Non-parametric Tests for Ordinal Data

- Non-parametric tests used most commonly on ordinal data (ranks)
- See **HCI:ERP** for discussion on limitations
- Type of test depends on
  - Number of conditions → 2 | 3+
  - Design → between-subjects | within-subjects

Design	Conditions	
	2	3 or more
Between-subjects (independent samples)	Mann-Whitney U	Kruskal-Wallis
Within-subjects (correlated samples)	Wilcoxon Signed-Rank	Friedman

Non-parametric – Example #1

- Research question:
  - *Is there a difference in the political leaning of Mac users and PC users?*
- Method:
  - 10 *Mac* users and 10 *PC* users randomly selected and interviewed
  - Participants assessed on a 10-point linear scale for political leaning
    - 1 = very left
    - 10 = very right
- Data (next slide)

Data (Example #1)

- Means:
  - 3.7 (*Mac* users)
  - 4.5 (*PC* users)
- Data suggest *PC* users more right-leaning, but is the difference statistically significant?
- Data are ordinal (at least), ∴ a non-parametric test is used
- Which test? (see below)

Design	Conditions	
	2	3 or more
Between-subjects (independent samples)	Mann-Whitney U	Kruskal-Wallis
Within-subjects (correlated samples)	Wilcoxon Signed-Rank	Friedman

Mac Users	PC Users
2	4
3	6
2	5
4	4
9	8
2	3
5	4
3	2
4	4
3	5
3.7	**4.5**

Mann Whitney U Test[1]

Mann-Whitney U for Response
Grouping Variable: Category for Response

U	31.000
U Prime	69.000
Z-Value	-1.436
P-Value	.1509
Tied Z-Value	-1.469
Tied P-Value	.1418
# Ties	4

Corrected for ties

Test statistic: *U*

Normalized *z* (calculated from *U*)

p (probability of the observed data, given the null hypothesis)

Mann-Whitney Rank Info for Response
Grouping Variable: Category for Response

	Count	Sum Ranks	Mean Rank
MAC	10	86.000	8.600
PC	10	124.000	12.400

Conclusion:
The null hypothesis remains tenable: No difference in the political leaning of *Mac* users and *PC* users ($U = 31.0$, $p > .05$)

See **HCI:ERP** for complete details and discussion

[1] Output table created by *StatView* (now marketed as *JMP*, a product of SAS; www.sas.com)

Non-parametric – Example #2

- Research question:
  - *Do two new designs for media players differ in "cool appeal" for young users?*
- Method:
  - 10 young tech-savvy participants recruited and given demos of the two media players (MPA, MPB)
  - Participants asked to rate the media players for "cool appeal" on a 10-point linear scale
    - 1 = not cool at all
    - 10 = really cool
- Data (next slide)

Data (Example #2)

- Means
  - 6.4 (MPA)
  - 3.7 (MPB)
- Data suggest MPA has more "cool appeal", but is the difference statistically significant?
- Data are ordinal (at least), ∴ a non-parametric test is used
- Which test? (see below)

Participant	MPA	MPB
1	3	3
2	6	6
3	4	3
4	10	3
5	6	5
6	5	6
7	9	2
8	7	4
9	6	2
10	8	3
	6.4	**3.7**

Design	Conditions	
	2	3 or more
Between-subjects (independent samples)	Mann-Whitney U	Kruskal-Wallis
Within-subjects (correlated samples)	Wilcoxon Signed-Rank	Friedman

Wilcoxon Signed-Rank Test

Wilcoxon Signed Rank Test for MPA, MPB

# 0 Differences	2
# Ties	2
Z-Value	-2.240
P-Value	.0251
Tied Z-Value	-2.254
Tied P-Value	.0242

Test statistic: Normalized *z* score

p (probability of the observed data, given the null hypothesis)

Wilcoxon Rank Info for MPA, MPB

	Count	Sum Ranks	Mean Rank
# Ranks < 0	1	2.000	2.000
# Ranks > 0	7	34.000	4.857

Conclusion:
The null hypothesis is rejected: Media player A has more "cool appeal" than media player B ($z = -2.254$, $p < .05$).

See **HCI:ERP** for complete details and discussion

Non-parametric – Example #3

- Research question:
  - *Is age a factor in the acceptance of a new GPS device for automobiles?*
- Method
  - 8 participants recruited from each of three age categories: 20-29, 30-39, 40-49
  - Participants demo'd the new GPS device and then asked if they would consider purchasing it for personal use
  - They respond on a 10-point linear scale
    - 1 = definitely no
    - 10 = definitely yes
- Data (next slide)

Data (Example #3)

- Means
  - 7.1 (20-29)
  - 4.0 (30-39)
  - 2.9 (40-49)
- Data suggest differences by age, but are differences statistically significant?
- Data are ordinal (at least), ∴ a non-parametric is used
- Which test? (see below)

A20-29	A30-39	A40-49
9	7	4
9	3	5
4	5	5
9	3	2
6	2	2
3	1	1
8	4	2
9	7	2
7.1	**4.0**	**2.9**

Design	Conditions	
	2	3 or more
Between-subjects (independent samples)	Mann-Whitney U	Kruskal-Wallis
Within-subjects (correlated samples)	Wilcoxon Signed-Rank	Friedman

Kruskal-Wallis Test

Kruskal-Wallis Test for Acceptability
Grouping Variable: Category for Preference

DF	2
# Groups	3
# Ties	7
H	9.421
P-Value	.0090
H corrected for ties	9.605
Tied P-Value	.0082

Test statistic: *H* (follows chi-square distribution)

p (probability of the observed data, given the null hypothesis)

Conclusion:
The null hypothesis is rejected: There is an age difference in the acceptance of the new GPS device. ($\chi^2 = 9.605$, $p < .01$).

Kruskal-Wallis Rank Info for Acceptability
Grouping Variable: Category for Preference

	Count	Sum Ranks	Mean Rank
A	8	148.000	18.500
B	8	88.500	11.063
C	8	63.500	7.938

See **HCI:ERP** for complete details and discussion

Non-parametric – Example #4

- Research question:
  - *Do four variations of a search engine interface (A, B, C, D) differ in "quality of results"?*
- Method
  - 8 participants recruited and demo'd the four interfaces
  - Participants do a series of search tasks on the four search interfaces (Note: counterbalancing is used, but this isn't important here)
  - Quality of results for each search interface assessed on a linear scale from 1 to 100
    - 1 = very poor quality of results
    - 100 = very good quality of results
- Data (next slide)

Data (Example #4)

- Means
  - 71.0 (A), 68.1 (B), 60.9 (C), 69.8 (D)
- Data suggest a difference in quality of results, but are the differences statistically significant?
- Data are ordinal (at least), ∴ a non-parametric test is used
- Which test? (see below)

Participant	A	B	C	D
1	66	80	67	73
2	79	64	61	66
3	67	58	61	67
4	71	73	54	75
5	72	66	59	78
6	68	67	57	69
7	71	68	59	64
8	74	69	69	66
	71.0	**68.1**	**60.9**	**69.8**

Design	Conditions	
	2	3 or more
Between-subjects (independent samples)	Mann-Whitney U	Kruskal-Wallis
Within-subjects (correlated samples)	Wilcoxon Signed-Rank	Friedman

Friedman Test

Friedman Test for 4 Variables

DF	3
# Groups	4
# Ties	2
Chi Square	8.475
P-Value	.0372
Chi Square corrected for ties	8.692
Tied P-Value	.0337

Friedman Rank Info for 4 Variables

	Count	Sum Ranks	Mean Rank
A	8	24.500	3.063
B	8	19.500	2.438
C	8	11.500	1.438
D	8	24.500	3.063

Test statistic: *H* (follows chi-square distribution)

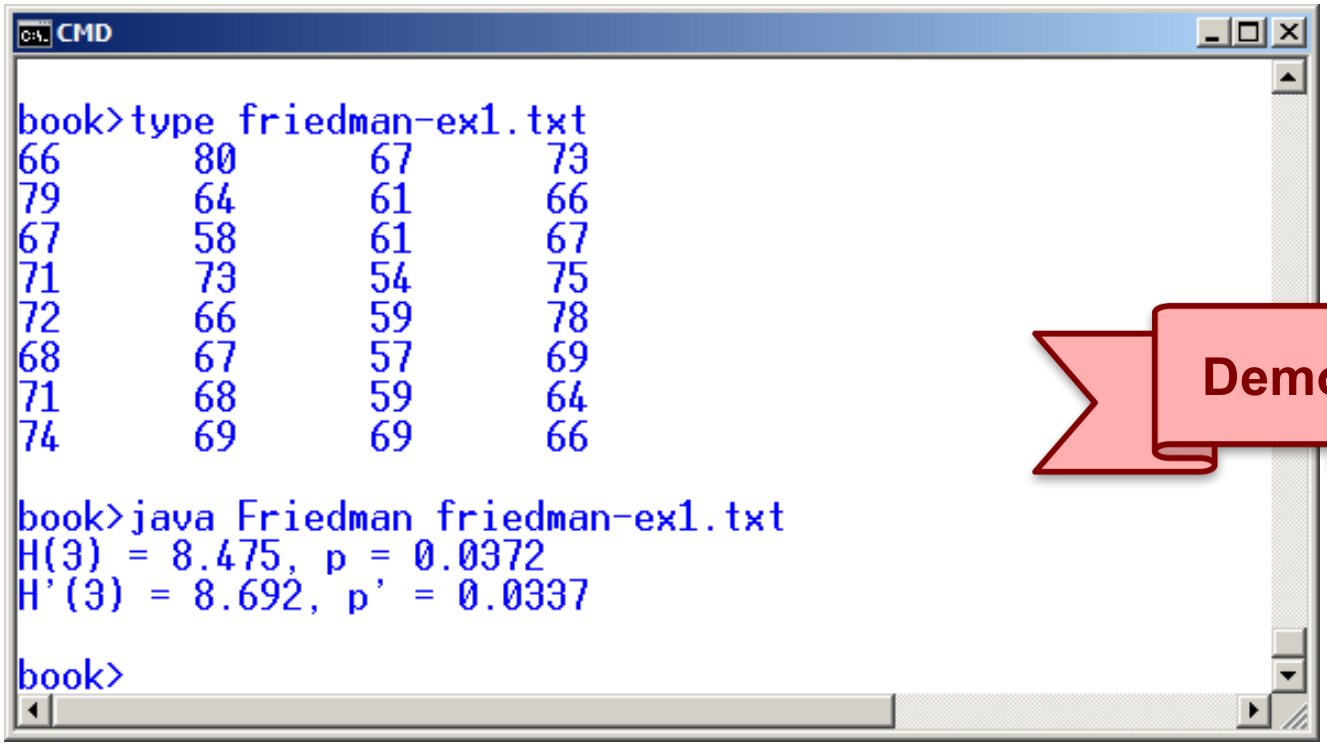
p (probability of the observed data, given the null hypothesis)

Conclusion:
The null hypothesis is rejected: There is a difference in the quality of results provided by the search interfaces ($\chi^2 = 8.692$, $p < .05$).

See **HCI:ERP** for complete details and discussion

`Friedman` Software

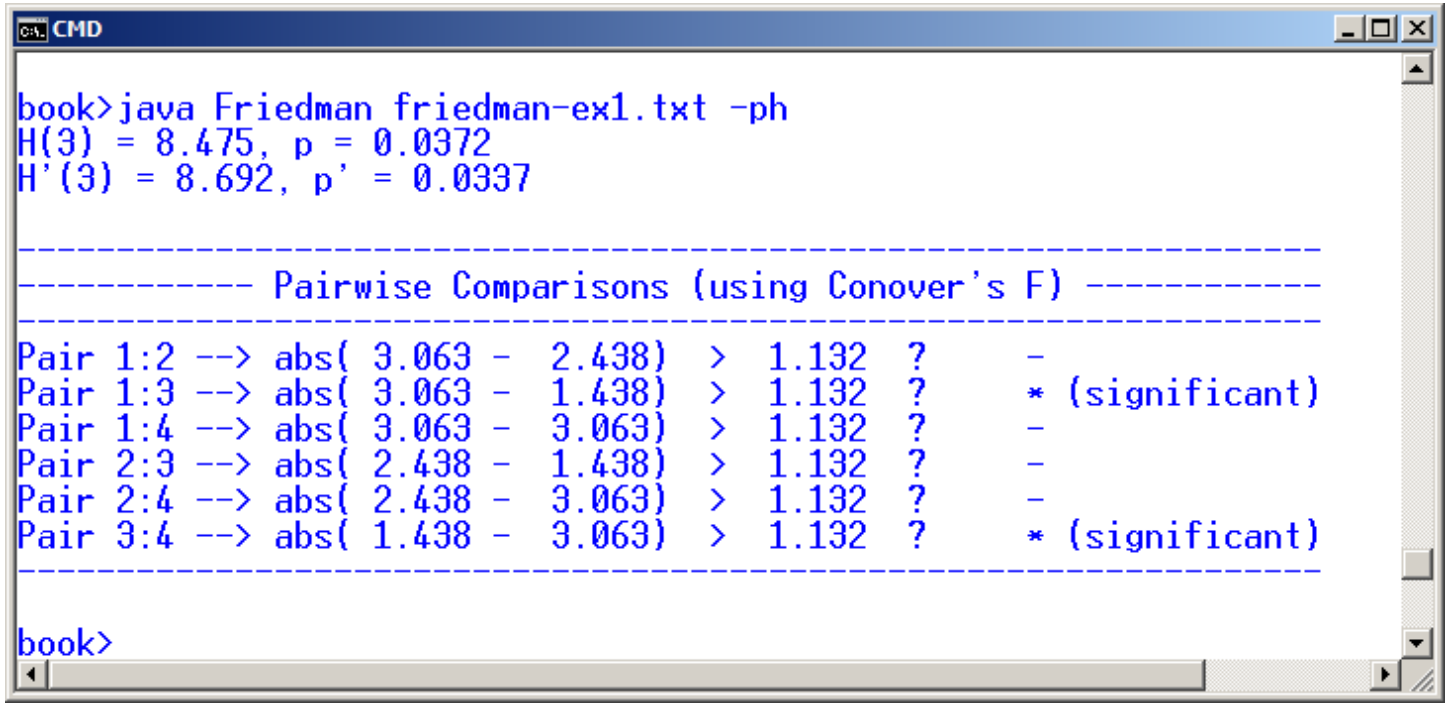
- Download `Friedman` Java software from **HCI:ERP** web site[1]



[1] `Friedman` files contained in `NonParametric.zip`.

Post Hoc Comparisons

- As with `KruskalWallis` application, available using the `-ph` option…

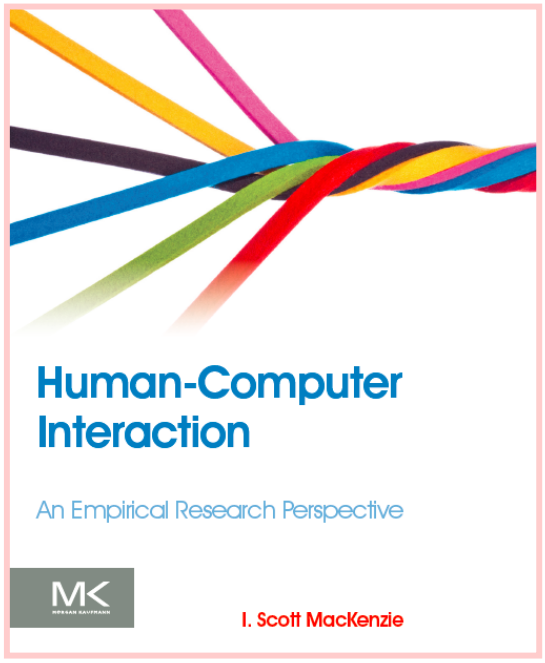


Points of Discussion

- Reporting the mean vs. median for scaled responses
- Non-parametric tests for multi-factor experiments
- Non-parametric tests for ratio-scale data

See **HCI:ERP** for complete details and discussion

Thank You


Human-Computer Interaction
An Empirical Research Perspective
MK
I. Scott MacKenzie