# 4 Statistical Methods

## 4.1 Choosing the Right Test

Different metrics require different statistical tests. The choice depends on:

1. **Metric type**: Binary, continuous, count, or ordinal
2. **Data distribution**: Normal, skewed, heavy-tailed
3. **Sample size**: Large (CLT applies) vs small
4. **Variance equality**: Equal vs unequal variances

**Note**

**Decision Tree for Test Selection**
**Step 1: Identify metric type**

- **Binary** (0/1, yes/no): Conversion, click, bounce → Proportion test (Section 4.2)
- **Continuous** (real numbers): Revenue, time on site, cart value → t-test or Mann-Whitney (Section 4.3)
- **Count** (non-negative integers): Page views, items purchased → Poisson test or Mann-Whitney (Section 4.4)
- **Categorical** (multiple categories): Product category, exit page → Chi-square test (Section 4.5)

**Step 2: Check distributional assumptions**

- If continuous and approximately normal: Use t-test (parametric)
- If continuous and heavily skewed: Use Mann-Whitney U (non-parametric)
- If count data: Consider distribution (Poisson vs negative binomial)

**Step 3: Select specific test**

- Two variants: Two-sample test
- $>2$ variants: ANOVA or Kruskal-Wallis

**Implementation**: See `analyze_metric()` in `statistical_analysis.py` for automatic test selection.

## 4.2 Two-Sample Proportion Test

**Use For**: Binary metrics (conversion rate, click-through rate, bounce rate)
**Key References**:

- Wald, A. (1943). "Tests of Statistical Hypotheses Concerning Several Parameters." *Transactions of the American Mathematical Society*, 54(3), 426–482.

### 4.2.1 The Test Statistic

For proportions $\hat{p}_1$ (treatment) and $\hat{p}_2$ (control):

$$Z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}(1-\hat{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \tag{17}$$

where $\hat{p} = \frac{x_1+x_2}{n_1+n_2}$ is the pooled proportion.
Under $H_0 : p_1 = p_2$, we have $Z \sim N(0,1)$ approximately (by CLT for large samples).
**Decision Rule**:

- Two-tailed test: Reject $H_0$ if $|Z| > Z_{\alpha/2}$ (e.g., $|Z| > 1.96$ for $\alpha = 0.05$)
- P-value: $p = 2 \times P(Z > |z_{\text{obs}}|)$

#### 4.2.2 Confidence Interval

95% CI for difference in proportions:

$$(\hat{p}_1 - \hat{p}_2) \pm 1.96 \times \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}} \tag{18}$$

**Note**

**Example: Conversion Rate Test**
**Data**:

- Control: 500 conversions out of 10,000 users ($\hat{p}_2 = 0.05$)
- Treatment: 575 conversions out of 10,000 users ($\hat{p}_1 = 0.0575$)

**Pooled proportion**:

$$\hat{p} = \frac{500 + 575}{10000 + 10000} = \frac{1075}{20000} = 0.05375$$

**Test statistic**:

$$\begin{aligned} Z &= \frac{0.0575 - 0.05}{\sqrt{0.05375(0.94625)\left(\frac{1}{10000} + \frac{1}{10000}\right)}} \\ &= \frac{0.0075}{\sqrt{0.050896 \times 0.0002}} \\ &= \frac{0.0075}{0.003193} = 2.348 \end{aligned}$$

**P-value**: $p = 2 \times P(Z > 2.348) \approx 0.019$
**Conclusion**: $p < 0.05 \Rightarrow$ Reject $H_0$, significant difference detected
**Effect size**: Relative lift $= (0.0575 - 0.05)/0.05 = 15\%$
**Implementation**: See `proportion_test()` in `statistical_analysis.py`

### 4.3 Two-Sample T-Test

**Use For**: Continuous metrics (revenue, time on site, cart value)
**Key References**:

- Student (W. S. Gosset). (1908). "The Probable Error of a Mean." *Biometrika*, 6(1), 1–25.
- Welch, B. L. (1947). "The Generalization of Student's Problem when Several Different Population Variances are Involved." *Biometrika*, 34(1/2), 28–35.

### 4.3.1 Welch's T-Test (Recommended)

**Why Welch's t-test?** It does NOT assume equal variances (robust to heteroscedasticity).

**Test statistic**:

$$t = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \tag{19}$$

**Degrees of freedom** (Welch-Satterthwaite approximation):

$$\nu = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{(s_1^2/n_1)^2}{n_1-1} + \frac{(s_2^2/n_2)^2}{n_2-1}} \tag{20}$$

**Decision Rule**:

- Reject $H_0$ if $|t| > t_{\alpha/2,\nu}$
- P-value from $t$-distribution with $\nu$ degrees of freedom

**Note**

**When Assumptions Fail: Robustness of T-Test**
The t-test assumes:

1. Independence (satisfied by randomization)
2. Normality of sampling distribution
3. Equal variances (not required for Welch's version)

**Good News**: Thanks to the Central Limit Theorem (CLT), the t-test is robust to non-normality when:

- Sample sizes are large ($n > 30$ per group)
- Distributions aren't extremely skewed

**E-Commerce Reality**: Revenue data is often:

- Right-skewed (many small purchases, few large ones)
- Heavy-tailed (outliers exist)
- Zero-inflated (many users don't convert)

**Solutions**:

- For moderate skewness + large samples: Welch's t-test is still valid
- For severe skewness: Use Mann-Whitney U test (non-parametric, Section 4.4)
- For revenue: Consider winsorizing outliers or log-transformation

**Implementation**: See `continuous_metric_test()` in `statistical_analysis.py`

## 4.4 Mann-Whitney U Test (Non-Parametric)

**Use For**: Continuous or ordinal data when:

- Distributions are heavily skewed

- Outliers are present
- Sample sizes are small
- Normality assumption violated

**Key References**:

- Mann, H. B., & Whitney, D. R. (1947). "On a Test of Whether one of Two Random Variables is Stochastically Larger than the Other." *Annals of Mathematical Statistics*, 18(1), 50–60.
- Wilcoxon, F. (1945). "Individual Comparisons by Ranking Methods." *Biometrics Bulletin*, 1(6), 80–83.

### 4.4.1 How It Works

Instead of comparing means, Mann-Whitney compares *ranks*:

1. Pool all observations from both groups
2. Rank them from smallest to largest
3. Sum ranks for each group
4. Test if rank sums are significantly different

**Null Hypothesis**: The two distributions are identical
**Alternative**: One distribution is stochastically larger (higher values more likely)
**Test Statistic**:

$$U = n_1 n_2 + \frac{n_1(n_1 + 1)}{2} - R_1 \tag{21}$$

where $R_1$ is sum of ranks for group 1.

**Note**

**Example: Revenue Test with Skewed Data**
**Scenario**: Revenue data with outliers
**Control** (10 users): \$0, \$0, \$25, \$30, \$35, \$40, \$45, \$50, \$55, \$500

- Mean = \$78
- Median = \$37.50

**Treatment** (10 users): \$0, \$20, \$40, \$45, \$50, \$55, \$60, \$65, \$70, \$75

- Mean = \$48
- Median = \$52.50

**Problem with t-test**: Control mean (\$78) is inflated by \$500 outlier, would incorrectly suggest control is better!
**Mann-Whitney Approach**:

- Ranks all 20 values
- Compares rank sums (robust to outliers)
- Correctly identifies treatment as better (higher median, more consistent performance)

**Implementation**: See `mann_whitney_test()` in `statistical_analysis.py`

## 4.5 ANOVA (Multiple Variant Tests)

**Use For**: Comparing > 2 variants on continuous metric
**Key References**:

- Fisher, R. A. (1925). *Statistical Methods for Research Workers*. Oliver and Boyd.

### 4.5.1 One-Way ANOVA

Tests: $H_0 : \mu_1 = \mu_2 = \cdots = \mu_k$ (all group means equal)
**Test Statistic** (F-ratio):

$$F = \frac{\text{Between-group variance}}{\text{Within-group variance}} = \frac{MS_{\text{between}}}{MS_{\text{within}}} \tag{22}$$

**Decomposition of variance**:

$$SS_{\text{total}} = SS_{\text{between}} + SS_{\text{within}} \tag{23}$$

$$SS_{\text{between}} = \sum_{i=1}^{k} n_i(\bar{X}_i - \bar{X})^2 \tag{24}$$

$$SS_{\text{within}} = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(X_{ij} - \bar{X}_i)^2 \tag{25}$$

If $F > F_{\alpha,k-1,N-k}$, reject $H_0$.

### 4.5.2 Post-Hoc Tests

If ANOVA rejects $H_0$ (at least one group differs), use post-hoc tests to identify which pairs differ:

**Options**:

- **Tukey HSD**: Controls family-wise error rate, best for all pairwise comparisons
- **Bonferroni**: Very conservative, simple
- **Dunnett**: Compares all treatments to single control

**Note**

**Example: 3-Variant Product Slider Test**
**Variants**:

- A: Social proof only
- B: Similar products
- C: Hybrid approach

**Metric**: Revenue per user
**ANOVA Result**: $F(2, 17997) = 12.5$, $p < 0.001$
**Interpretation**: At least one variant differs significantly
**Post-Hoc (Tukey HSD)**:

- A vs B: $p = 0.023$ (significant)
- A vs C: $p = 0.001$ (significant)
- B vs C: $p = 0.412$ (not significant)

**Business Conclusion**: C (hybrid) performs significantly better than A, similar to B. Deploy C.
**Implementation**: See `anova_test()` in `statistical_analysis.py`

## 4.6 Chi-Square Test for Independence

**Use For**: Categorical outcomes (e.g., exit page, product category chosen)

**Key References**:

- Pearson, K. (1900). "On the Criterion that a Given System of Deviations." *Philosophical Magazine*, 50, 157–175.

### 4.6.1 Contingency Table Analysis

Tests independence of two categorical variables.

**Example**: Does variant affect which product category users browse?

| | Electronics | Clothing | Home | Total |
|---|---|---|---|---|
| Control | 150 | 200 | 150 | 500 |
| Treatment | 180 | 180 | 140 | 500 |
| Total | 330 | 380 | 290 | 1000 |

**Expected Counts** (under independence):

$$E_{ij} = \frac{(\text{Row}_i \text{ Total}) \times (\text{Column}_j \text{ Total})}{\text{Grand Total}} \tag{26}$$

**Test Statistic**:

$$\chi^2 = \sum_{i,j} \frac{(O_{ij} - E_{ij})^2}{E_{ij}} \tag{27}$$

Degrees of freedom: $(r-1)(c-1)$ where $r$ = rows, $c$ = columns.

**Note**

**Implementation**
See `chi_square_test()` in `statistical_analysis.py`
**Cramér's V** (effect size for chi-square):

$$V = \sqrt{\frac{\chi^2}{n \times \min(r-1, c-1)}} \tag{28}$$

**Interpretation**:

- $V < 0.1$: Weak association
- $0.1 \leq V < 0.3$: Moderate association
- $V \geq 0.3$: Strong association

## 4.7 Multiple Testing Correction

**The Problem**: When testing multiple metrics, false positive rate increases!
**Example**: Testing 10 independent metrics at $\alpha = 0.05$:

- Probability of NO false positives: $(1 - 0.05)^{10} = 0.599$
- Probability of $\geq 1$ false positive: $1 - 0.599 = 0.401$ (40%!)

**Key References**:

- Bonferroni, C. (1936). *Teoria statistica delle classi e calcolo delle probabilità.*
- Holm, S. (1979). "A Simple Sequentially Rejective Multiple Test Procedure." *Scandinavian Journal of Statistics*, 6(2), 65–70.
- Benjamini, Y., & Hochberg, Y. (1995). "Controlling the False Discovery Rate." *Journal of the Royal Statistical Society B*, 57(1), 289–300.

### 4.7.1 Correction Methods

**1. Bonferroni Correction** (Most Conservative)
Adjust significance level: $\alpha_{\text{adjusted}} = \frac{\alpha}{m}$
For $m = 10$ tests and $\alpha = 0.05$: Use $\alpha = 0.005$ for each test.
**Pros**: Simple, controls family-wise error rate (FWER)
**Cons**: Very conservative (low power) when many tests
**2. Holm-Bonferroni** (Recommended for 5–10 tests)
Sequential procedure:

1. Order p-values: $p_{(1)} \leq p_{(2)} \leq \cdots \leq p_{(m)}$
2. Test sequentially:
   - Compare $p_{(1)}$ to $\alpha/m$
   - Compare $p_{(2)}$ to $\alpha/(m-1)$
   - Continue until first non-rejection

**Pros**: More powerful than Bonferroni, still controls FWER
**3. Benjamini-Hochberg FDR** (For $> 10$ tests)
Controls False Discovery Rate (proportion of false positives among rejections).
Procedure:

1. Order p-values: $p_{(1)} \leq \cdots \leq p_{(m)}$
2. Find largest $k$ where: $p_{(k)} \leq \frac{k}{m} \times \alpha$
3. Reject $H_0$ for all $i \leq k$

**Pros**: More powerful for many tests
**Cons**: Allows some false positives (by design)

**Note**

**Which Method to Use?**
**Our 5 E-Commerce Tests**: We analyze 5–10 metrics per test
**Recommendation**: Use **Holm-Bonferroni**

- Appropriate for 5–10 tests
- Controls FWER (no false positives)
- More powerful than Bonferroni

**Alternative**: If testing 20+ metrics (e.g., full metric suite), use **Benjamini-Hochberg FDR**
**Implementation**: See `multiple_testing_correction()` in `validation.py`
**Example Usage**:

```
# Test 5 metrics, get p-values
pvalues = [0.023, 0.041, 0.087, 0.012, 0.156]

# Apply Holm correction
result = validator.multiple_testing_correction(
    pvalues,
    method='holm',
    alpha=0.05
)

# Result shows which tests remain significant
# after correction
```