PODS Lab 10: Non-Parametric Significance Tests

Hamza Alshamy

Center for Data Science, NYU
ha2486@nyu.edu

04/04/25

1. Chi-squared Test (χ^2)
   - Goodness-of-fit
   - Test of independence
2. Mann-Whitney U Test (AKA Wilcoxon Rank-sum Test)
3. Kolmogorov-Smirnov (KS) Test
4. Permutation Test
5. **Quiz + Discussion**

Parametric vs. Non-Parametric

- **Parametric Tests:** Statistical tests that assume the data follows a specific probability distribution (e.g., Normal distribution).
  - **Used:** When data can be reduced to sample means.
  - The underlying assumption is that the **mean** is a sufficient statistic of the data distribution.
  - **Examples:** Z-tests, t-tests, ANOVA.

Parametric vs. Non-Parametric

- **Parametric Tests:** Statistical tests that assume the data follows a specific probability distribution (e.g., Normal distribution).
  - **Used:** When data can be reduced to sample means.
  - The underlying assumption is that the **mean** is a sufficient statistic of the data distribution.
  - **Examples:** Z-tests, t-tests, ANOVA.

- **Non-Parametric Tests:** Statistical tests that **do not assume** the data follows a specific distribution or when the distribution is unknown.
  - **Used:** When the data cannot be reduced to sample means or when the use of the mean is not appropriate or robust.
  - **Instead,** focus on ranks, medians, and distribution-free metrics.
  - **Examples:** Chi-squared Test, Mann-Whitney U Test, Kolmogorov-Smirnov (KS) Test, Permutation Test.

Chi-squared Test (χ^2)

A significance test for categorical data

- **Purpose:** To assess whether observed frequencies differ from expected frequencies under a null hypothesis.
- **Use Cases:**
  1. **Goodness-of-fit Test:** Comparing observed categorical data to a theoretical distribution.
     - **E.g.** Zodiac signs and serial killer counts
  2. **Test of Independence:** Testing for association between two categorical variables.
     - **E.g.** Preferences for pineapple on pizza across generations

Chi-squared (χ^2) Steps

χ^2 Test Statistic

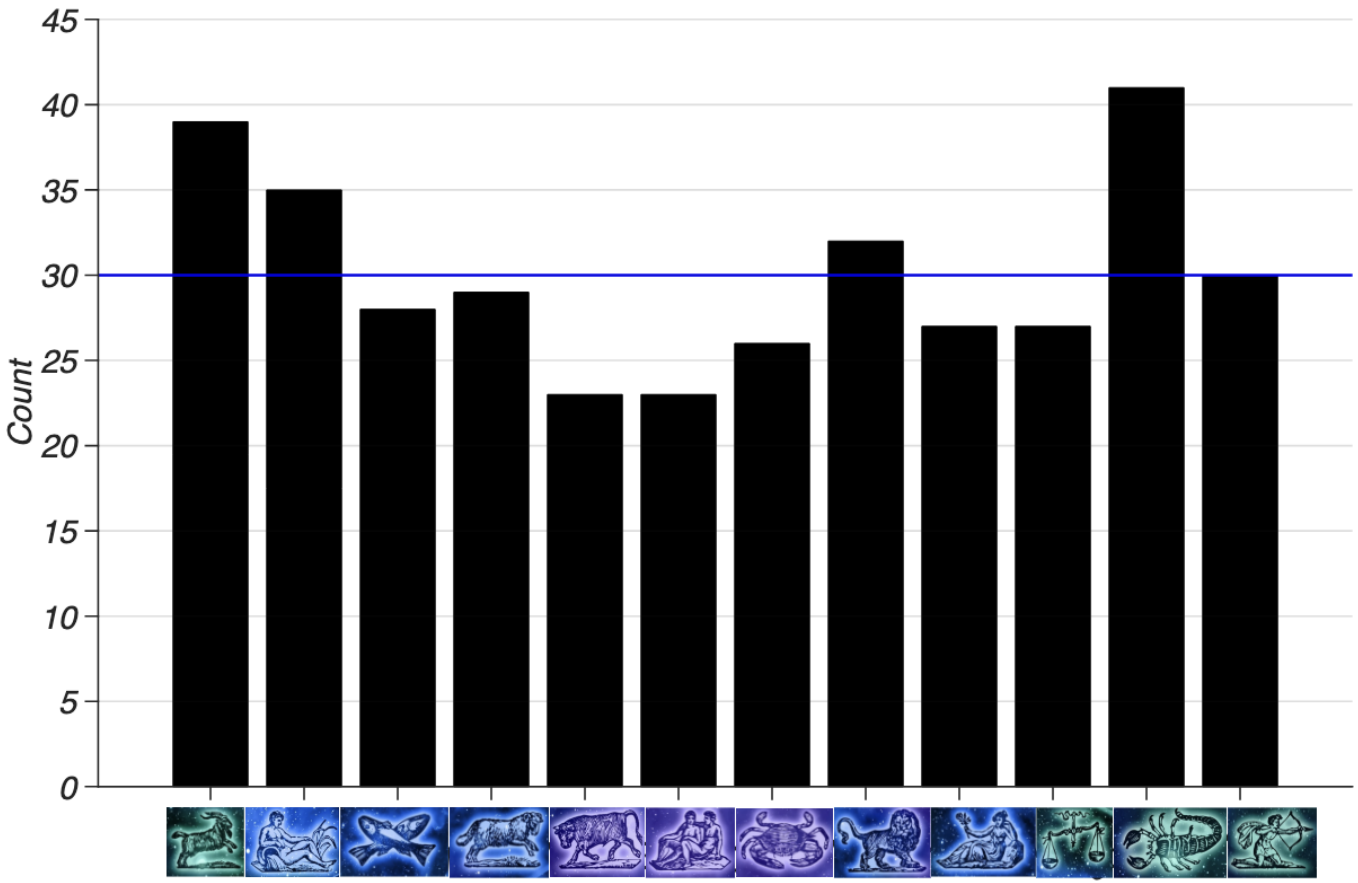
$$\chi^2 = \sum \frac{(\text{Observed Count} - \text{Expected Count})^2}{\text{Expected Count}}$$

Steps

1. Formulate the null hypothesis (H_0)
2. Calculate the expected counts for each category based on H_0.
3. Compute the test statistic χ^2 using the observed and expected counts.
4. Determine the degrees of freedom and critical value for a chosen significance level α.
   - **DF for Goodness-of-fit:** $K - 1$ (K is # of categories)
   - **DF for Test of Independence:** $(r-1) \cdot (c-1)$ (rows, columns)
5. If $\chi^2 > \chi^2_{crit}$, reject H_0. Otherwise, fail to reject H_0.

(1) Goodness-of-fit Test Example: Zodiac Signs and Serial Killers (Full example in lecture!)

Question: Is Zodiac sign predictive of being a serial killer?



(1) Goodness-of-fit Test Example: Zodiac Signs and Serial Killers (Full example in lecture!)

- **Idea of goodness-of-fit:** Comparing observed data to a theoretical or expected distribution.

Question: Is Zodiac sign predictive of being a serial killer?

- **Null Hypothesis:** Serial killer counts are evenly distributed across all 12 Zodiac signs.
- **Alternative Hypothesis:** Some signs are disproportionately represented.
- **Categories:** 12 Zodiac signs.
- **Expected Count:** If N is the total number of serial killers, then each sign is expected to have $\frac{N}{12}$.
- **Degrees of Freedom:** $DF = K - 1 = 12 - 1 = \boxed{11}$

(1) Goodness-of-fit Test Example: Zodiac Signs and Serial Killers (Full example in lecture!)

Sign	Observed Count	Expected Count (n/12)	Deviation	Plugged in to equation	Contribution to χ^2
Capricorn	39	30	9	$9^2/30$	2.7
Aquarius	35	30	5	$5^2/30$	0.83
Pisces	28	30	-2	$-2^2/30$	0.13
Aries	29	30	-1	$-1^2/30$	0.03
Taurus	23	30	-7	$-7^2/30$	1.63
Gemini	23	30	-7	$-7^2/30$	1.63
Cancer	26	30	-4	$-4^2/30$	0.53
Leo	32	30	2	$2^2/30$	0.13
Virgo	27	30	-3	$-3^2/30$	0.3
Libra	27	30	-3	$-3^2/30$	0.3
Scorpio	41	30	11	$11^2/30$	4.03
Sagittarius	30	30	0	$0^2/30$	0

$\chi^2 = 2.7+0.83+0.13+0.03+1.63+1.63+0.53+0.13+0.3+0.3+4.03+0$

$\chi^2 = 12.24$

(1) Goodness-of-fit Test Example: Zodiac Signs and Serial Killers (Full example in lecture!)

Gathered information:

- $DF = 11$, $\chi^2 = 12.24$, $\alpha = 0.05$
- **Find** χ^2_{crit} from table (clickable link!)

df	$\chi^2_{.995}$	$\chi^2_{.990}$	$\chi^2_{.975}$	$\chi^2_{.950}$	$\chi^2_{.900}$	$\chi^2_{.100}$	$\chi^2_{.050}$	$\chi^2_{.025}$	$\chi^2_{.010}$	$\chi^2_{.005}$
1	0.000	0.000	0.001	0.004	0.016	2.706	3.841	5.024	6.635	7.879
2	0.010	0.020	0.051	0.103	0.211	4.605	5.991	7.378	9.210	10.597
3	0.072	0.115	0.216	0.352	0.584	6.251	7.815	9.348	11.345	12.838
4	0.207	0.297	0.484	0.711	1.064	7.779	9.488	11.143	13.277	14.860
5	0.412	0.554	0.831	1.145	1.610	9.236	11.070	12.833	15.086	16.750
6	0.676	0.872	1.237	1.635	2.204	10.645	12.592	14.449	16.812	18.548
7	0.989	1.239	1.690	2.167	2.833	12.017	14.067	16.013	18.475	20.278
8	1.344	1.646	2.180	2.733	3.490	13.362	15.507	17.535	20.090	21.955
9	1.735	2.088	2.700	3.325	4.168	14.684	16.919	19.023	21.666	23.589
10	2.156	2.558	3.247	3.940	4.865	15.987	18.307	20.483	23.209	25.188
11	2.603	3.053	3.816	4.575	5.578	17.275	19.675	21.920	24.725	26.757
12	3.074	3.571	4.404	5.226	6.304	18.549	21.026	23.337	26.217	28.300

- $\chi^2 \overset{?}{>} \chi^2_{crit} \rightarrow 12.24 < 19.675 \implies$ Fail to reject H_0

(2) Test of Independence Example: Pineapple Pizza Preferences

- **Idea of test of independence:** Testing whether one categorical variable (X) is associated with another categorical variable (Y).
- Does knowing the value of X gives you any information about the value of Y?

(2) Test of Independence Example: Pineapple Pizza Preferences

- **Idea of test of independence:** Testing whether one categorical variable (X) is associated with another categorical variable (Y).
- Does knowing the value of X gives you any information about the value of Y?

Question: Is preference for pineapple pizza independent of generation?

- **Null Hypothesis:** Preference for pineapple pizza is independent of generation.
- **Alternative Hypothesis:** Preference varies across generations.
- **Categories:** Generations (Millennials, Gen Z) vs. Preference (For, Undecided, Against)
- $DF = (r-1)\cdot(c-1) = (2-1)\cdot(3-1) = 1\cdot 2 = \boxed{2}$

(2) Test of Independence Example: Pineapple Pizza Preferences

Observed Count:

Observed	For	Undecided	Against	Total
Millennials	92	18	90	200
Gen Z	68	22	110	200
Total	160	40	200	400

Expected Count:

Expectations	For	Undecided	Against
Millennials	80	20	100
Gen Z	80	20	100

$$\chi^2 = \frac{(92-80)^2}{80} + \frac{(18-20)^2}{20} + \ldots + \frac{(110-100)^2}{100} =$$
$$1.8 + 0.2 + 1 + 1.8 + 0.2 + 1 = \boxed{6}$$

(2) Test of Independence Example: Pineapple Pizza Preferences

Gathered Information:

- $DF = 2$, $\chi^2 = 6$, $\alpha = 0.05$
- **Find** χ^2_{crit} from table (clickable link!)

df	$\chi^2_{.995}$	$\chi^2_{.990}$	$\chi^2_{.975}$	$\chi^2_{.950}$	$\chi^2_{.900}$	$\chi^2_{.100}$	$\chi^2_{.050}$	$\chi^2_{.025}$	$\chi^2_{.010}$	$\chi^2_{.005}$
1	0.000	0.000	0.001	0.004	0.016	2.706	3.841	5.024	6.635	7.879
2	0.010	0.020	0.051	0.103	0.211	4.605	5.991	7.378	9.210	10.597
3	0.072	0.115	0.216	0.352	0.584	6.251	7.815	9.348	11.345	12.838
4	0.207	0.297	0.484	0.711	1.064	7.779	9.488	11.143	13.277	14.860
5	0.412	0.554	0.831	1.145	1.610	9.236	11.070	12.833	15.086	16.750
6	0.676	0.872	1.237	1.635	2.204	10.645	12.592	14.449	16.812	18.548
7	0.989	1.239	1.690	2.167	2.833	12.017	14.067	16.013	18.475	20.278
8	1.344	1.646	2.180	2.733	3.490	13.362	15.507	17.535	20.090	21.955
9	1.735	2.088	2.700	3.325	4.168	14.684	16.919	19.023	21.666	23.589
10	2.156	2.558	3.247	3.940	4.865	15.987	18.307	20.483	23.209	25.188
11	2.603	3.053	3.816	4.575	5.578	17.275	19.675	21.920	24.725	26.757

- $\chi^2 \overset{?}{>} \chi^2_{crit} \rightarrow 6 > 5.991 \implies$ Reject H_0
- **Note:** Notice that if α was slightly smaller, the χ^2_{crit} would've been larger $\implies$ Fail to reject!

Scenario and Three Types of Data

- **Scenario:** You want to compare the ratings of two movies.

Scenario and Three Types of Data

- **Scenario:** You want to compare the ratings of two movies.
- **Solution?** An independent t-test might seem reasonable because you are comparing the means of two groups (the two movies).

Scenario and Three Types of Data

- **Scenario:** You want to compare the ratings of two movies.
- **Solution?** An independent t-test might seem reasonable because you are comparing the means of two groups (the two movies).
- **Question:** Is it reasonable to reduce movie ratings to means?

Scenario and Three Types of Data

- **Scenario:** You want to compare the ratings of two movies.
- **Solution?** An independent t-test might seem reasonable because you are comparing the means of two groups (the two movies).
- **Question:** Is it reasonable to reduce movie ratings to means?
- **Answer:** Probably not. Movie ratings are **ordinal data** where the psychological difference between 3 and 4 may not equal the difference between 4 and 5.

Scenario and Three Types of Data

- **Scenario:** You want to compare the ratings of two movies.
- **Solution?** An independent t-test might seem reasonable because you are comparing the means of two groups (the two movies).
- **Question:** Is it reasonable to reduce movie ratings to means?
- **Answer:** Probably not. Movie ratings are **ordinal data** where the psychological difference between 3 and 4 may not equal the difference between 4 and 5.

Three Types of Data:

1. **Nominal:** Unordered categories (e.g., gender, colors).
2. **Ordinal:** Ordered categories **without** consistent differences (e.g., movie ratings, education levels).
   - Indicates position or rank, but differences between ranks are not necessarily consistent.
3. **Cardinal (Interval/Ratio):** Numerical values where arithmetic operations are meaningful (e.g., temperature, height, age).

Mann-Whitney U Test (AKA Wilcoxon Rank-sum Test)

- **Mann-Whitney U Test:** A non-parametric test used to see if two samples come from populations with the same **median**.
- It is an alternative to the T-test when the assumptions of normality or equal variance are violated.
- The t-test compares means, while the Mann-Whitney U Test compares medians (under certain conditions).

When to Use the Mann-Whitney U Test:

- When it is inappropriate to reduce a dataset to its mean.
- For **ordinal data** (e.g. movie ratings) where the difference between values may not be consistent.
- When there are outliers, as the median is more robust to extreme values than the mean.

Mann-Whitney U Test (AKA Wilcoxon Rank-sum Test)

Steps

1. Combine all values from both samples and arrange them in rank order (from smallest to largest).
2. Assign ranks to the values, ignoring which sample they belong to. Assign average ranks in case of ties.
3. Calculate the sum of ranks for each sample (T_A and T_B).
4. Compute the test statistic U:

Test Statistic (U)

$$U = n_A n_B + \frac{n_X(n_X + 1)}{2} - T_X$$

Where:

- n_A, n_B: Sizes of the two samples.
- n_X: Size of the sample with the **larger rank sum** (either n_A or n_B).
- T_X: Sum of ranks for the sample with the **larger rank sum**.

Kolmogorov-Smirnov (KS) Test

- **Kolmogorov-Smirnov (KS) Test:** A non-parametric test used to determine whether two samples come from the same distribution.
- Unlike tests that compare means or medians, the KS test focuses on the **overall distribution shape**.

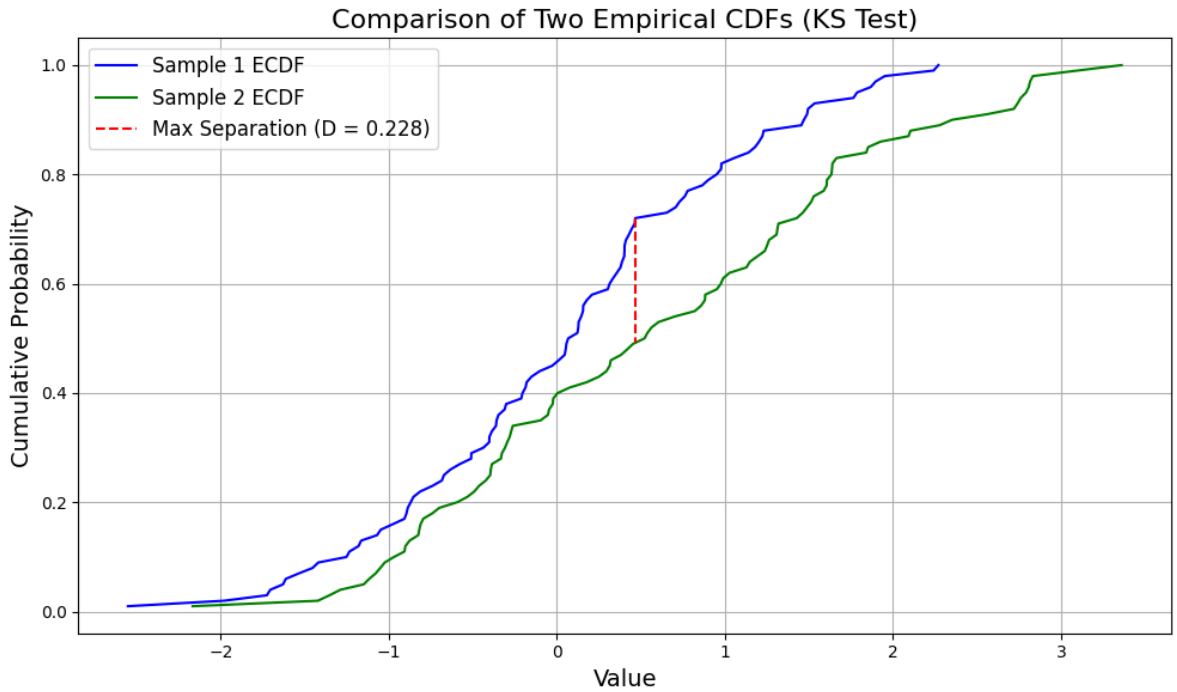
Prupose:

- Test whether the underlying distributions of two samples are the same.
- Serves as a **goodness-of-fit test**, comparing a sample's distribution to a reference.

KS Test Steps

Steps:

1. Compute the empirical cumulative distribution function (eCDF) for both samples.
2. plot the the eCDF of the two samples.
3. **Test Statistic (D):** The value of the largest separation between the two eCDF.



Permutation Test

- **Permutation Tests:** A non-parametirc tests that works by **shuffling** data and computing the test statistic under many repitions in order to generate a **custom null distribution.**

Permutation testing is useful under the following conditions:

1. Data does not meet assumptions for parametric testing.
2. Other non-parametric tests are not applicable.

Steps of Permutation Test

1. **Define the Data Distributions:** Identify the two groups to be compared (e.g., ratings for the first and second Harry Potter movies).
2. **Combine and Shuffle the Data:** Pool the two datasets into one large set and shuffle randomly.
3. **Split the Shuffled Data:** Divide the shuffled data back into two groups of the original sizes.
4. **Compute the Test Statistic:** Calculate a relevant test statistic (e.g., difference in means, medians, etc.) for the new groups.
5. **Generate the Null Distribution:** Repeat Steps 3 and 4 many times to build a distribution of the test statistic under the null hypothesis.
6. **Compute the Test Statistic for Original Data:** Calculate the test statistic for the original (unshuffled) data.
7. **Compare and Compute p-value:** Compare the observed statistic to the null distribution and compute the p-value.

Summary of Non-parametric Tests

Test	Purpose	Data Type
Chi-square	Test of Independence or Goodness-of-Fit	**Categorical**
Mann-Whitney U	Comparing Ranks (often interpreted as comparing medians)	**Ordinal**
KS	Comparing Overall Distribution	**Typically cardinal**
Permutation Test	General-purpose test depending on resampling	**Any Type***

- *: Depends on test statistic.

Quiz + Discussion