Mathematics for Machine Learning

MAD-B3-2526-S2-MAT0611

Hypothesis Testing II: Two-Sample & Nonparametric Tests

Agenda

1. Two-Sample t-tests
2. Paired t-tests
3. Mann-Whitney U Test
4. Kolmogorov-Smirnov Test
5. The Chi-Square Distribution
6. Chi-Square Tests
    - Goodness of Fit
    - Test of Independence
    - Test of Homogeneity
7. Likelihood Ratio Tests

Two-Sample t-test: Independent Samples

Research Question

Are the means of two independent populations different?

Example: Do students in Group A score differently than students in Group B?

Hypotheses:

$$H_0 : \mu_1 = \mu_2 \quad \text{vs.} \quad H_1 : \mu_1 \neq \mu_2$$

Equivalently:

$$H_0 : \mu_1 - \mu_2 = 0 \quad \text{vs.} \quad H_1 : \mu_1 - \mu_2 \neq 0$$

Two-Sample t-test: Assumptions

Key Assumptions:

1. **Independence**: Observations within and between samples are independent
2. **Normality**: Each population is normally distributed (or n is large enough for CLT)
3. **Equal Variances** (for standard version): $\sigma_1^2 = \sigma_2^2$

Relaxations:

- **Welch's t-test**: Does not assume equal variances (more robust)
- **Large sample sizes**: CLT makes normality less critical

Test Statistic: Equal Variances

When we assume $\sigma_1^2 = \sigma_2^2$:

$$t = \frac{\bar{X}_1 - \bar{X}_2}{s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

Pooled standard deviation:

$$s_p = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}}$$

Under H_0: $t \sim t_{n_1+n_2-2}$

Test Statistic: Unequal Variances (Welch's t-test)

When we do **not** assume equal variances:

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

Degrees of freedom (Welch-Satterthwaite equation):

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

Recommendation: Use Welch's t-test as default (more robust)

Paired t-test

When to Use

Use when measurements are **paired** or **matched**:

- Before and after measurements on the same subjects
- Matched pairs (e.g., twins, matched controls)
- Repeated measurements under different conditions

Key difference: Accounts for correlation between pairs

Hypotheses:

$$H_0 : \mu_D = 0 \quad \text{vs.} \quad H_1 : \mu_D \neq 0$$

where $D_i = X_{1i} - X_{2i}$ are the paired differences

Paired t-test: Test Statistic

Compute differences: $D_i = X_{1i} - X_{2i}$ for $i = 1, \ldots, n$

$$t = \frac{\bar{D}}{s_D/\sqrt{n}}$$

where:

- $\bar{D} = \frac{1}{n}\sum_{i=1}^{n} D_i$ is the mean difference
- $s_D = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(D_i - \bar{D})^2}$ is the standard deviation of differences

Under H_0: $t \sim t_{n-1}$

Note: This is just a one-sample t-test on the differences

Paired vs. Unpaired t-test

Why is pairing important?

Pairing typically **increases statistical power** by:

- Removing between-subject variability
- Focusing on within-subject changes

Example: Testing a weight loss program

- **Unpaired**: Compare weights of two different groups (treatment vs. control)
  - High variability due to natural differences between people
- **Paired**: Compare weights before and after treatment in the same people
  - Lower variability because we measure change within each person

If data can be paired, use paired test.

Nonparametric Tests: Introduction

Use nonparametric tests when:

1. **Normality assumption is violated** (and sample size is small)
2. **Data are ordinal** (ranks, ratings)
3. **Outliers are present** (nonparametric tests are more robust)
4. **Distribution is unknown** or heavily skewed

Key Features:

- Distribution-free: **no parameters to estimate**, fewer assumptions
- Often based on ranks or counts rather than actual values
- Generally less powerful than parametric tests when assumptions are met
- More robust to outliers and violations of assumptions

Mann-Whitney U Test (Wilcoxon Rank-Sum Test)

Nonparametric Alternative to Independent Two-Sample t-test

Research Question: Do two independent samples come from the same distribution?

Hypotheses:

$$H_0 : \text{Distributions are identical} \quad \text{vs.} \quad H_1 : \text{Distributions differ}$$

Procedure:

1. Combine all observations from both groups
2. Rank all observations from smallest to largest
3. Sum ranks for each group (R_1, R_2)
4. Compute U statistic

Mann-Whitney U Statistic

$$U_1 = n_1 n_2 + \frac{n_1(n_1+1)}{2} - R_1$$

$$U_2 = n_1 n_2 + \frac{n_2(n_2+1)}{2} - R_2$$

Take $U = \min(U_1, U_2)$

For large samples ($n_1, n_2 > 20$), the test statistic is approximately normal:

$$z = \frac{U - \mu_U}{\sigma_U}$$

where $\mu_U = \frac{n_1 n_2}{2}$ and $\sigma_U = \sqrt{\frac{n_1 n_2 (n_1+n_2+1)}{12}}$

Kolmogorov-Smirnov Test

Testing Distribution Equality

Purpose: Nonparametric test to compare:

1. **One-sample KS**: Does a sample follow a specified continuous distribution?
2. **Two-sample KS**: Do two samples come from the same distribution?

Advantages:

- No assumptions about distribution shape
- Tests entire distribution (not just location/scale)
- Works with continuous data

Key Idea: Compares empirical cumulative distribution functions (ECDFs)

Empirical Cumulative Distribution Function (ECDF)

Definition: For a sample $X_1, \ldots, X_n$, the ECDF is:

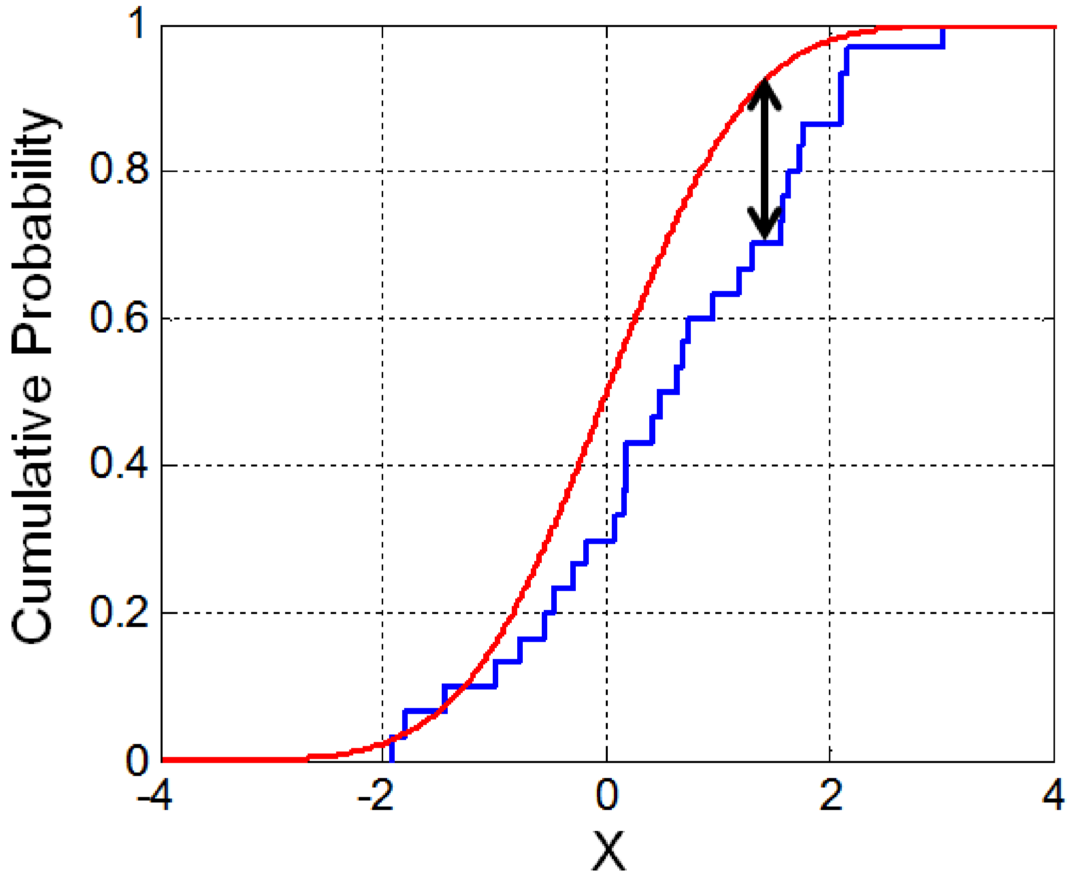
$$\hat{F}_n(x) = \frac{1}{n} \sum_{i=1}^{n} 1(X_i \leq x)$$

where $1(X_i \leq x)$ is 1 if $X_i \leq x$ and 0 otherwise.

Properties:

- Step function that jumps by $1/n$ at each observation
- $\hat{F}_n(x) \to F(x)$ as $n \to \infty$ (fundamental theorem of statistics)
- At each data point, equals the proportion of data $\leq$ that point

KS Test: Visual Example



The KS statistic D is the maximum vertical distance between the two cumulative distribution functions.

In this example:

- Red line: theoretical CDF
- Blue line: empirical CDF
- Green arrow: KS statistic D

The test measures how far the observed data deviates from the expected distribution.

One-Sample Kolmogorov-Smirnov Test

Research Question: Does a sample follow a specified distribution F_0?

Hypotheses:

$$H_0 : F = F_0 \quad \text{vs.} \quad H_1 : F \neq F_0$$

Test Statistic:

$$D_n = \sup_x \left|\hat{F}_n(x) - F_0(x)\right|$$

where $\sup$ denotes the supremum (maximum vertical distance)

Interpretation: Largest absolute difference between empirical and theoretical CDFs

Two-Sample Kolmogorov-Smirnov Test

Research Question: Do two samples come from the same **continuous** distribution?

Hypotheses:

$$H_0 : F_1 = F_2 \quad \text{vs.} \quad H_1 : F_1 \neq F_2$$

Test Statistic:

$$D_{n,m} = \sup_x \left|\hat{F}_n(x) - \hat{F}_m(x)\right|$$

where $\hat{F}_n$ and $\hat{F}_m$ are the ECDFs of the two samples.

Properties:

- Sensitive to differences in location, scale, and shape
- More powerful than Mann-Whitney when distributions differ in shape

KS Test: Distribution and Critical Values

Kolmogorov Distribution: Under H_0, for large n:

$$P(\sqrt{n}D_n \leq x) \to K(x) = 1 - 2\sum_{k=1}^{\infty}(-1)^{k-1}e^{-2k^2x^2}$$

Critical Values (one-sample, $\alpha = 0.05$):

- $n = 20$: $D_{crit} \approx 0.294$
- $n = 50$: $D_{crit} \approx 0.188$
- Large n: $D_{crit} \approx 1.36/\sqrt{n}$

For two-sample test with sizes n and m:

$$D_{crit} \approx c(\alpha)\sqrt{\frac{n+m}{nm}}$$

When to Use KS vs Other Tests

Use Kolmogorov-Smirnov when:

- Need to test entire distribution (not just means)
- Distribution shape is unknown
- Want to detect any type of difference (location, scale, shape)
- Data is continuous

Use other tests when:

- Comparing only location (means/medians): Use t-test or Mann-Whitney
- Categorical data: Use chi-square tests
- Small sample sizes with specific alternatives: KS has lower power
- Data has ties: KS assumes continuous distributions

The Chi-Square Distribution

Origin and Definition

Let $Z_1, Z_2, \ldots, Z_k$ be independent standard normal random variables: $Z_i \sim N(0, 1)$

The **chi-square distribution** with k degrees of freedom is:

$$\chi_k^2 = Z_1^2 + Z_2^2 + \cdots + Z_k^2 = \sum_{i=1}^{k} Z_i^2$$

Properties:

- Always non-negative ($\chi^2 \geq 0$)
- Right-skewed for small k
- Approaches normal distribution as k increases

Probability Density Function

The χ_k^2 is a special case of the gamma distribution with shape $\alpha = k/2$ and scale $\beta = 2$.

For χ_k^2 distribution with k degrees of freedom:

$$f(x;k) = \frac{1}{2^{k/2}\Gamma(k/2)} x^{k/2-1} e^{-x/2}, \quad x > 0$$

where Γ is the gamma function.

Properties

- Mean: $E[\chi_k^2] = k$
- Variance: $\mathrm{Var}(\chi_k^2) = 2k$

You can use Geogebra for interactive exploration of the chi-square distribution.

Why Chi-Square Appears Everywhere

Connection to Normal Distribution

When we have normally distributed data and compute sample variance or standardized squared deviations, chi-square naturally appears.

Sample Variance Result: If $X_1, \ldots, X_n \sim N(\mu, \sigma^2)$ independently, then:

$$\frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{n-1}$$

where $S^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X})^2$

Standardized Squared Deviations: If we standardize observations and square them:

$$\sum_{i=1}^{n} \left(\frac{X_i - \mu}{\sigma} \right)^2 \sim \chi^2_n$$

Chi-Square Test Statistic

$$\chi^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i}$$

where:

- O_i = observed frequency in category i
- E_i = expected frequency in category i
- k = number of categories

Under H_0: $\chi^2 \sim \chi^2_{k-1-p}$ where p is the number of estimated parameters

Assumptions:

- Minimum expected frequency for all categories (typically $E_i \geq 5$)
- Independent observations

Chi-Square Test: Goodness of Fit

Testing Distribution Fit

Research Question: Does observed data follow a specified distribution?

Example: A die is rolled 60 times. Is it fair?

Outcome	1	2	3	4	5	6
Observed	8	11	9	12	10	10
Expected	10	10	10	10	10	10

Hypotheses:

$$H_0 : \text{Die is fair} \quad \text{vs.} \quad H_1 : \text{Die is not fair}$$

Chi-Square in Goodness of Fit

Why it works: When testing categorical data, under H_0:

$$\frac{O_i - E_i}{\sqrt{E_i}} \approx N(0,1) \quad \text{(by CLT for counts)}$$

Therefore:

$$\chi^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i} = \sum_{i=1}^{k} \left(\frac{O_i - E_i}{\sqrt{E_i}} \right)^2 \approx \chi^2_k$$

$k - p - 1$ **Degrees of Freedom**:

- Start with k categories
- Subtract 1 for the constraint $\sum O_i = \sum E_i = n$
- Subtract 1 for each estimated parameter (p in total)

Chi-Square in Contingency Tables

Under H_0 (independence), each standardized residual:

$$\frac{O_{ij} - E_{ij}}{\sqrt{E_{ij}}} \approx N(0,1)$$

Summing squared residuals:

$$\chi^2 = \sum_{i=1}^{r}\sum_{j=1}^{c}\frac{(O_{ij} - E_{ij})^2}{E_{ij}} \sim \chi^2_{(r-1)(c-1)}$$

Degrees of Freedom for a $r \times c$ table (independence test):

- $r \times c$ cells
- Constraints: r row totals + c column totals - 1 (grand total counted twice)
- $rc - r - c + 1 = (r-1)(c-1)$

Chi-Square Test of Independence

Testing Association Between Two Categorical Variables

Research Question: Are two categorical variables independent?

Example: Is smoking status independent of lung disease?

	Disease	No Disease	Total
Smoker	50	100	150
Non-smoker	20	130	150
Total	70	230	300

Hypotheses:

$$H_0 : \text{Variables are independent} \quad \text{vs.} \quad H_1 : \text{Variables are dependent}$$

Expected Frequencies Under Independence

For each cell (i, j):

$$E_{ij} = \frac{(\text{row } i \text{ total}) \times (\text{column } j \text{ total})}{\text{grand total}}$$

Example:

$$E_{11} = \frac{150 \times 70}{300} = 35$$

Test Statistic:

$$\chi^2 = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

Under H_0: $\chi^2 \sim \chi^2_{(r-1)(c-1)}$

Chi-Square Test of Homogeneity

Testing if Multiple Populations Have Same Distribution

Example: Do three different regions have the same political preference distribution?

	Party A	Party B	Party C	Total
Region 1	45	55	30	130
Region 2	60	40	50	150
Region 3	50	65	35	150
Total	155	160	115	430

$$H_0 : \text{All populations have identical distributions}$$

$$H_1 : \text{At least one population differs}$$

Test of Homogeneity vs. Test of Independence

Test of Independence:

- **One sample** from a single population
- Classify by two variables
- Question: Are the two variables associated?

Test of Homogeneity:

- **Multiple samples** from different populations
- Compare distribution across populations
- Question: Do populations have the same distribution?

Mathematical Equivalence: Both have the same formula and distribution

Likelihood Ratio Tests (LRT)

General Framework for Hypothesis Testing

Idea: Compare how well two models fit the data

- **Null model**: Restricted model under H_0
- **Alternative model**: More general model

Likelihood Ratio:

$$\Lambda = \frac{L(\hat{\theta}_0|\text{data})}{L(\hat{\theta}_1|\text{data})} = \frac{\max_{\theta \in \Theta_0} L(\theta)}{\max_{\theta \in \Theta} L(\theta)}$$

- Numerator: maximum likelihood under H_0 (restricted)
- Denominator: maximum likelihood under H_1 (unrestricted)

LRT Test Statistic

Test Statistic (log-likelihood ratio):

$$G = -2\log(\Lambda) = 2[\ell(\hat{\theta}_1) - \ell(\hat{\theta}_0)]$$

where ℓ is the log-likelihood.

Wilks' Theorem: Under H_0 and regularity conditions:

$$G \sim \chi^2_d$$

where d is the difference in number of parameters between the models.

Decision Rule: Reject H_0 if $G > \chi^2_{d,\alpha}$ (critical value)

LRT in Regression Models

Testing Nested Models

Setup: Compare two nested regression models

- **Reduced model**: $Y = \beta_0 + \beta_1 X_1 + \epsilon$
- **Full model**: $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \epsilon$

Hypotheses:

$$H_0 : \beta_2 = \beta_3 = 0 \quad \text{vs.} \quad H_1 : \text{at least one } \beta_j \neq 0$$

Test Statistic:

$$G = 2[\ell(\text{full}) - \ell(\text{reduced})] \sim \chi^2_d$$

where d is the number of additional parameters in the full model.

Additional Resources

Books:

- Wasserman: *All of Statistics* (Ch. 10-11)
- Agresti: *Categorical Data Analysis*
- Casella & Berger: *Statistical Inference* (Ch. 8-9)
- Lehmann & Romano: *Testing Statistical Hypotheses*

Online Resources:

- scipy.stats documentation
- statsmodels documentation
- Penn State STAT 415: Introduction to Mathematical Statistics