's work through an example problem using Bayes's Theorem to calculate the probability of an event occurring given prior knowledge and new evidence.

**Sample Problem:**
Suppose you are a doctor trying to diagnose a rare disease called Disease X. You know that Disease X affects 1 in 1000 people in the general population. There's a specific blood test for Disease X, but the test is not perfect. It has a false positive rate of 5% (meaning it incorrectly indicates the disease when the person doesn't have it) and a false negative rate of 2% (meaning it fails to detect the disease when the person actually has it).

Now, a patient comes in who tested positive for Disease X. What is the probability that the patient actually has the disease?

**Solution:**
Let's use Bayes's Theorem to calculate the probability that the patient has the disease given the positive test result.

1. **Given Information:**
   - Prior Probability of Disease (\(P(\text{Disease})\)): 0.001 (1 in 1000)
   - False Positive Rate (\(P(\text{Positive Result|No Disease})\)): 0.05
   - False Negative Rate (\(P(\text{Negative Result|Disease})\)): 0.02

- Positive Result Given Disease (\(P(\text{Positive Result|Disease})\)): 1 - False Negative Rate = 0.98

2. **Calculate the Denominator (Probability of Positive Result):**
Using the Law of Total Probability, the probability of a positive result is the sum of two cases:
- Positive Result Given Disease (\(P(\text{Positive Result|Disease})\)) multiplied by Prior Probability of Disease (\(P(\text{Disease})\)).
- Positive Result Given No Disease (\(P(\text{Positive Result|No Disease})\)) multiplied by Prior Probability of No Disease (\(1 - P(\text{Disease})\)).

\(P(\text{Positive Result}) = P(\text{Positive Result|Disease}) \cdot P(\text{Disease}) + P(\text{Positive Result|No Disease}) \cdot (1 - P(\text{Disease})) \\
= 0.98 \cdot 0.001 + 0.05 \cdot (1 - 0.001) \\
= 0.05092\)

3. **Apply Bayes's Theorem:**
Now, use Bayes's Theorem to calculate the probability of having the disease given the positive test result.

\(P(\text{Disease|Positive Result}) = \frac{P(\text{Positive Result|Disease}) \cdot P(\text{Disease})}{P(\text{Positive Result})} \\
= \frac{0.98 \cdot 0.001}{0.05092} \\
\approx 0.0192\)

So, even though the test is positive, the probability that the patient actually has the disease is only around 1.92%. This demonstrates how Bayes's Theorem helps us update our beliefs about the event's probability based on new evidence.

**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 with a mean of 50 and a standard deviation of 5, we can use the formula for the confidence interval for the population mean when the sample standard deviation is known:

\[ \text{Confidence Interval} = \text{Sample Mean} \pm \text{Margin of Error} \]

The margin of error depends on the desired level of confidence (95% in this case), the sample size, the standard deviation, and the critical value from the standard normal distribution (Z-distribution). For a 95% confidence level, the critical Z-value is approximately 1.96.

Given:

- Sample Mean (\( \text{X̄} \)): 50
- Standard Deviation (\( \sigma \)): 5
- Sample Size (\( n \)): Not provided (we'll assume a reasonable sample size of, say, 30 for demonstration purposes)

**Calculations:**

1. Calculate the standard error (\( SE \)) using the formula: \( \frac{\sigma}{\sqrt{n}} \)

2. Calculate the margin of error (\( \text{MOE} \)) using the formula: \( Z \times SE \), where \( Z \) is the critical Z-value for a 95% confidence level (approximately 1.96).

3. The confidence interval is then: \( \text{Sample Mean} \pm \text{MOE} \)

Assuming a sample size of 30 for demonstration:
\[ SE = \frac{5}{\sqrt{30}} \approx 0.9129 \]
\[ \text{MOE} = 1.96 \times 0.9129 \approx 1.7897 \]

So, the 95% confidence interval is:
\[ 50 \pm 1.7897 \]
\[ \text{Lower Limit} \approx 48.2103 \]
\[ \text{Upper Limit} \approx 51.7897 \]

**Interpretation:**

With 95% confidence, we can say that the true population mean lies within the interval of approximately 48.2103 to 51.7897. This means that if we were to repeatedly sample and calculate confidence intervals from the same population, we would expect that 95% of those intervals would contain the true population mean of the data. The wider the confidence interval, the higher the uncertainty in our estimate, and vice versa.

**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 with a specified level of confidence. It represents the amount by which the point estimate (e.g., sample mean) could vary due to random sampling variability.

The formula to calculate the margin of error depends on several factors:
\[ \text{MOE} = \text{Critical Value} \times \text{Standard Error} \]

- **Critical Value:** This is determined by the chosen confidence level and the distribution being used (usually the standard normal distribution Z-distribution or the t-distribution). A higher confidence level requires a larger critical value.

- **Standard Error:** This accounts for the variability in the sample data and is often calculated as the ratio of the population standard deviation to the square root of the sample size (\( \frac{\sigma}{\sqrt{n}} \)).

**Effect of Sample Size on Margin of Error:**
A larger sample size generally leads to a smaller margin of error. When the sample size is larger, the variability in the sample mean becomes smaller, and as a result, the estimate of the true population parameter becomes more precise.

**Example Scenario:**
Let's consider a scenario involving a political survey. Suppose you want to estimate the proportion of voters who support a certain candidate. You take two samples: one with a sample size of 200 and another with a sample size of 1000.

1. Sample with 200 people:
   - Standard Error (\( SE \)) is larger due to the smaller sample size.
   - Consequently, the margin of error (\( MOE \)) will be relatively larger, leading to a wider confidence interval.

2. Sample with 1000 people:
   - Standard Error (\( SE \)) is smaller due to the larger sample size.
   - Consequently, the margin of error (\( MOE \)) will be relatively smaller, resulting in a narrower confidence interval.

In this example, the larger sample size (1000) yields a smaller margin of error and a more precise estimate of the proportion of voters who support the candidate. This means that with the larger sample, the confidence interval will be narrower, and you'll have more confidence that the interval contains the true population proportion.

**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 z-score (also known as the standard score) measures how many standard deviations a data point is away from the mean of a population. It's calculated using the formula:

\[ \text{z-score} = \frac{\text{data point value} - \text{population mean}}{\text{population standard deviation}} \]

Given the values:
- Data point value: 75
- Population mean: 70
- Population standard deviation: 5

Let's calculate the z-score:

\[ \text{z-score} = \frac{75 - 70}{5} = 1 \]

**Interpretation:**
The calculated z-score is 1. This means that the data point with a value of 75 is 1 standard deviation above the mean of the population. In other words, it's relatively higher than the average value of the population by one standard deviation. The positive sign of the z-score indicates that the data point is above the mean. Z-scores allow us to standardise data and compare values from different distributions on a common scale.

**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.**

Null hypothesis (H0): The new drug is not significantly effective (\(\mu = 0\)).
Alternative hypothesis (H1): The new drug is significantly effective (\(\mu \neq 0\)).

Using a t-test for a one-sample mean:
Calculate the t-statistic:
\[ t = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}} = \frac{6 - 0}{\frac{2.5}{\sqrt{50}}} \approx 12.72 \]

Degrees of freedom: \(n - 1 = 50 - 1 = 49\)

At a 95% confidence level, the critical t-value is approximately \(2.0096\) for a two-tailed test.

Since \(12.72 > 2.0096\), we reject the null hypothesis.

**Interpretation:** At a 95% confidence level, the study provides strong evidence to suggest that the new weight loss drug is significantly effective.

**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.**

Using the formula for the confidence interval for a proportion:
\[ \text{Confidence Interval} = p \pm Z \times \sqrt{\frac{p(1-p)}{n}} \]

At a 95% confidence level, the critical Z-value is approximately \(1.96\).

Calculate the confidence interval:
\[ \text{Confidence Interval} = 0.65 \pm 1.96 \times \sqrt{\frac{0.65 \times 0.35}{500}} \approx (0.610, 0.690) \]

**Interpretation:** With 95% confidence, the true proportion of people satisfied with their job lies within the interval of approximately 61.0% to 69.0%.

**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.**

Certainly, let's conduct a hypothesis test to determine if there is a significant difference in student performance between two teaching methods. We'll use a two-sample t-test for this scenario.

Given:
- Sample A mean (\(\bar{x}_A\)): 85
- Sample A standard deviation (\(s_A\)): 6
- Sample A size (\(n_A\)): Not provided (we'll assume it's a reasonable value, say, 30)
- Sample B mean (\(\bar{x}_B\)): 82
- Sample B standard deviation (\(s_B\)): 5
- Sample B size (\(n_B\)): Not provided (we'll assume it's a reasonable value, say, 30)
- Significance level (\(\alpha\)): 0.01

Null hypothesis (H0): There is no significant difference in student performance (\(\mu_A - \mu_B = 0\)).
Alternative hypothesis (H1): There is a significant difference in student performance (\(\mu_A - \mu_B \neq 0\)).

Using a t-test for two independent samples:
Calculate the pooled standard deviation (\(s_p\)):
\[ s_p = \sqrt{\frac{(n_A - 1)s_A^2 + (n_B - 1)s_B^2}{n_A + n_B - 2}} \]

Calculate the t-statistic:
\[ t = \frac{(\bar{x}_A - \bar{x}_B)}{s_p \sqrt{\frac{1}{n_A} + \frac{1}{n_B}}} \]

Degrees of freedom: \(n_A + n_B - 2\)

At a significance level of 0.01, the critical t-value for a two-tailed test and the calculated degrees of freedom is obtained from a t-distribution table.

Calculate the t-statistic, degrees of freedom, and the critical t-value.

If the calculated t-statistic is beyond the critical t-value in either tail, you can reject the null hypothesis.

Please provide the values of \(n_A\) and \(n_B\) for a more accurate calculation.

**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.**

Certainly, to calculate the 90% confidence interval for the true population mean, you can use the formula for the confidence interval for the population mean when the population standard deviation is known:

\[ \text{Confidence Interval} = \text{Sample Mean} \pm \text{Margin of Error} \]

The margin of error depends on the desired level of confidence (90% in this case), the population standard deviation, and the critical value from the standard normal distribution (Z-distribution). For a 90% confidence level, the critical Z-value is approximately 1.645.

Given:
- Population mean (\( \mu \)): 60
- Population standard deviation (\( \sigma \)): 8
- Sample mean (\( \bar{x} \)): 65
- Sample size (\( n \)): 50
- Confidence level: 90%

**Calculations:**

1. Calculate the standard error (\( SE \)) using the formula: \( \frac{\sigma}{\sqrt{n}} \)

2. Calculate the margin of error (\( \text{MOE} \)) using the formula: \( Z \times SE \), where \( Z \) is the critical Z-value for a 90% confidence level (approximately 1.645).

3. The confidence interval is then: \( \text{Sample Mean} \pm \text{MOE} \)

\[ SE = \frac{8}{\sqrt{50}} \approx 1.131 \]
\[ \text{MOE} = 1.645 \times 1.131 \approx 1.863 \]

So, the 90% confidence interval is:
\[ 65 \pm 1.863 \]
\[ \text{Lower Limit} \approx 63.137 \]
\[ \text{Upper Limit} \approx 66.863 \]

**Interpretation:**

With 90% confidence, we can say that the true population mean lies within the interval of approximately 63.137 to 66.863. This means that if we were to repeatedly sample and calculate confidence intervals from the same population, we would expect that 90% of those intervals would contain the true population mean of the data. The wider the confidence interval, the higher the uncertainty in our estimate, and vice versa.

**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.**

Certainly, let's conduct a hypothesis test to determine if caffeine has a significant effect on reaction time using a t-test. We'll perform a one-sample t-test for this scenario.

Given:
- Sample size (\( n \)): 30
- Sample mean (\( \bar{x} \)): 0.25 seconds
- Sample standard deviation (\( s \)): 0.05 seconds
- Confidence level: 90%

Assumptions:
- The sample is randomly selected and comes from a normally distributed population.
- The sample size is sufficiently large for the Central Limit Theorem to apply.

Null hypothesis (H0): Caffeine has no significant effect on reaction time (\( \mu = \mu_0 \)).
Alternative hypothesis (H1): Caffeine has a significant effect on reaction time (\( \mu \neq \mu_0 \)).

Here, \( \mu_0 \) represents the hypothesized population mean reaction time without caffeine.

Using a t-test for a one-sample mean:
Calculate the t-statistic:

$$t = \frac{\bar{x} - \mu_0}{\frac{s}{\sqrt{n}}} = \frac{0.25 - \mu_0}{\frac{0.05}{\sqrt{30}}}$$

Degrees of freedom: $n - 1 = 30 - 1 = 29$

At a 90% confidence level, the critical t-value is obtained from a t-distribution table.

If the calculated t-statistic is beyond the critical t-value in either tail, you can reject the null hypothesis.

For this calculation, you would need to provide the hypothesized population mean ($\mu_0$) or perform a two-tailed test with a range of values around the sample mean ($\bar{x}$) to compare against the critical t-value.

Please provide the hypothesized population mean ($\mu_0$) or the range around the sample mean that you're interested in testing.