Statistics and Data Analysis
Unit 03 – Lecture 06 Notes
Non-parametric Tests and p-value Interpretation

Tofik Ali

February 17, 2026

Topic

Rank-based tests and p-value interpretation; statistical vs practical significance.

How to Use These Notes

These notes are written for students who are seeing the topic for the first time. They follow the slide order, but add the missing 'why', interpretation, and common mistakes. If you get stuck, look at the worked exercises and then run the Python demo.

Course repository (slides, demos, datasets): `https://github.com/tali7c/Statistics-and-Data-Analysis`

Time Plan (55 minutes)

- 0–10 min: Attendance + recap of previous lecture
- 10–35 min: Core concepts (this lecture's sections)
- 35–45 min: Exercises (solve 1–2 in class, rest as practice)
- 45–50 min: Mini demo + interpretation of output
- 50–55 min: Buffer / wrap-up (leave 5 minutes early)

Slide-by-slide Notes

Title Slide

State the lecture title clearly and connect it to what students already know. Tell students what they will be able to do by the end (not just what you will cover).

Quick Links / Agenda

Explain the structure of the lecture and where the exercises and demo appear.

- Overview

- When to Use
- Common Tests
- Exercises
- Demo
- Summary

Learning Outcomes

- Explain why non-parametric tests are used
- Choose Mann-Whitney / Wilcoxon / Kruskal-Wallis
- Interpret p-values carefully
- Discuss statistical vs practical significance
- Explain multiple testing risk

Why these outcomes matter. A **p-value** is computed assuming the null hypothesis H_0 is true. It measures how surprising the observed data (or something more extreme) would be under H_0. A small p-value suggests the data is hard to explain by H_0 alone, but it does not tell you how large the effect is or whether it is practically important. **Non-parametric** tests rely on ranks rather than raw values. They are useful when data is skewed, has outliers, or is ordinal (e.g., ratings). The trade-off is that they may be less powerful than parametric tests when assumptions of parametric tests actually hold.

When to Use: Key Points

- Skewed data/outliers
- Ordinal scales
- Small sample and doubtful normality

Common Tests: Key Points

- Two independent groups: Mann-Whitney U
- Paired samples: Wilcoxon signed-rank
- 3+ groups: Kruskal-Wallis

Exercises (with Solutions)

Attempt the exercise first, then compare with the solution. Focus on interpretation, not only arithmetic.

Exercise 1: Choose test

Same students before/after training (skewed). Which test?

Solution

- Wilcoxon signed-rank

Exercise 2: Practical vs statistical

Very small p-value but tiny difference: what should you report?

Solution

- Report effect size and context; significance != importance.

Walkthrough. A **p-value** is computed assuming the null hypothesis H_0 is true. It measures how surprising the observed data (or something more extreme) would be under H_0. A small p-value suggests the data is hard to explain by H_0 alone, but it does not tell you how large the effect is or whether it is practically important. **Effect size** quantifies *how big* a difference/relationship is (e.g., Cohen's d, correlation r). With large samples, even tiny effects can be statistically significant, so reporting effect size prevents over-claiming.

Exercise 3: Multiple testing

20 tests at alpha=0.05: expected false positives?

Solution

- About 1 on average.

Walkthrough. The **significance level** α is the maximum Type I error rate you are willing to tolerate: the probability of rejecting H_0 when H_0 is actually true. Common choices are 0.05 or 0.01, but the right value depends on consequences of false alarms vs missed detections. In chi-square tests, **expected counts** are what you would expect to see if H_0 were true (e.g., independence). Very small expected counts can break the approximation used by the test; a common rule of thumb is that most expected counts should be at least 5.

Mini Demo (Python)

Run from the lecture folder:

```
python demo/demo.py
```

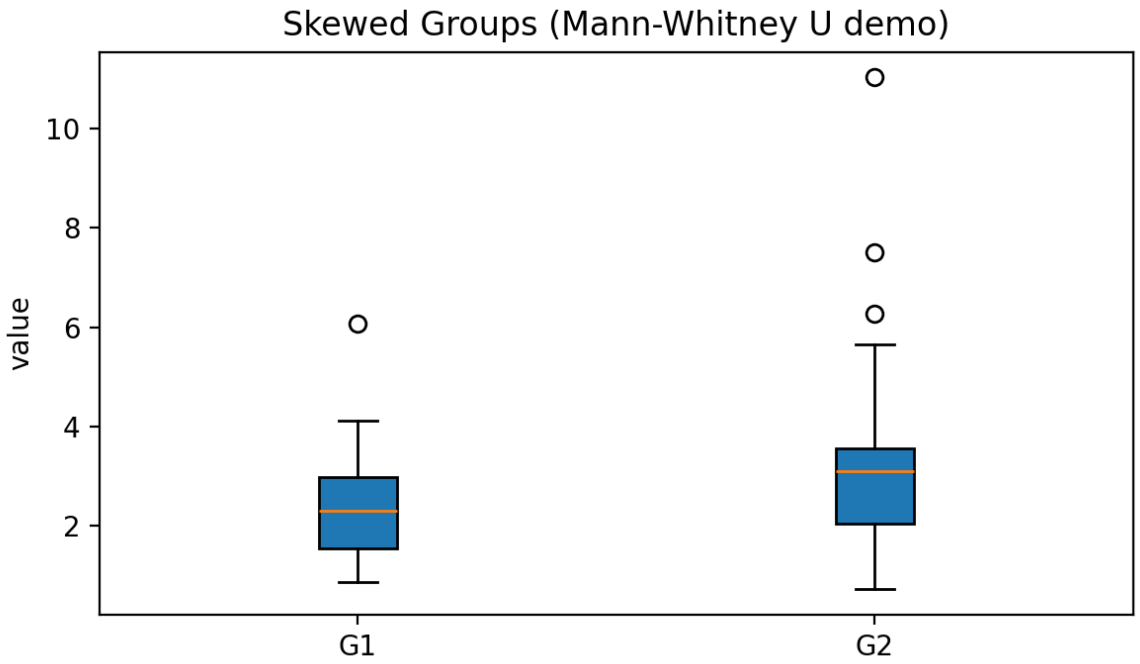
Output files:

- `images/demo.png`
- `data/results.txt`

What to show and say.

- Generates two skewed groups and runs Mann-Whitney U (rank-based) test.
- Shows a boxplot and reports medians to emphasize robustness under skew/outliers.
- Compare with a mean-based test to discuss when non-parametric is preferred.

Demo Output (Example)



Summary

- Key definitions and the main formula.
- How to interpret results in context.
- How the demo connects to the theory.

Exit Question

Give one reason to prefer a rank-based test over a mean-based test.

Suggested answer (for revision). Rank-based tests reduce sensitivity to outliers/skew and work for ordinal data, so they are safer when normality is doubtful.

References

- Montgomery, D. C., & Runger, G. C. *Applied Statistics and Probability for Engineers*, Wiley.
- Devore, J. L. *Probability and Statistics for Engineering and the Sciences*, Cengage.
- McKinney, W. *Python for Data Analysis*, O'Reilly.

Appendix: Slide Deck Content (Reference)

The material below is a reference copy of the slide deck content. Exercise solutions are explained in the main notes where applicable.

Title Slide

https://github.com/tali7c/Statistics-and-Data-Analysis

Quick Links

Overview When to Use Common Tests Exercises Demo Summary

Agenda

- Overview
- When to Use
- Common Tests
- Exercises
- Demo
- Summary

Learning Outcomes

- Explain why non-parametric tests are used
- Choose Mann-Whitney / Wilcoxon / Kruskal-Wallis
- Interpret p-values carefully
- Discuss statistical vs practical significance
- Explain multiple testing risk

When to Use: Key Points

- Skewed data/outliers
- Ordinal scales
- Small sample and doubtful normality

Common Tests: Key Points

- Two independent groups: Mann-Whitney U
- Paired samples: Wilcoxon signed-rank
- 3+ groups: Kruskal-Wallis

Exercise 1: Choose test

Same students before/after training (skewed). Which test?

Solution 1

- Wilcoxon signed-rank

Exercise 2: Practical vs statistical

Very small p-value but tiny difference: what should you report?

Solution 2

- Report effect size and context; significance != importance.

Exercise 3: Multiple testing

20 tests at alpha=0.05: expected false positives?

Solution 3

- About 1 on average.

Mini Demo (Python)

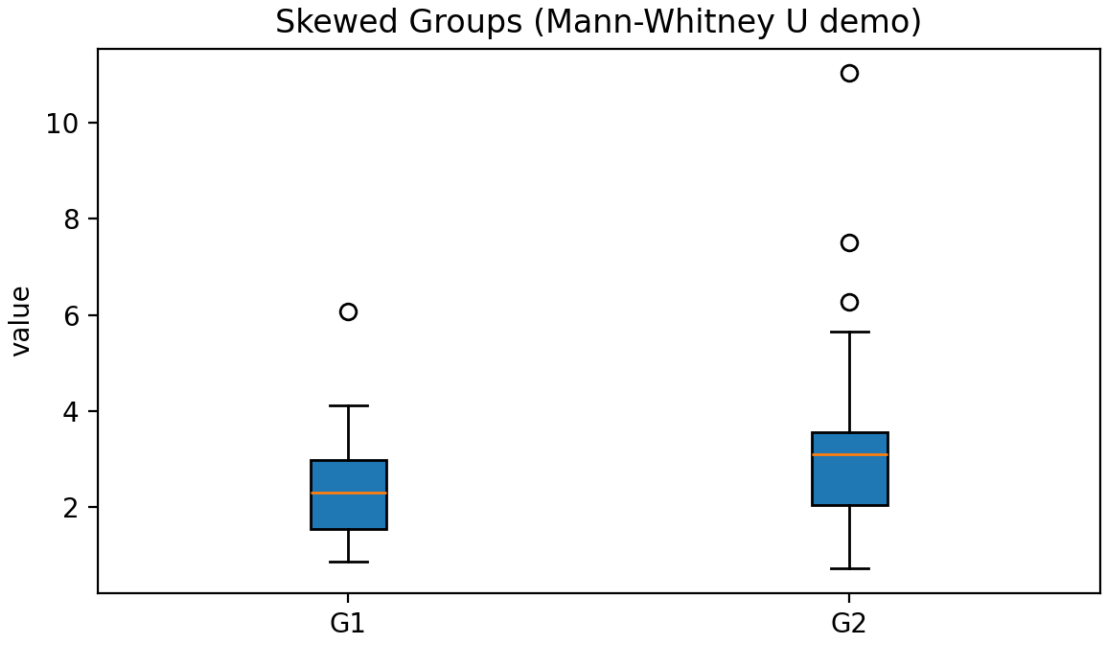
Run from the lecture folder:

```
python demo/demo.py
```

Outputs:

- `images/demo.png`
- `data/results.txt`

Demo Output (Example)



Summary

- Key definitions and the main formula.
- How to interpret results in context.
- How the demo connects to the theory.

Exit Question

Give one reason to prefer a rank-based test over a mean-based test.