Advanced statistics

Reading Material

PW SKILLS

Topics:

1. Central Limit Theorem
2. Hypothesis Testing
3. Confidence Interval (CI)
4. P-Value
5. Z-test
6. T-test
7. Chi-square Test
8. ANOVA Test

1. Central Limit Theorem

Definition and Importance:
The Central Limit Theorem (CLT) states that when you take a big enough sample of independent identical random variables, their average tends to follow a normal distribution. This holds true no matter what the original distribution looks like. The CLT plays a key role in statistics because it lets us use normal distribution methods even when our data isn't normally distributed. Many statistical techniques and hypothesis tests rely on the CLT as their foundation.

Assumptions and Conditions:

To apply the Central Limit Theorem, you need to meet certain requirements:

1. **Independence:** The sample observations need to be free from each other's influence. No observation should have an impact on the others.
2. **Distributed:** The random variables must come from the same probability distribution. This means they should share the same statistical properties.
3. **Sample Size:** You need a big enough sample size. While no hard rule exists, experts often suggest at least 30 samples for the CLT to work well. Keep in mind, distributions that are very skewed might require even more samples.

Applications in Statistical Inference:

The Central Limit Theorem plays a key role in statistical inference, including:

1. **Confidence Intervals:** The CLT helps build confidence intervals for population parameters. The sample mean follows a near-normal distribution, so we can calculate confidence intervals using the standard normal distribution.
2. **Hypothesis Testing:** The CLT backs up hypothesis testing. It lets us use normal distribution estimates for test statistics. This makes it easier to figure out p-values and make decisions in hypothesis testing.
3. **Estimation:** The CLT makes sure sample means give a fair estimate of the population mean. Their distribution looks like a normal one. This makes estimation and analysis simpler.

Sampling Distribution of the Mean:

The sampling distribution of the mean shows how sample means spread out when you take many random samples from the same group. The Central Limit Theorem tells us:

1. **Normal Distribution:** For big samples, the sampling distribution of the mean looks like a normal distribution. This happens no matter what shape the original group's distribution has.
2. **Mean and Variance:** The average of the sampling distribution of the mean matches the group's average. The spread of the sampling distribution of the mean equals the group's spread divided by the sample size (sigma^2/n). One can call the standard deviation of the sampling distribution the standard error.

2. Hypothesis testing

Null and Alternative Hypotheses:

- **Null Hypothesis (H_0):** The null hypothesis states that no effect or difference exists. It acts as the starting point in hypothesis testing. Tests aim to challenge or disprove this assumption. For instance, H_0 might claim that the means of two groups show no difference.
- **Alternative Hypothesis (H_1 or Ha):** The alternative hypothesis states that an effect or difference exists. It represents the idea researchers want to provide evidence for. As an example, H_1 might claim that a meaningful difference exists between the means of two groups.

Types of Errors (Type I and Type II):

- **Type I Error (α):** People also call this a false positive. A Type I error happens when someone rejects the null hypothesis even though it's true. The chance of making a Type I error is called α (alpha). This α is also the significance level of the test.
- **Type II Error (β):** This one goes by the name false negative too. A Type II error takes place when someone doesn't reject the null hypothesis, but it's false. We use β (beta) to show the chance of making a Type II error. To figure out the power of the test, we subtract β from 1.

Statistical Power:

- **Definition:** Statistical power has an influence on the chance of rejecting the null hypothesis when it's false. It gauges the test's ability to spot a difference or effect if one exists. More power means you're more likely to find a real effect.

- **Factors Affecting Power:**
  - **Sample Size:** Bigger samples boost the power of the test.
  - **Effect Size:** It's easier to spot larger effects leading to more power.
  - **Significance Level (α):** Bumping up α can increase power, but it also ups the risk of a Type I error.
  - **Variability:** Less scatter in the data can make the test more powerful.

Steps to Test a Hypothesis:

1. **Set up your Guesses:** Write down what you think might happen and what might not.
2. **Pick how sure you want to be:** Decide how confident you need to be to say your guess is wrong. People often go with 0.05.
3. **Get and Look at Info:** Gather all the facts and numbers then run the right math on them.
4. **Do the Math:** Figure out your test number and p-value to see how likely your first guess is right.
5. **Choose what to believe:** Look at your p-value and how sure you wanted to be. If p is smaller, your first guess is wrong. If not, you can't say it's wrong yet.
6. **Sum it all up:** Explain what all this means for your big question.

One-tailed vs. Two-tailed Tests:

One-tailed Test:
- **Definition:** A one-tailed test checks the hypothesis in a single direction. It determines if the sample average is much higher or lower than a predicted value.
- **Application:** You should use a one-tailed test when your research idea points to a specific direction of the effect (for example, to find out if a new medicine works better than an existing one).

Two-tailed Test:
- **Definition:** A two-tailed test looks at the hypothesis from both sides. It figures out if the sample average differs a lot from a predicted value, no matter which way the difference goes.
- **Application:** Go for a two-tailed test when your research question doesn't point to a specific direction (for example when you're checking if there's any difference in how two groups perform).

3. Confidence Interval

- **Definition**: A confidence interval (CI) is a range of values from a dataset that has the true value of a population parameter with a specific level of confidence. For instance, a 95% confidence interval for a mean shows that if we took many samples and figured out a CI for each sample about 95% of those intervals would have the true population mean.

Calculating Confidence Intervals:

1. **To calculate a Mean (when you know the population standard deviation):**
   - **Formula:** $CI=\bar{x} \pm Z \times \sigma/\sqrt{n}$
     $\bar{x}$ =Sample mean
     Z=Z-score corresponding to the confidence level (e.g., 1.96 for 95% CI)
     σ= The standard deviation of the whole population
     n = Sample size

2. **To calculate a Mean (when you don't know the population standard deviation):**
- **Formula:** $CI = \bar{x} \pm t \times s/\sqrt{n}$
  $\bar{x}$ = Sample mean
  t = t-score from the t-distribution table based on the confidence level you want and degrees of freedom (n - 1)
  s = Sample standard deviation
  n = Sample size

3. **For Proportions:**
- **Formula:** $CI=\hat{p} \pm Z \times \sqrt{\hat{p}(1-\hat{p})/n}$
  $\hat{p}$ = Sample proportion
  Z = Z-score corresponding to the confidence level
  n = Sample size

Factors That Affect CI Width:

1. **Sample Size (n):** Bigger samples result in slimmer confidence intervals. When you increase the sample size, the standard error goes down. This leads to a more accurate guess of the population parameter.
2. **Confidence Level:** Higher confidence levels make confidence intervals wider. For instance, a 99% CI will be broader than a 95% CI. This happens because it needs a bigger range to make sure the true parameter is included with more certainty.
3. **Population Variability:** More variety in the population causes wider confidence intervals. Bigger standard deviations create larger standard errors, which expand the CI.

Link to Hypothesis Testing:

- **Connection:** Confidence intervals and hypothesis testing have a link as both aim to estimate population parameters and evaluate evidence against hypotheses.
- **Hypothesis Testing and CI:** When a null hypothesis value lies outside the confidence interval, we consider it significant at the chosen confidence level. On the other hand, if the null hypothesis value falls within the confidence interval, we don't have enough proof to reject the null hypothesis.
- **Decision Making:** Confidence intervals give a range of likely values for a parameter, which offers more insights than a simple hypothesis test result. They help us grasp how precise and reliable the estimate is.
- **Interpretation:** The confidence level shows how often confidence intervals would include the true parameter if you repeated the experiment many times. A 95% CI means there's a 95% chance that the interval contains the true parameter. Keep in mind that this doesn't mean there's a 95% chance the true parameter is within the calculated interval for any one sample.

4. P-Value

- **Definition:** The p-value shows the likelihood of seeing a test statistic as extreme as the one observed, if we assume the null hypothesis is true. It measures how much proof we have against the null hypothesis.
- **Interpretation:** A p-value that's smaller points to stronger evidence against the null hypothesis. Let's say we get a p-value of 0.03. This means there's a 3% chance of getting the data we saw, or something even more extreme, if the null hypothesis were true. if the p-value is less than the significance level we set beforehand (like 0.05), we reject the null hypothesis and go with the alternative hypothesis instead.

Calculating P-Values:

1. **Calculate the Test Statistic:** Work out the test statistic (like z-score or t-score) using your sample data and the null hypothesis.
2. **Get the P-Value:** Use the test statistic and the right distribution (such as normal distribution for z-tests or t-distribution for t-tests) to find the p-value. You can do this with statistical tables, software, or calculators.
3. **One-Tailed vs. Two-Tailed Tests:**
   - For a **one-tailed test**, the p-value is the area under the distribution curve to the right (or left) of the test statistic.
   - In a **two-tailed test**, we calculate the p-value by adding up the areas in both tails of the distribution beyond the absolute value of the test statistic.

Common Misinterpretations:

1. **P-Value Doesn't Show How Likely the Null Hypothesis Is:** A p-value doesn't tell you how probable the null hypothesis is. Rather, it shows how you'd see your data if the null hypothesis were true.
2. **P-Value Isn't the Chance of Getting It Wrong:** The p-value doesn't represent how likely you are to make a Type I or Type II mistake.
3. **A Small P-Value Doesn't Prove the Other Side:** When you get a low p-value, it hints that the null hypothesis might not hold up. But this doesn't prove the alternative hypothesis. You need to look at other things too, like how big your sample is and how strong the effect is.
4. **P-Values Change:** P-values shift depending on the sample. They don't give a final answer but show the data from one specific test. Keep in mind that different experiments can lead to different p-values for the same hypothesis.

P-Value vs. Significance Level:

- **Significance Level (α):** The significance level shown as α, is a limit the researcher sets before they run the hypothesis test. It shows the chance of rejecting the null hypothesis when it's true (Type I error). People often use 0.05, 0.01, and 0.10 as significance levels.
- **P-Value and How to Decide:**
  - **If p-value ≤ α:** Turn down the null hypothesis. This means the data gives enough proof to say that the effect or difference you see is real in a statistical sense.
- **If p-value > α:** We can't reject the null hypothesis. This means the data doesn't give us enough proof to say the effect or difference we see is significant.
- **Relationship:** The p-value shows how unusual our data is if the null hypothesis is true. The significance level is a cutoff point we choose beforehand to make decisions. When we compare the p-value to the significance level, we figure out if we should reject the null hypothesis or not.

5. Z-Test

Assumptions and Conditions:

1. **Normality:** Your data should look like it follows a normal distribution when you don't have many samples. If you have a lot of samples (more than 30), the Central Limit Theorem kicks in. This means that even if your data isn't normal, the average of your samples will be close to normal.

2. **Known Population Variance:** You use a Z-test when you know the population variance (σ2). If you don't know this variance, you'll want to use a t-test instead.
3. **Independence:** Each observation in your data should stand on its own. This means you need to collect your sample data so one observation doesn't influence another.
4. **Random Sampling:** To ensure the sample represents the population, researchers must gather data through random sampling.
5. **Sample Size:** The sample size needs to be big enough for accurate results. While you can use the Z-test with smaller samples, it works best with larger ones (more than 30 participants).

One-Sample Z-Test:

- **Purpose:** This test helps to figure out if the average of one sample is different from a known average of the whole group.

- **Hypotheses:**
  - **Null Hypothesis (H0):** The sample mean is equal to the population mean ($\mu=\mu0$).
  - **Alternative Hypothesis (Ha):** The sample mean is different from the population mean ($\mu\neq\mu0$).

- **Test Statistic:** $Z = (\bar{x} - \mu_0) / (\sigma / \sqrt{n})$
  - $\bar{x}$ = Sample mean
  - μ_0 = Population mean
  - σ = Population standard deviation
  - n = Sample size
  -
- **Decision Rule:** Compare the Z value you got with the critical Z value from the standard normal distribution table at the significance level you chose (e.g., 1.96 for a 95% confidence level). If the Z value you calculated is higher than the critical value, reject the null hypothesis.

Two-Sample Z-Test:

- **Purpose:** This test helps figure out if there's a big difference between the averages of two separate groups.

- **Hypotheses:**
  - **Null Hypothesis (H0H_0H0):** The averages of both groups are the same ($\mu1=\mu2$).
  - **Alternative Hypothesis (HaH_aHa):** The averages of both groups are not the same ($\mu1\neq\mu2$).

- **Test Statistic:** $Z = (\bar{x_1} - \bar{x_2}) / \sqrt{(\sigma_1^2/n_1 + \sigma_2^2/n_2)}$
  - $\bar{x_1}$, $\bar{x_2}$ = Means of the two samples
    σ_1^2 σ_2^2= The two groups' population variances
- n_1n_2= The two groups' sample sizes

- **Decision Rule:** Just like the one-sample Z-test, compare the Z value you calculated with the critical Z value. If the calculated Z is greater than the critical value, reject the null hypothesis.

Z-Test for Proportions:

- **Purpose:** Helps to figure out if the success rate in a sample differs a lot from a known population rate.
- **Hypotheses:**
  - **Null Hypothesis (H0):** The sample proportion is equal to the population proportion ($p=p0$).
- **Alternative Hypothesis (Ha):** The sample proportion is different from the population proportion ($p\neq p0$).

- **Test Statistic:** Z = (p^ - p0) / sqrt((p0 * (1 - p0)) / n)
  - p^ = Sample proportion
  - p0 = Population proportion
  - n = Sample size

- **Decision Rule:** Compare the Z value you calculated with the critical Z value from the standard normal distribution. If your Z value is higher than the critical value, reject the null hypothesis.

6. T-Test

Student's t-Distribution:

- **Definition:** The t-distribution, also known as Student's t-distribution, helps to test hypotheses when you have a small sample and don't know the population standard deviation. It looks like the normal distribution but has fatter tails, which means it's more likely to give values far from the average.

- **Characteristics:**
  - **Mean:** The t-distribution has a mean of 0.
  - **Shape:** The degrees of freedom (df) determine the t-distribution's shape. As you increase the sample size, the t-distribution starts to look more like the normal distribution.

- **Degrees of Freedom (df):** The df relates to the sample size. For a sample of size n, the df for a t-test is n-1.

One-Sample t-Test:

- **Purpose:** The one-sample t-test helps to figure out if a sample's average differs a lot from what we know or think the population's average is when we don't know how spread out the population data is.

- **Hypotheses:**
  - **Null Hypothesis (H0H_0H0):** The average of our sample matches the average of the whole population ($\mu=\mu0$).
  - **Alternative Hypothesis (HaH_aHa):** The average of our sample doesn't match the average of the whole population ($\mu\neq\mu0$).

- **Test Statistic:** $t = (\bar{x} - \mu_0) / (s / \sqrt{n})$
  - $\bar{x}$ = Sample mean
  - μ_0 = Population mean
  - s = Sample standard deviation
  - n = Sample size

- **Decision Rule:** Compare the t value you calculated to the critical t value from the t-distribution table. Use the degrees of freedom and significance level to find this value. If your calculated t is higher than the critical value, reject the null hypothesis.

Independent Samples t-Test:

- **Purpose:** The independent samples t-test (also called the two-sample t-test) helps to figure out if the averages of two separate groups differ a lot from each other.

- **Hypotheses:**
  - **Null Hypothesis (H0):** The averages of the two groups are the same ($\mu1=\mu2$).
  - **Alternative Hypothesis (Ha):** The averages of the two groups are not the same ($\mu1\neq\mu2$).

- **Test Statistic:** $t = (\bar{x_1} - \bar{x_2}) / \sqrt{[(s_1^2/n_1) + (s_2^2/n_2)]}$
  - $\bar{x_1}, \bar{x_2}$ = Averages of the two groups
  - $s_1^2 s_2^2$= How much the two groups vary
  - $n_1 n_2$ = Sample sizes of the two groups

Decision Rule: Look at the t value you got and the critical t value from the t-distribution table. If your t value is bigger than the critical one, throw out the null hypothesis.

Welch's t-Test:

- **Purpose:** Welch's t-test has an impact on adapting the independent samples t-test to compare two groups with unequal variances and unequal sample sizes. It stands stronger than the traditional independent t-test in these situations.

- **Hypotheses:**
  - **Null Hypothesis (H0):** The means of the two groups are equal ($\mu1=\mu2$).
  - **Alternative Hypothesis (Ha):** The means of the two groups differ ($\mu1\neq\mu2$).

- **Decision Rule:** Compare the t value you calculated to the critical t value from the t-distribution table. Use the degrees of freedom you worked out to find this. If your t value is bigger than the critical one, throw out the null hypothesis.

7. Chi-square Test

- The Chi-square test helps statisticians figure out if two types of data have a strong connection. People often use it to check theories and see how actual numbers compare to what they thought might happen.

Goodness of Fit Test

- **Definition:** The Goodness of Fit test checks if the observed frequency distribution of a categorical variable matches what we expect.
- **Purpose:** This test helps us figure out if a sample data set fits a population with a particular distribution.
- **Example:** Let's say you want to check if a six-sided die is fair. You'd compare how often each face comes up with how often you'd expect it to (which would be the same for all faces if the die is fair).
- **Formula:** $\chi2 = \Sigma (Oi - Ei)^2 / Ei$

In this equation, Oi stands for the observed frequency and Ei represents the expected frequency.
Test of Independence

- **Definition:** The Test of Independence checks if two categorical variables don't depend on each other.
- **Purpose:** It figures out if there's a link between the variables in the population the sample came from.
- **Example:** You could use a Chi-square Test of Independence to see if gender has an impact on how people vote in an election.
- **Contingency Table:** A grid that shows how often the variables occur together.
- **Formula:** $\chi2 = \Sigma (Oij - Eij)^2 / Eij$

Oij stands for the observed frequency in cell ij. Eij represents the expected frequency in cell ij.

Test of Homogeneity

- **Definition:** The Test of Homogeneity checks if different groups share the same spread for a category-based variable.
- **Purpose:** This test compares how category data spreads across several groups to see if they're alike.
- **Example:** You might use this test to check if customer happiness levels are similar across various store sites.
- **Contingency Table:** It's used like the Test of Independence, but it looks at how data spreads across different groups.
- **Formula:** The same math as the Test of Independence applies here.

Assumptions and Limitations

- **Assumptions:**
  - The data falls into categories.
  - Each observation stands on its own.
  - The sample is big enough to make sure each cell in a contingency table has an expected frequency of at least 5.

- **Limitations:**
  - You can't use the Chi-square test if the expected frequencies are too low (under 5), as this can lead to wrong results.
  - It shows if there's a big link but doesn't tell you how strong or which way the link goes.
  - It reacts to sample size; with a huge sample even a tiny difference can look like it matters a lot.

8. ANOVA Test

- ANOVA, which stands for Analysis of Variance, is a statistical technique that compares the averages of three or more groups to see if at least one group average differs from the rest. It builds on the t-test and researchers often use it in experiments to test ideas about differences between group averages.

One-Way ANOVA

- **Definition:** One-Way ANOVA tests the difference among the means of three or more independent groups based on a single factor (independent variable).
- **Purpose:** It helps determine if the means of different groups are equal.
- **Example:** Let's say a scientist wants to check how well three diets work for weight loss. The One-Way ANOVA would compare the average weight loss across these three diet groups.

- **Hypotheses:**
  - **Null Hypothesis (H0H_0H0):** All group means are equal.
- **Alternative Hypothesis (HaH_aHa):** One or more group means differ.
- **Formula:** F=Between-group varianceWithin-group varianceF = \frac{\text{Between-group variance}}{\text{Within-group variance}}F=Within-group varianceBetween-group varianceA high F-value points to a big difference between group means.

Two-Way ANOVA

- **Definition:** Two-Way ANOVA helps us understand how two independent factors affect the dependent variable. It also shows us how these factors interact with each other.
- **Purpose:** This method is helpful to see how two factors together influence the end result.
- **Example:** Let's say a scientist wants to know how diet and exercise together impact weight loss. The Two-Way ANOVA would look at the main effects of both diet and exercise, and also how they work together.
- **Hypotheses:**
  - **Null Hypothesis (H0):** The averages are the same across all levels of each factor, and the factors don't interact with each other.
  - **Alternative Hypothesis (Ha):** A significant main effect or interaction effect exists.
- **Interaction:** The interaction term in Two-Way ANOVA checks if one factor's effect changes based on the other factor's level.

Repeated Measures ANOVA

- **Definition:** Researchers use Repeated Measures ANOVA when they measure the same subjects under different conditions or at several points in time.
- **Purpose:** This method takes into account how repeated measurements on the same subjects relate to each other, which boosts statistical power.
- **Example:** A scientist might check patients' blood pressure before, during, and after treatment. Repeated Measures ANOVA would then compare the average blood pressure at these different times.
- **Hypotheses:**
  - **Null Hypothesis (H0):** The averages across the repeated measures are the same.
  - **Alternative Hypothesis (Ha):** At least one average differs from the others.

Assumptions of ANOVA

- **Normality:** Each group's data should follow a normal distribution more or less.
- **Homogeneity of Variances:** All groups should have the same variance (also called homoscedasticity).
- **Independence:** One observation shouldn't affect another.
- **Sphericity (for Repeated Measures ANOVA):** The differences between all related group pairs should have equal variances.

Post-hoc Tests

- **Definition:** Post-hoc tests happen after we find a significant ANOVA result to figure out which group means differ from each other.
- **Purpose:** They help control Type I errors when we make multiple comparisons.
- **Common Post-hoc Tests:**
  - **Tukey's HSD (Significant Difference):** This test looks at all possible pairs of group means to spot the differences.
  - **Bonferroni Correction:** This method tweaks the significance level to account for multiple comparisons.
  - **Scheffé Test:** This is a more careful approach that works for complex comparisons.
- **Example:** When an ANOVA shows that the means of four groups are different, researchers might use a Tukey's HSD test to figure out which specific pairs of groups are not the same.