Non-Parametric Within-Participants ANOVA

PSYC5141: Experimental Methods for Psychology

Mark Hurlstone, Lancaster University

Week 4, Lecture 2

Lecture Overview

- **Introduction to the Friedman test**
  - When and why we might use Friedman
  - How it differs from Kruskal–Wallis
- **Calculating the Friedman test statistic**
  - Ranking observations within participants
  - Calculating and interpreting the test statistic
  - Measuring effect size using Kendall's W
- **Follow-up comparisons**
  - Wilcoxon signed-rank tests
  - Planned comparisons, post-hoc tests and multiple-comparison corrections
- **Summary**
  - Friedman versus parametric within-participants ANOVA
  - Key distinctions between the non-parametric procedures

Introducing the Friedman Test

From Kruskal–Wallis to Friedman

Kruskal–Wallis	Friedman
Independent groups	Repeated measures
Different participants in each condition	Same participants in every condition
Rank observations across all groups	Rank observations within each participant
Tests differences among group distributions	Tests differences in condition rankings

Both tests use ranks, but the way ranks are assigned reflects the experimental design

When might Friedman be appropriate?

- Parametric repeated-measures ANOVA models differences between condition means
- Its assumptions include independence between participants, normally distributed errors and sphericity
- When the data are ordinal or the assumptions of parametric analysis are poorly supported, a rank-based alternative may be useful
- The Friedman test does not require normality or sphericity

Introducing the Friedman test

The Friedman test is a non-parametric alternative to single-factor within-participants ANOVA

- Appropriate for three or more related conditions
- Each participant provides an observation in every condition
- Observations are ranked separately within each participant
- Tests whether the conditions differ systematically in their rankings
- Produces a test statistic that can be evaluated using a chi-squared approximation

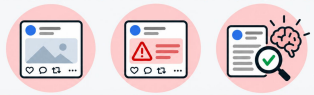
A Within-Participants Experiment

- A psychologist investigates whether misinformation warnings improve people's ability to distinguish between true and false information on social media
- **Independent variable:** Type of misinformation warning, with three levels:
  - A_1: No warning
  - A_2: Generic warning
  - A_3: Specific warning
- **Dependent variable:** Percentage of social media posts correctly classified as true or false
- Design: all five participants complete all three conditions

Example data

Misinformation warnings and accuracy

The same participants complete all three conditions (within-participants design).



Participant	A_1 No warning	A_2 Generic warning	A_3 Specific warning
P1	35	48	74
P2	42	55	77
P3	38	49	79
P4	44	65	84
P5	70	68	90

The logic of Friedman

- Suppose each participant receives ranks of 1, 2 and 3 according to their accuracy in the three conditions
- If warning condition has no systematic effect, we wouldn't expect one condition to consistently receive higher ranks than the others
- However, if specific warnings improve performance, we'd expect the specific-warning condition to receive relatively high ranks across participants

The Friedman test evaluates **how consistently the conditions receive different rankings**

Understanding the ranking procedure

Step 1: Rank observations within each participant

Participant	A_1 score	A_2 score	A_3 score	A_1 rank	A_2 rank	A_3 rank
P1	35	48	74	1	2	3
P2	42	55	77	1	2	3
P3	38	49	79	1	2	3
P4	44	65	84	1	2	3
P5	70	68	90	2	1	3

- Rank each participant's scores separately, from lowest (1) to highest (3)
- Each participant contributes one rank to each condition
- The ranks within each participant always sum to 6 (assuming three conditions, including when ties receive average ranks)
- Unlike Kruskal–Wallis, we do not pool scores across participants before ranking

Step 2: Calculate the rank totals for each condition

Participant	A_1 rank	A_2 rank	A_3 rank
P1	1	2	3
P2	1	2	3
P3	1	2	3
P4	1	2	3
P5	2	1	3
Sum of ranks	6	9	15

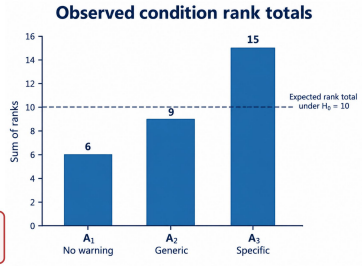
- Each condition has five ranks, and the total of all rank sums is 30
- The expected rank total for each condition under the null hypothesis is:

$$\frac{n(k+1)}{2} = \frac{5(4)}{2} = 10$$

What do the rank totals tell us?

- If warning condition has no systematic influence, the conditions should have similar rank totals
- The expected rank total for each condition is 10
- Our observed rank totals are 6, 9 and 15
- The specific-warning condition consistently receives the highest rank

Main question: Are the observed differences in rank totals sufficiently large to reject the null hypothesis?



Introducing the Friedman formula

$$Q = \frac{12}{nk(k+1)} \sum_{j=1}^{k} R_j^2 - 3n(k+1)$$

Symbols:

- n = number of participants
- k = number of conditions
- R_j = sum of ranks for condition j

For our example:

- $n = 5$
- $k = 3$
- $R_1 = 6,\ R_2 = 9,\ R_3 = 15$

Introducing the Friedman formula

“ Square each condition's rank total and sum across conditions

$$Q = \frac{12}{nk(k+1)} \sum_{j=1}^{k} R_j^2 - 3n(k+1)$$

Symbols:

- n = number of participants
- k = number of conditions
- R_j = sum of ranks for condition j

For our example:

- $n = 5$
- $k = 3$
- $R_1 = 6,\ R_2 = 9,\ R_3 = 15$

Introducing the Friedman formula

$$Q = \frac{12}{nk(k+1)} \sum_{j=1}^{k} R_j^2 - 3n(k+1)$$

“ Adjusts the calculation for the number of participants and conditions

Symbols:

- n = number of participants
- k = number of conditions
- R_j = sum of ranks for condition j

For our example:

- $n = 5$
- $k = 3$
- $R_1 = 6,\ R_2 = 9,\ R_3 = 15$

Introducing the Friedman formula

For our example:

- $n = 5$
- $k = 3$
- $R_1 = 6,\ R_2 = 9,\ R_3 = 15$

$$Q = \frac{12}{nk(k+1)} \sum_{j=1}^{k} R_j^2 - 3n(k+1)$$

Introducing the Friedman formula

For our example:

- $n = 5$
- $k = 3$
- $R_1 = 6,\ R_2 = 9,\ R_3 = 15$

$$Q = \frac{12}{nk(k+1)} \sum_{j=1}^{k} R_j^2 - 3n(k+1)$$

Introducing the Friedman formula

For our example:

- $n = 5$
- $k = 3$
- $R_1 = 6,\ R_2 = 9,\ R_3 = 15$

$$Q = \frac{12}{nk(k+1)} \sum_{j=1}^{k} R_j^2 - 3n(k+1)$$

Introducing the Friedman formula

For our example:

- $n = 5$
- $k = 3$
- $R_1 = 6,\ R_2 = 9,\ R_3 = 15$

$$Q = \frac{12}{nk(k+1)} \sum_{j=1}^{k} R_j^2 - 3n(k+1)$$

Introducing the Friedman formula

For our example:

- $n = 5$
- $k = 3$
- $R_1 = 6,\ R_2 = 9,\ R_3 = 15$

$$Q = \frac{12}{5(3)(4)}(6^2 + 9^2 + 15^2) - 3(5)(4)$$

Introducing the Friedman formula

For our example:

- $n = 5$
- $k = 3$
- $R_1 = 6,\ R_2 = 9,\ R_3 = 15$

$$Q = \frac{12}{5(3)(4)}(6^2 + 9^2 + 15^2) - 3(5)(4)$$

$$Q = \frac{12}{60}(36 + 81 + 225) - 60$$

Introducing the Friedman formula

For our example:

- $n = 5$
- $k = 3$
- $R_1 = 6,\ R_2 = 9,\ R_3 = 15$

$$Q = \frac{12}{5(3)(4)}(6^2 + 9^2 + 15^2) - 3(5)(4)$$

$$Q = \frac{12}{60}(36 + 81 + 225) - 60$$

$$Q = \frac{12}{60}(342) - 60$$

Introducing the Friedman formula

For our example:

- $n = 5$
- $k = 3$
- $R_1 = 6,\ R_2 = 9,\ R_3 = 15$

$$Q = \frac{12}{5(3)(4)}(6^2 + 9^2 + 15^2) - 3(5)(4)$$

$$Q = \frac{12}{60}(36 + 81 + 225) - 60$$

$$Q = \frac{12}{60}(342) - 60$$

$$Q = 68.4 - 60 = \boxed{8.40}$$

Interpreting the omnibus result

$$df = k - 1 = 3 - 1 = 2$$

$$Q = 8.40 > 5.991$$

Chi-squared critical values (χ^2)

Degrees of freedom (df)	α = .10	α = .05	α = .01
1	2.706	3.841	6.635
2	4.605	**5.991**	9.210
3	6.251	7.815	11.345
4	7.779	9.488	13.277
5	9.236	11.070	15.086

Conclusion: The Friedman test is statistically significant, $\chi^2(2) = 8.40$, $p < .05$. The warning conditions differ systematically in their rankings.

Effect size: Kendall's W

Kendall's W is a measure of effect size for the Friedman test:

- Measures the degree of agreement in how participants rank the conditions
- Ranges from 0 (no overall agreement) to 1 (complete agreement)
- Larger values indicate greater consistency in condition rankings across participants

$$W = \frac{Q}{n(k-1)}$$

Symbols:

- Q = Friedman test statistic (8.40)
- n = number of participants (5)
- k = number of conditions (3)

For our example:

$$W = \frac{8.40}{5(3-1)}$$

$$W = \frac{8.40}{10} = \mathbf{0.84}$$

Guidelines for interpreting Kendall's W

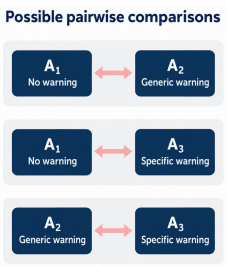
Kendall's W	Interpretation
Close to 0	Little agreement
Around .5	Moderate agreement
Close to 1	Strong agreement

- **These are descriptive guides rather than established cut-offs**
- For our example, W = .84 indicates strong agreement in how participants ranked the three warning conditions.

Follow-up comparisons

Why do we need follow-up comparisons?

- The Friedman test is an omnibus test: it evaluates whether conditions differ systematically in their rankings
- A significant Friedman result does not identify **which conditions differ**
- As with parametric within-participants ANOVA, we can investigate specific differences using planned comparisons or post-hoc tests



29

Planned comparisons: Wilcoxon signed-rank test

- The **Wilcoxon signed-rank test** is a non-parametric alternative to the paired-samples *t*-test
  - Compares two related conditions using paired observations
  - Calculates the difference between each participant's two scores
  - Ranks the absolute differences, excluding zero differences
  - Uses the signs of the differences to assess whether there is a systematic tendency for one condition to produce higher scores
- **Example:** We might predict that participants receiving a specific warning (A_3) will achieve higher accuracy than when receiving no warning (A_1)

30

Post-hoc comparisons

- Following a significant Friedman test, we may want to investigate all possible pairwise differences
  - Pairwise Wilcoxon signed-rank tests are a common post-hoc procedure
  - Each test compares scores from two related conditions
  - Adjustments for multiple comparisons help control the familywise error rate
- As in the previous lecture, we can use **Holm-adjusted *p*-values**

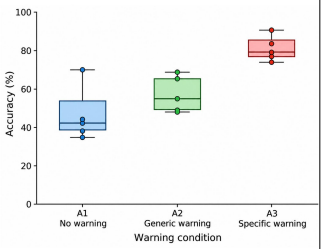
Interpreting follow-up comparisons

The Friedman omnibus test was statistically significant:

$$\chi^2(2) = 8.40, \quad p = .015$$

However, the pairwise Wilcoxon signed-rank tests, using exact two-sided *p*-values with Holm adjustment, produced the following results:

Comparison	Adjusted p	Significant?
A_1 vs A_2	.188	No
A_1 vs A_3	.188	No
A_2 vs A_3	.188	No



A significant omnibus test does not guarantee significant follow-up comparisons.

Within-Participants ANOVA versus Friedman

Feature	Within-Participants ANOVA	Friedman
Design	Repeated measures	Repeated measures
Data analysed	Original scores	Within-participant ranks
Primary comparison	Condition means	Condition rankings
Normality	Assumes normally distributed model errors	Does not require normality
Sphericity	Required for the usual uncorrected F test	Not required
Test statistic	F	Q (χ^2)
Effect size	Partial eta-squared	Kendall's W
Follow-up tests	Paired t-tests / adjusted comparisons	Wilcoxon signed-rank tests

Summary

- The Friedman test is a non-parametric alternative to single-factor within-participants ANOVA
- It ranks observations within each participant, rather than across the entire dataset
- The Friedman statistic assesses whether the conditions differ systematically in their rankings
- Wilcoxon signed-rank tests can be used to investigate specific differences between related conditions, with appropriate adjustments for multiple comparisons

That's all folks!