KIN 610: Quantitative Methods in Kinesiology

Chapter 19: Nonparametric Methods

Ovande Furtado Jr., PhD.

2026-04-15

Table of contents

1 FYI

This presentation is based on the following books. The references are coming from these books unless otherwise specified.

Main sources:

- Weir, J. P., & Vincent, W. J. (2021). *Statistics in kinesiology* (5th ed.). Human Kinetics.
- Field, A. (2018). *Discovering statistics using IBM SPSS statistics* (5th ed.). SAGE Publications.
- Furtado, O., Jr. (2026). *Statistics for movement science: A hands-on guide with SPSS* (1st ed.). https://drfurtado.github.io/sms

ClassShare App

You may be asked in class to go to the ClassShare App to answer questions.

- https://classshare-b44e7.web.app/

SPSS Tutorial

- SPSS Tutorial: Nonparametric Tests

2 Intro Question

A physical therapist rates the functional recovery of 20 patients on a 0–100 scale before and after a 6-week program. The scores are bounded, skewed toward the floor at pre-test, and the sample is small. She wants to know whether patients improved. What test should she use and why?

Click to reveal answer

The **Wilcoxon signed-rank test** is appropriate. The outcome is bounded and ordinal-like (not a true continuous interval scale), the distribution is skewed, and the sample is too small to rely on the Central Limit Theorem. Wilcoxon converts the difference scores to ranks, making it robust to the non-normality and floor effects that would compromise a paired t-test.

- Parametric tests (t, ANOVA, Pearson r) assume specific distributional forms
- **Nonparametric tests** make no distributional assumptions — they are distribution-free
- The key skill: **knowing when to switch**, and **which test to pick**

3 Learning Objectives

By the end of this chapter, you should be able to:

- Identify when nonparametric tests are appropriate alternatives to parametric procedures
- Explain how rank transformation forms the basis of most nonparametric methods
- Select the correct nonparametric test for a given design using the decision guide
- Interpret output from chi-square, Spearman , Mann-Whitney U, Wilcoxon, Kruskal-Wallis, and Friedman's tests in SPSS
- Report rank-biserial r, Cramér's V, Kendall's W, and Spearman as nonparametric effect sizes
- Write APA-style results paragraphs for each nonparametric test covered in this chapter

4 When to Use Nonparametric Tests

Use nonparametric when:

1. **Ordinal measurement** — pain ratings, Borg RPE, Likert scales: intervals between values are not equal
2. **Severe non-normality + small samples** ($n < 20$ per group) — CLT cannot rescue you
3. **Categorical outcome** — injured/not injured, group membership: chi-square is the only option
4. **Floor or ceiling effects** — scores pile up at scale boundaries, violating symmetry assumptions

What nonparametric tests give up:

- **Power** — converting to ranks discards information when data are well-behaved
- **Flexibility** — fewer options for covariates, complex factorial designs
- **Interpretability** — U, W, H are less intuitive than t or F alone

> ⚠ Warning
>
> Using nonparametric tests "just to be safe" when assumptions *are* met unnecessarily reduces power. Choose based on evidence, not habit.

5 The Logic of Ranks

Most nonparametric tests share one core mechanism: convert raw scores → ranks, then compute on ranks

Participant	Raw score	Rank
A	48	1
B	62	2
C	71	3
D	85	4
E	91	5

Key insight: After ranking, the test depends only on the *ordering*, not the exact magnitudes. An outlier of 48 and an outlier of −200 receive the **same rank (1)** — this is the source of robustness.

Tied ranks: When participants share the same raw score, each receives the *average* of the ranks they would have occupied (e.g., two tied scores at ranks 3 and 4 → both receive rank 3.5). SPSS handles this automatically.

6 Chi-Square Tests: Two Designs

Goodness-of-Fit

- One categorical variable
- H : observed frequencies match a hypothesized distribution
- Example: Is the sex split in our sample 50/50?
- df = $k - 1$ (where k = number of categories)
- Effect size: **Cramér's *V*** (or phi for 2 categories)

Test of Independence

- Two categorical variables in a contingency table
- H : the two variables are independent (not associated)
- Example: Is group assignment independent of sex?
- df = (rows − 1)(cols − 1)
- Effect size: **phi ( )** for 2×2; **Cramér's *V*** for larger tables

Critical assumption: All expected cell counts 5. If violated → use **Fisher's exact test** (2×2 tables).

7 Chi-Square: Worked Example

Research question: Is the distribution of participants across training and control groups independent of sex?

	Control	Training	Row total
Female	21	12	33
Male	9	18	27
Col total	**30**	**30**	**60**

- All expected cell counts 5 → chi-square is appropriate
- Result: $^2(1, N = 60) = 5.45$, $p = .020$, $= .302$

See Figure 1 in the SMS textbook (Ch. 19): mosaic plot of sex by group assignment — tile height represents the proportion of each sex within each group, visualizing the overrepresentation of females in the control group.

APA write-up:

> A chi-square test of independence revealed a statistically significant association between participant sex and group assignment, $^2(1, N = 60) = 5.45$, $p = .020$, $= .302$. Females were proportionally more common in the control group (70%) than the training group (40%).

8 Spearman Rank-Order Correlation

When to use Spearman instead of Pearson r:

- One or both variables are **ordinal** (e.g., RPE, injury severity ratings)
- The relationship is **monotonic but not linear** (e.g., diminishing returns)
- **Outliers** substantially distort Pearson r

What Spearman measures: Whether one variable *tends to increase* as the other increases (or decreases) — regardless of whether the pattern is a straight line.

Example from Ch. 19: Is there an association between post-test muscular strength and 20-m sprint time?

> **See Figure 2 (fig-spearman-scatter) in the SMS textbook (Ch. 19):** scatter plot with lowess smoother showing the strong negative monotonic relationship between post-test strength and sprint time ($N = 60$).

- Result: $(53) = -.71$, $p < .001$ ($N = 55$) — stronger participants tended to sprint faster (lower times)
- Report df as $n - 2$ in parentheses after

9 Mann-Whitney U Test

Nonparametric alternative to: Independent-samples t-test

Use when: Two independent groups; outcome is ordinal or severely non-normal in small samples

How it works:

- Pool all scores → rank together
- U statistic = number of times group 1 score exceeds a group 2 score (and vice versa)
- A large discrepancy in rank sums → one group consistently outranks the other

Effect size: Rank-biserial r (range -1 to $+1$; same magnitude benchmarks as Pearson r)

Example from Ch. 19: Do training and control groups differ in post-test RPE (Borg 6–20 scale)?

> **See Figure 3 (fig-mwu-boxplot) in the SMS textbook (Ch. 19):** boxplots of post-test RPE by group — the control group shows higher RPE values ($Mdn = 13$) than the training group ($Mdn = 12$).

- Result: $U = 249$, $p = .030$, $r = -.34$ (medium effect); $n\_ctrl = 30$, $n\_train = 25$
- Report medians, not means, as the central tendency measure

10 Wilcoxon Signed-Rank Test

Nonparametric alternative to: Paired-samples t-test

Use when: Two related measurements per participant (pre/post, left/right); difference scores are not normally distributed or are ordinal

How it works:

- Compute difference score for each participant
- Rank the *absolute values* of differences
- W statistic = sum of ranks for positive differences − sum for negative differences
- If H is true, positive and negative ranks roughly cancel

Wilcoxon vs. sign test: Wilcoxon uses the *magnitude and direction* of differences → more powerful. Sign test uses *direction only* → use only when magnitudes are untrustworthy.

Example from Ch. 19: Did functional ability improve from pre- to post-test in the training group?

> **See Figure 4 (fig-wilcoxon-paired) in the SMS textbook (Ch. 19):** paired scatter plot connecting each participant's pre and post functional scores — nearly all lines slope upward, consistent with the large effect.

- Result: $W = 69$, $p < .001$, $r = -.71$ (large effect); $n = 25$ complete pairs; pre $Mdn = 67.1$, post $Mdn = 74.9$

11 Kruskal-Wallis Test

Nonparametric alternative to: One-way between-subjects ANOVA

Use when: Three or more independent groups; outcome is ordinal or severely non-normal

How it works: - Pool all scores across groups → rank together - H statistic based on rank sums within each group (approximately chi-square distributed, df $= k - 1$) - Significant H → at least one group's distribution tends to be higher or lower than the others

After a significant result → post hoc comparisons: - **Dunn's test (Bonferroni-corrected)** pairwise Mann-Whitney comparisons - Same logic as post hoc in ANOVA: omnibus first, then pairwise

Note: When $k = 2$, Kruskal-Wallis is mathematically equivalent to Mann-Whitney U — use Mann-Whitney directly.

Example from Ch. 19: Do balance errors differ across three time points in the full sample?

- Result: $H(2) = 1.03$, $p = .597$ — non-significant; no evidence of systematic time-related change in balance errors at the group level.

12 Friedman's Test

Nonparametric alternative to: One-way repeated-measures ANOVA

Use when: Same participants measured under k 3 conditions; outcome is ordinal or non-normal

How it works: - Rank scores *within each participant* across conditions (each participant has their own ranks 1 through k) - Test statistic based on whether condition rank sums differ more than chance predicts - Significant result → at least one condition differs from the others

Effect size: Kendall's *W* (concordance coefficient, 0–1) - Near 0: participants show no consistent ordering across conditions - Near 1: all participants rank conditions in the same order

After a significant result: Pairwise Wilcoxon signed-rank tests with Bonferroni correction

Example from Ch. 19: Does RPE change across pre, mid, post in the training group?

> **See Figure 5 (fig-friedman-profile) in the SMS textbook (Ch. 19):** individual RPE profiles across time — criss-crossing lines show no consistent directional trend, consistent with a non-significant result.

- Result: $^2(2) = 4.46$, $p = .107$, $W = .05$ — non-significant; no systematic change in RPE over time.

13 Test Selection Decision Guide

See Table 1 (tbl-nonparam-guide) in the SMS textbook (Ch. 19) for the complete decision guide. Key mappings:

Design	Parametric analogue	Nonparametric test	Effect size
1 sample vs. median	One-sample t	Sign test	$\hat{p}$
Categorical outcome, 1 variable	—	Chi-square GoF	Cramér's V
Categorical outcome, 2 variables	—	Chi-square independence	or Cramér's V
Monotonic relationship	Pearson r	Spearman	
2 independent groups	Independent t	Mann-Whitney U	Rank-biserial r
2 related samples	Paired t	Wilcoxon signed-rank	Rank-biserial r
3 independent groups	One-way ANOVA	Kruskal-Wallis	^{2}_H

Design	Parametric analogue	Nonparametric test	Effect size
3 related conditions	Repeated-measures ANOVA	Friedman's	Kendall's W
2+ factors, between-subjects	Factorial ANOVA	Scheirer-Ray-Hare	^{2}_H per effect

The key decision questions: 1. How many groups/conditions? (2 vs. 3) 2. Independent or related/repeated measurements? 3. Is the outcome continuous, ordinal, or categorical?

14 Effect Sizes for Nonparametric Tests

Always report an effect size alongside the p-value. A non-significant result does not mean the effect is zero.

Test	Effect size	Range	Benchmarks
Chi-square (2×2)	Phi ( )	0–1	.10 small / .30 medium / .50 large
Chi-square (larger)	Cramér's V	0–1	same
Spearman	itself	−1 to +1	same as Pearson r
Mann-Whitney U	Rank-biserial r	−1 to +1	.10 / .30 / .50
Wilcoxon signed-rank	Rank-biserial r	−1 to +1	.10 / .30 / .50
Kruskal-Wallis	^{2}_H	0–1	.01 / .06 / .14
Friedman's	Kendall's W	0–1	concordance, not ^{2}

> ⚠ Warning
>
> **Kendall's W eta-squared.** W measures concordance in rank ordering across participants. Do not benchmark it against ANOVA effect size standards.

15 Common Pitfalls

1. **Using nonparametric "just to be safe."** Rank transformation discards information and reduces power when parametric assumptions are met. Base the choice on evidence, not habit.

2. **Treating non-significance as equivalence.** A non-significant result with a small sample may simply lack power. Always report the effect size so readers can evaluate practical significance.

3. **Ignoring expected cell count requirements.** Running chi-square with expected counts < 5 produces an unreliable p-value. Check SPSS output — if any expected cell is < 5, switch to Fisher's exact test.

4. **Running multiple nonparametric follow-ups without correction.** Post-hoc Wilcoxon or Mann-Whitney tests after a significant Friedman or Kruskal-Wallis inflate familywise Type I error. Apply Bonferroni or Holm correction.

5. **Misinterpreting Kendall's *W* as eta-squared.** $W = .26$ means moderate concordance in rank ordering — it does not mean "26% of variance explained."

6. **Omitting effect sizes.** Statistical significance and effect size are both required for a complete APA report.

16 APA Reporting Templates

Chi-square independence: > "A chi-square test of independence revealed a [significant/non-significant] association between [var1] and [var2], $^2(df, N = n)$ = [value], p = [value], /V = [value]."

Spearman : > "Spearman rank-order correlation indicated a [strong/moderate/weak] [positive/negative] [significant/non-significant] association between [var1] and [var2], (df) = [value], p = [value]."

Mann-Whitney U: > "A Mann-Whitney U test indicated that [group 1] (Mdn = [value]) [significantly/did not significantly] differ from [group 2] (Mdn = [value]), U = [value], p = [value], r = [value]."

Wilcoxon signed-rank: > "A Wilcoxon signed-rank test revealed a [significant/non-significant] change from [time 1] (Mdn = [value]) to [time 2] (Mdn = [value]), W = [value], p = [value], r = [value]."

Kruskal-Wallis: > "A Kruskal-Wallis test indicated [significant/non-significant] differences across groups, $H(df)$ = [value], p = [value], ^{2}_H = [value]. [Post-hoc Dunn-Bonferroni comparisons revealed that…]"

Friedman's test: > "Friedman's test revealed [significant/non-significant] differences across conditions, $^2(df)$ = [value], p = [value], W = [value]."

17 Check Questions

Q1. A researcher measures muscle soreness (0–10 VAS) in 20 participants at pre-, mid-, and post-intervention (same participants at all three time points). Which test should they use?

Click to reveal answer

Friedman's test — three related conditions (same participants), ordinal-like outcome (bounded VAS), and small sample. Effect size: Kendall's W. Follow up with Bonferroni-corrected pairwise Wilcoxon tests if significant.

Q2. A sport scientist uses a chi-square test of independence on a 3×4 contingency table and finds three cells with expected counts of 2. What should they do?

Click to reveal answer

The chi-square approximation is unreliable when expected cell counts < 5. Options: (1) collapse categories to create a smaller table with all expected counts 5; (2) collect more data; (3) use an exact permutation test. Fisher's exact test is designed for 2×2 tables only; for larger tables, collapsing or permutation approaches are needed.

Q3. A Kruskal-Wallis test comparing four groups yields $H(3) = 11.84$, $p = .008$. What does this tell you, and what must you do next?

Click to reveal answer

The significant H indicates that **at least one group** tends to have higher or lower scores than the others — but it does not identify *which* pairs differ. The next step is **Dunn's test with Bonferroni correction** for all pairwise comparisons. Do not run uncorrected Mann-Whitney U tests, which would inflate the familywise error rate.

1. Furtado, O., Jr. (2026). *Statistics for movement science: A hands-on guide with SPSS* (1st ed.). https://drfurtado.github.io/sms/