exam

A comprehensive guide to hypothesis testing, covering z-tests, t-tests, ANOVA, correlation, and regression. Useful for quick reference and exam preparation.

Hypothesis Testing Fundamentals

Basic Concepts

Hypothesis: A statement about a population parameter.

Null Hypothesis (H_0): A statement of no effect or no difference; it is assumed to be true until evidence indicates otherwise.

Alternative Hypothesis (H_1 or H_a): A statement that contradicts the null hypothesis; it represents what we are trying to find evidence for.

Types of Tests:

- **One-Tailed (Right-Tailed):** H_1: $\mu >$ value
- **One-Tailed (Left-Tailed):** H_1: $\mu <$ value
- **Two-Tailed:** H_1: $\mu \neq$ value

Type I Error (α): Rejecting H_0 when it is actually true (False Positive).

Type II Error (β): Failing to reject H_0 when it is actually false (False Negative).

Significance Level (α): The probability of making a Type I error.

Power ($1-\beta$): The probability of correctly rejecting H_0 when it is false.

Steps in Hypothesis Testing

1. **State the Hypotheses:** Define H_0 and H_1.
2. **Determine the Test Statistic:** Choose the appropriate test statistic (z, t, F, etc.).
3. **Set the Significance Level:** Determine α (e.g., 0.05, 0.01).
4. **Calculate the Test Statistic:** Compute the value of the test statistic from the sample data.
5. **Determine the p-value or Critical Value:**
   - **p-value:** The probability of observing a test statistic as extreme as, or more extreme than, the one computed if H_0 is true.
   - **Critical Value:** The value(s) that define the rejection region.
6. **Make a Decision:**
   - **p-value Method:** If p-value $\leq \alpha$, reject H_0.
   - **Critical Value Method:** If the test statistic falls in the rejection region, reject H_0.
7. **State the Conclusion:** Interpret the decision in the context of the problem.

Z-Test vs T-Test

Z-Test

- Population standard deviation (σ) is known.
- Sample size is large ($n \geq 30$).
- Used for testing hypotheses about a single mean or comparing two means.

Test Statistic:

$z = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}}$

T-Test

- Population standard deviation (σ) is unknown.
- Sample size is small ($n < 30$).
- Used for testing hypotheses about a single mean or comparing two means.

Test Statistic:

$t = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}}$

Two-Sample Tests

Two-Sample Z-Test for Means

Purpose: To test if there is a significant difference between the means of two independent populations when the population standard deviations are known or sample sizes are large.

Hypotheses:

- H_0: $\mu_1 = \mu_2$
- H_1: $\mu_1 \neq \mu_2$ (two-tailed)
- H_1: $\mu_1 > \mu_2$ (right-tailed)
- H_1: $\mu_1 < \mu_2$ (left-tailed)

Test Statistic:

$$z = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

Where:

- $\bar{x}_1, \bar{x}_2$ are the sample means.
- μ_1, μ_2 are the population means.
- σ_1, σ_2 are the population standard deviations.
- n_1, n_2 are the sample sizes.

Example: Testing the difference in weight loss between a low-carb diet (n_1 = 80, $\bar{x}_1$ = 13.5 lbs, s_1 = 6.59 lbs) and a low-fat diet (n_2 = 76, $\bar{x}_2$ = 15.1 lbs, s_2 = 6.38 lbs) at $\alpha = 0.01$.

- H_0: $\mu_1 = \mu_2$
- H_1: $\mu_1 \neq \mu_2$

$$z = \frac{(13.5 - 15.1) - 0}{\sqrt{\frac{6.59^2}{80} + \frac{6.38^2}{76}}} \approx -1.54$$

Since $-2.58 < -1.54 < 2.58$, fail to reject H_0. There is not sufficient evidence to conclude that the mean weight loss differs.

Paired Sample T-Test

Purpose: To test if there is a significant difference between two related populations (e.g., before and after treatment).

Hypotheses:

- H_0: $\mu d = 0$
- H_1: $\mu d \neq 0$ (two-tailed)
- H_1: $\mu d > 0$ (right-tailed)
- H_1: $\mu d < 0$ (left-tailed)

Test Statistic:

$$t = \frac{\bar{d} - \mu_d}{\frac{s_d}{\sqrt{n}}}$$

Where:

- $\bar{d}$ is the mean of the differences.
- μ_d is the hypothesized mean difference (usually 0).
- s_d is the standard deviation of the differences.
- n is the number of pairs.

Example: Testing if runners ran faster after eating spaghetti ($n = 6$, $\bar{d} = 3.0$, $sd = 3.742$) at $\alpha = 0.05$.

- H_0: $\mu d = 0$
- H_1: $\mu d > 0$

$$t = \frac{3.0 - 0}{\frac{3.742}{\sqrt{6}}} \approx 1.963$$

Since $1.963 < 2.015$, fail to reject H_0. There is not sufficient evidence to support the claim that runners run faster after eating spaghetti.

Correlation and Regression

Correlation

Purpose: To measure the strength and direction of the linear relationship between two variables.

Pearson Correlation Coefficient (r):
$r = \frac{n(\sum xy) - (\sum x)(\sum y)}{\sqrt{[n(\sum x^2) - (\sum x)^2][n(\sum y^2) - (\sum y)^2]}}$

Interpretation:

- $-1 \le r \le 1$
- $r > 0$: Positive correlation
- $r < 0$: Negative correlation
- $r \approx 0$: No correlation
- $|r|$ close to 1: Strong correlation

Example: Calculating correlation between non-member price (x) and member price (y) for n = 8 pairs.

$r \approx 0.857$ indicates a strong positive correlation. As non-member price increases, member price tends to increase as well.

Linear Regression

Purpose: To model the relationship between two variables and predict the value of one variable based on the other.

Regression Line Equation:
$\hat{y} = a + bx$

Where:

- $\hat{y}$ is the predicted value of the dependent variable.
- x is the independent variable.
- a is the y-intercept.
- b is the slope.

Formulas for Slope (b) and Y-Intercept (a):
$b = \frac{n(\sum xy) - (\sum x)(\sum y)}{n(\sum x^2) - (\sum x)^2}$

$a = \bar{y} - b\bar{x}$

Example: Predicting number of mistakes (y) based on hours without sleep (x).

$\hat{y} = -2.769 + 0.288x$

For x = 40 hours:
$\hat{y} = -2.769 + 0.288(40) \approx 8.751 \approx 9$ mistakes.

Scatter Plots

Visual Representation: A scatter plot is a graph that displays the relationship between two variables.

Types of Relationships:

- **Positive Linear:** As x increases, y increases.
- **Negative Linear:** As x increases, y decreases.
- **No Relationship:** Points are scattered randomly.
- **Nonlinear:** Points follow a curved pattern.

Example: Plotting age (x) vs. income (y) data points on a scatter plot can reveal a positive linear relationship, indicating that income tends to increase with age.

Analysis of Variance (ANOVA)

ANOVA Fundamentals

Purpose: To test if there is a significant difference between the means of three or more independent groups.

Hypotheses:

- H_0: $\mu_1 = \mu_2 = \mu_3 = \ldots = \mu k$ (all means are equal)
- H_1: At least one mean is different

Test Statistic: F = MSbetween / MSwithin

Where:

- MSbetween is the mean square between groups.
- MSwithin is the mean square within groups.

Degrees of Freedom:

- dfNumerator = k - 1 (k is the number of groups)
- dfDenominator = N - k (N is the total number of observations)

F-Distribution:

- Always right-tailed.
- Critical value is found using α, dfNumerator, and dfDenominator.

Sums of Squares:

- SSTotal = $\sum x^2 - (\sum x)^2 / N$
- SSBetween = $\sum ni\ (x\hat{\imath} - \bar{X}GM)^2$
- SSWithin = SSTotal - SSBetween

ANOVA Calculation Steps

1. **Calculate the Group Sums, Means, and Total Mean.**
2. **Calculate the Sum of Squares Total (SSTotal).**
3. **Calculate the Sum of Squares Between (SSBetween).**
4. **Calculate the Sum of Squares Within (SSWithin).**
5. **Calculate the Mean Squares:**
   - MSBetween = SSBetween / dfNumerator
   - MSWithin = SSWithin / dfDenominator
6. **Calculate the F statistic:** F = MSBetween / MSWithin
7. **Find the Critical Value:** Using α, dfNumerator, and dfDenominator.
8. **Make a Decision:** If F > Fcrit, reject H_0.
9. **State the Conclusion.**

Example

Example: Testing if average number of students in an English course differs by time of day (Morning, Afternoon, Evening) at α = 0.05.

F ≈ 4.776, Fcrit = 3.89. Since 4.776 > 3.89, reject H_0. There is sufficient evidence to conclude that mean attendance differs by time of day.