Hypothesis Testing

1. **Introduction to Hypothesis Testing**

Hypothesis testing is a fundamental concept in statistical inference used to make decisions about populations based on sample data. It involves assuming about a population parameter and then using sample data to assess the plausibility of this assumption.

Key Components:

- Null Hypothesis (H_0): The initial assumption about a population parameter.
- Alternative Hypothesis (H_1 or H_a): The claim to be tested against the null hypothesis.
- Test Statistic: A value calculated from sample data used to make the decision.
- Significance Level (α): The probability of rejecting the null hypothesis when it's true.
- p-value: The probability of obtaining test results at least as extreme as the observed results, assuming the null hypothesis is true.

2. Steps in Hypothesis Testing

1. State the Hypotheses:

   Formulate H_0 and H_1 clearly and in terms of population parameters.

2. Choose the Significance Level (α):

   Common choices are 0.05, 0.01, or 0.1.

3. Select the Appropriate Test Statistic:

   Depends on the nature of the data and the hypothesis.

   Common test statistics: z-score, t-statistic, chi-square, F-statistic.

4. Determine the Critical Region:

   Based on α and the distribution of the test statistic.

5. Calculate the Test Statistic:

   Use sample data to compute the value of the chosen test statistic.

6. Make a Decision:

   Compare the calculated test statistic to the critical value(s).

   Alternatively, compare the p-value to α.

7. Draw a Conclusion:

   Interpret the results in the context of the original problem.

3. Types of Hypotheses

One-Tailed Tests:

- Right-tailed test: $H_1: \theta > \theta_0$
- Left-tailed test: $H_1: \theta < \theta_0$

Two-Tailed Tests:

- $H_1: \theta \neq \theta_0$

Where θ represents the population parameter and θ_0 is a specific value.

4. Types of Errors in Hypothesis Testing

- Type I Error: Rejecting H_0 when it's actually true. Probability = α
- Type II Error: Failing to reject H_0 when it's actually false. Probability = β
- Power of the Test: $1 - \beta$, the probability of correctly rejecting a false null hypothesis

5. Common Hypothesis Tests

1. Z-Test

Description: Used when the population standard deviation is known and the sample size is large ($n \geq 30$).

Example:

A factory claims that their light bulbs last an average of 1000 hours. A researcher tests 100 bulbs and finds a mean life of 950 hours. Assuming a known population standard deviation of 100 hours, test if the factory's claim is valid at a 5% significance level.

Steps:

1. $H_0: \mu = 1000$, $H_1: \mu \neq 1000$

2. $\alpha = 0.05$

3. $Z = (\bar{x} - \mu) / (\sigma / \sqrt{n}) = (950 - 1000) / (100 / \sqrt{100}) = -5$

4. Critical value for two-tailed test: ± 1.96

5. $|-5| > 1.96$, so reject H_0

Conclusion: There's significant evidence to doubt the factory's claim.

2. T-Test

Description: Used when the population standard deviation is unknown and/or the sample size is small ($n < 30$).

Example:

A new diet claims to reduce weight by 10 pounds on average. In a study of 25 participants, the mean weight loss was 8 pounds with a sample standard deviation of 4 pounds. Test if the diet's claim is supported at a 5% significance level.

Steps:

1. $H_0: \mu = 10$, $H_1: \mu \neq 10$

2. $\alpha = 0.05$, $df = 24$

3. $t = (\bar{x} - \mu) / (s / \sqrt{n}) = (8 - 10) / (4 / \sqrt{25}) = -2.5$

4. Critical value for two-tailed test with $df=24$: ± 2.064

5. $|-2.5| > 2.064$, so reject H_0

Conclusion: There's significant evidence to doubt the diet's claim.

3. Chi-Square Test

Description: Used to determine if there is a significant association between two categorical variables.

Example:

A study examines if there's a relationship between gender and preference for tea or coffee. The data collected:

	Tea	**Coffee**
Male	30	70
Female	50	50

Test if there's a significant association at a 5% significance level.

Steps:

1. H_0: No association between gender and drink preference

   H_1: There is an association

2. $\alpha = 0.05$, df = (r-1)(c-1) = 1

3. Calculate expected frequencies and chi-square statistic

4. $\chi^2 = 8.33$ (calculation details omitted for brevity)

5. Critical value for df=1, α=0.05: 3.841

6. 8.33 > 3.841, so reject H_0

Conclusion: There's a significant association between gender and drink preference.

4. F-Test

Description: Used to compare the variances of two populations.

Example:

Two manufacturing processes are being compared for consistency. Process A produces parts with a variance of 25 mm^2, while Process B has a variance of 16 mm^2. Each process was tested on 30 parts. Determine if there's a significant difference in variability at a 5% level.

Steps:

1. H_0: $\sigma^2A = \sigma^2B$, H_1: $\sigma^2A \neq \sigma^2B$

2. $\alpha = 0.05$, df1 = df2 = 29

3. $F = s^2A / s^2B = 25 / 16 = 1.5625$

4. Critical values for F(29,29) at α=0.05: 0.5017 and 1.9924

5. $0.5017 < 1.5625 < 1.9924$, so fail to reject H_0

Conclusion: There's no significant difference in variability between the two processes.

5. ANOVA (Analysis of Variance)

Description: Used to compare means of three or more groups.

Example:

A researcher wants to know if three different study methods (A, B, C) lead to different test scores. Each method is used by 20 students. The mean scores are: A=75, B=70, C=80.

Steps:

1. H_0: $\mu A = \mu B = \mu C$, H_1: At least one mean is different

2. $\alpha = 0.05$

3. Calculate F-statistic (details omitted for brevity)

4. If F > F-critical, reject H_0

Conclusion: If F > F-critical, conclude that at least one study method leads to significantly different test scores.

6. A/B Test

Description: Used to compare two versions of a single variable, typically for decision-making in business settings.

Example:

An e-commerce site tests two versions of a "Buy Now" button: red and green. They want to know which color leads to more conversions. 1000 visitors see each version.

Red button: 150 conversions

Green button: 180 conversions

Steps:

1. H_0: $p_r = p^g$, H_1: $p_r \neq p^g$
2. Use a two-proportion z-test
3. Calculate z-statistic and p-value
4. If p-value $< \alpha$ (usually 0.05), reject H_0

Conclusion: If p-value < 0.05, conclude that there's a significant difference in conversion rates between the two button colors.

Comparison of Tests

1. Z-Test vs T-Test:

- ✓ Z-test is used when population standard deviation is known; t-test when it's unknown.
- ✓ Z-test is typically used for larger samples ($n \geq 30$); t-test can be used for smaller samples.

2. Chi-Square vs T-Test:

- Chi-square test is for categorical data; t-test is for continuous data.
- Chi-square tests for independence or goodness of fit; t-test compares means.

3. ANOVA vs T-Test:

- ANOVA compares means of three or more groups; t-test compares means of two groups.
- ANOVA is an extension of the t-test for multiple groups.

4. F-Test vs ANOVA:

- F-test compares variances of two populations; ANOVA uses F-distribution to compare means of multiple groups.
- F-test is a component of ANOVA.

5. A/B Test vs Other Tests:

- A/B test is a practical application of statistical tests (often z-test or t-test) in business settings.
- It's focused on comparing two versions of a single variable, while other tests can have broader applications.

6. Assumptions and Conditions

Most parametric tests assume:

1. Independence of observations

2. Normality of the population distribution

3. Homogeneity of variances (for some tests)

It's crucial to check these assumptions before conducting a hypothesis test.

7. Interpreting p-values

- p-value $< \alpha$: Reject H_0

- p-value $\geq \alpha$: Fail to reject H_0

Remember: Failing to reject H_0 is not the same as accepting H_0.

8. Confidence Intervals and Hypothesis Testing

There's a close relationship between confidence intervals and two-tailed hypothesis tests:

- If a $(1-\alpha)100\%$ confidence interval doesn't contain the hypothesized value, we reject H_0 at significance level α.

9. Effect Size

While hypothesis testing tells us if there's a statistically significant difference, effect size measures tell us about the magnitude of that difference. Common effect size measures include:

- Cohen's d

- Pearson's r

- Odds ratio

10. Practical vs. Statistical Significance

A result can be statistically significant ($p < \alpha$) but not practically significant. Always consider the context and real-world implications of your findings.

11. Analysis of Variance (ANOVA)

ANOVA is a statistical method used to test differences between two or more means. It's an extension of the t-test to more than two groups.

Types of ANOVA:

1. One-way ANOVA: Compares means across one factor with multiple levels.

2. Two-way ANOVA: Examines the influence of two different factors on a dependent variable.

3. MANOVA (Multivariate ANOVA):

Tests for differences in two or more vectors of means.

Key Concepts in ANOVA:

- Factor: The independent variable being tested.

- Levels: The different categories or values of the factor.
- Between-group variability: Differences between the group means.
- Within-group variability: Differences within each group (error).

ANOVA Assumptions:

1. Independence of observations
2. Normality of residuals
3. Homogeneity of variances

Example of One-way ANOVA:

Suppose a researcher wants to test if there's a significant difference in the effectiveness of three different teaching methods on student test scores.

Null Hypothesis (H_0): $\mu_1 = \mu_2 = \mu_3$ (all group means are equal)

Alternative Hypothesis (H_1): At least one group mean is different

Data:

- Method A: 75, 82, 78, 80, 79
- Method B: 85, 87, 82, 89, 84
- Method C: 79, 81, 80, 83, 82

Using statistical software to perform ANOVA:

Source	SS	df	MS	F	p-value
Between	126.13	2	63.07	6.94	0.008
Within	109.20	12	9.10		
Total	235.33	14			

Interpretation: With a p-value of 0.008 < 0.05 (assuming $\alpha = 0.05$), we reject the null hypothesis. There is significant evidence to suggest that at least one teaching method leads to different test scores.

12. A/B Testing

A/B testing, also known as split testing, is a method of comparing two versions of a webpage, app interface, marketing email, or other product to determine which one performs better in terms of a specified metric.

Key Concepts:

1. Control (A): The current version or baseline.
2. Variant (B): The new version being tested.
3. Conversion Rate: The percentage of users who take a desired action.
4. Statistical Significance: Determined by p-value, usually with $\alpha = 0.05$.

Steps in A/B Testing:

1. Formulate a hypothesis
2. Determine the sample size
3. Run the experiment
4. Analyze results
5. Draw conclusions and implement changes

Statistical Analysis in A/B Testing:

Typically uses a z-test for proportions or a t-test for continuous metrics.

Example of A/B Test:

An e-commerce website wants to test if changing the color of the "Buy Now" button from blue to green increases the click-through rate (CTR).

Null Hypothesis (H_0): $p_1 = p_2$ (CTR is the same for both button colors)

Alternative Hypothesis (H_1): $p_1 \neq p_2$ (CTR is different between the two button colors)

Data:

- Control (Blue): 10,000 visitors, 300 clicks
- Variant (Green): 10,000 visitors, 360 clicks

Calculations:

- p_1 (Blue CTR) = 300 / 10,000 = 0.03 (3%)
- p_2 (Green CTR) = 360 / 10,000 = 0.036 (3.6%)
- Pooled proportion: $p = (300 + 360) / (10{,}000 + 10{,}000) = 0.033$
- Standard Error: $SE = \sqrt{[p(1-p)(1/n_1 + 1/n_2)]} = 0.00255$
- Z-score: $z = (p_2 - p_1) / SE = 2.35$

Using a z-table or calculator, we find that the p-value for this two-tailed test is approximately 0.019.

Interpretation: With a p-value of $0.019 < 0.05$, we reject the null hypothesis. There is significant evidence to suggest that changing the button color from blue to green affects the click-through rate. The green button appears to perform better, with a 20% relative increase in CTR (from 3% to 3.6%).

Remember that statistical significance doesn't always imply practical significance. The team should consider if a 0.6 percentage point increase in CTR is meaningful for their business before implementing the change.