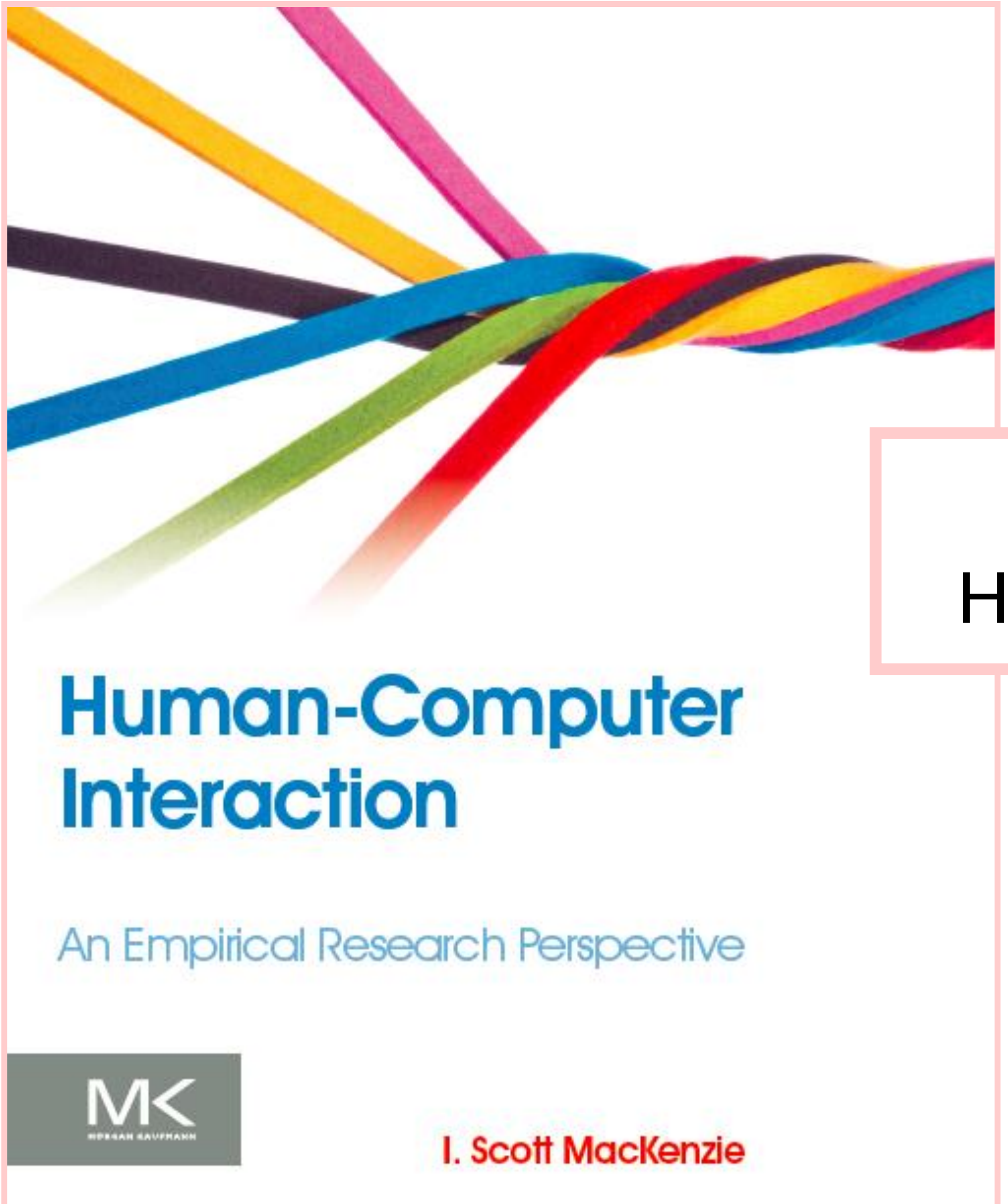


Chapter 6
Hypothesis Testing

Standard Deviation

$$\sigma = \sqrt{\frac{\sum (x - \overline{x})^2}{n}}$$

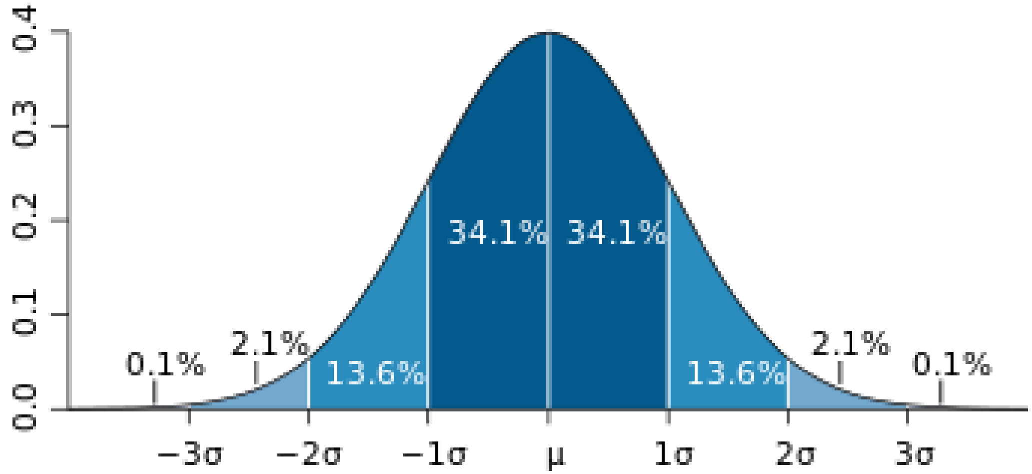
σ = standard deviation

$\sum$ = sum of

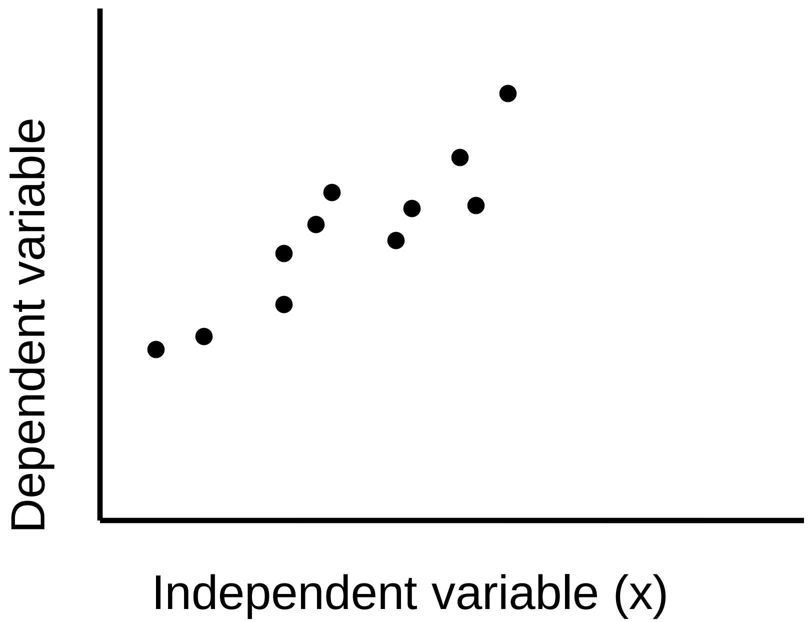
x = each value in the data set

$\overline{x}$ = mean of all values in the data set

n = number of value in the data set



Regression

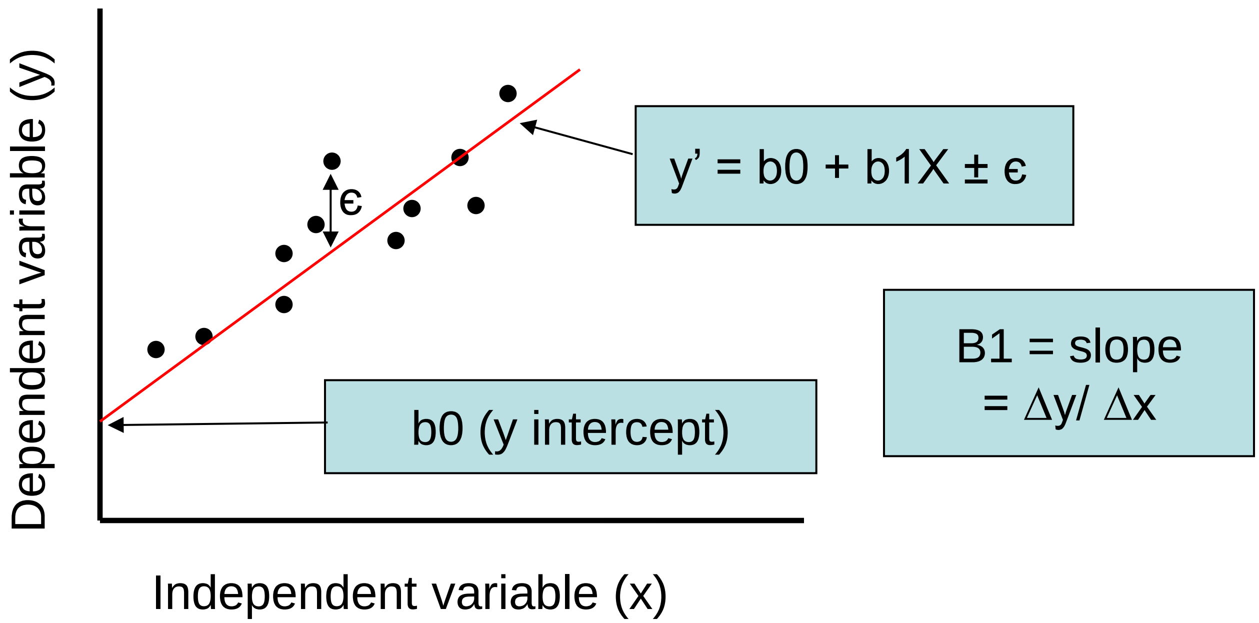


Regression is the attempt to explain the variation in a dependent variable using the variation in independent variables.

Regression is thus an explanation of causation.

If the independent variable(s) sufficiently explain the variation in the dependent variable, the model can be used for prediction.

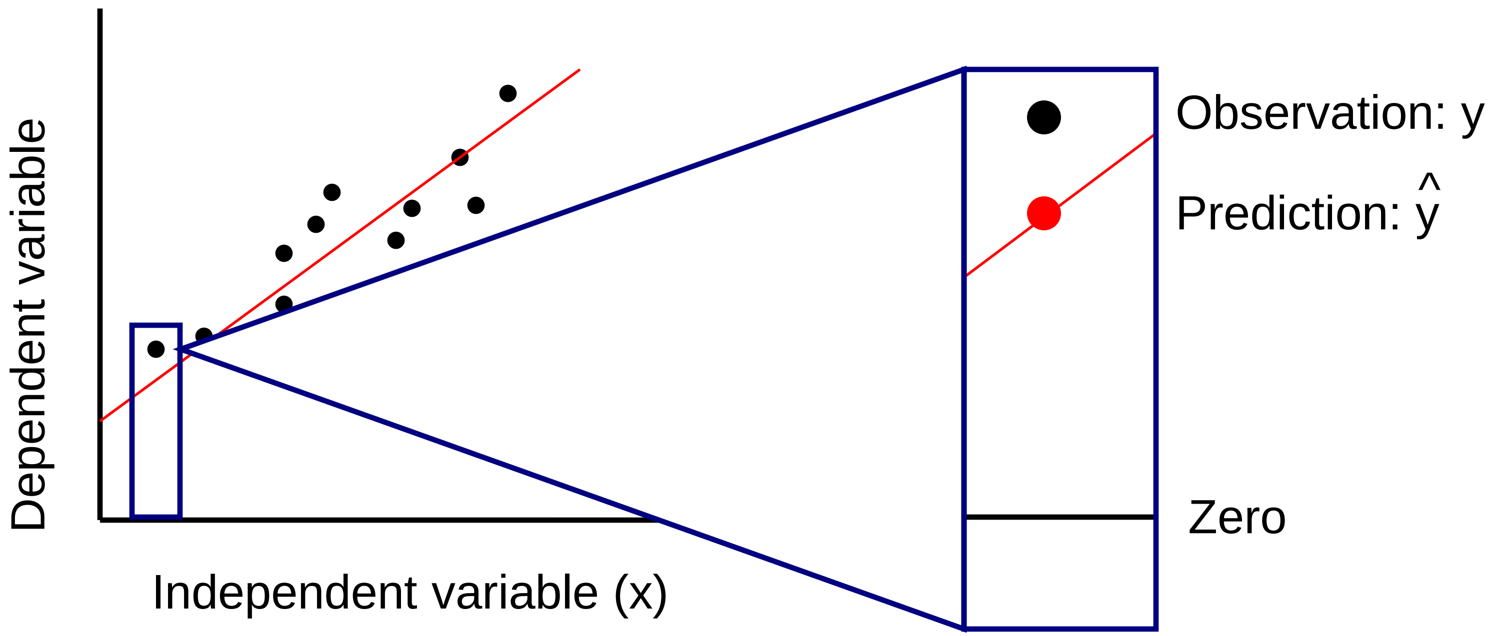
Simple Linear Regression



The output of a regression is a function that predicts the dependent variable based upon values of the independent variables.

Simple regression fits a straight line to the data.

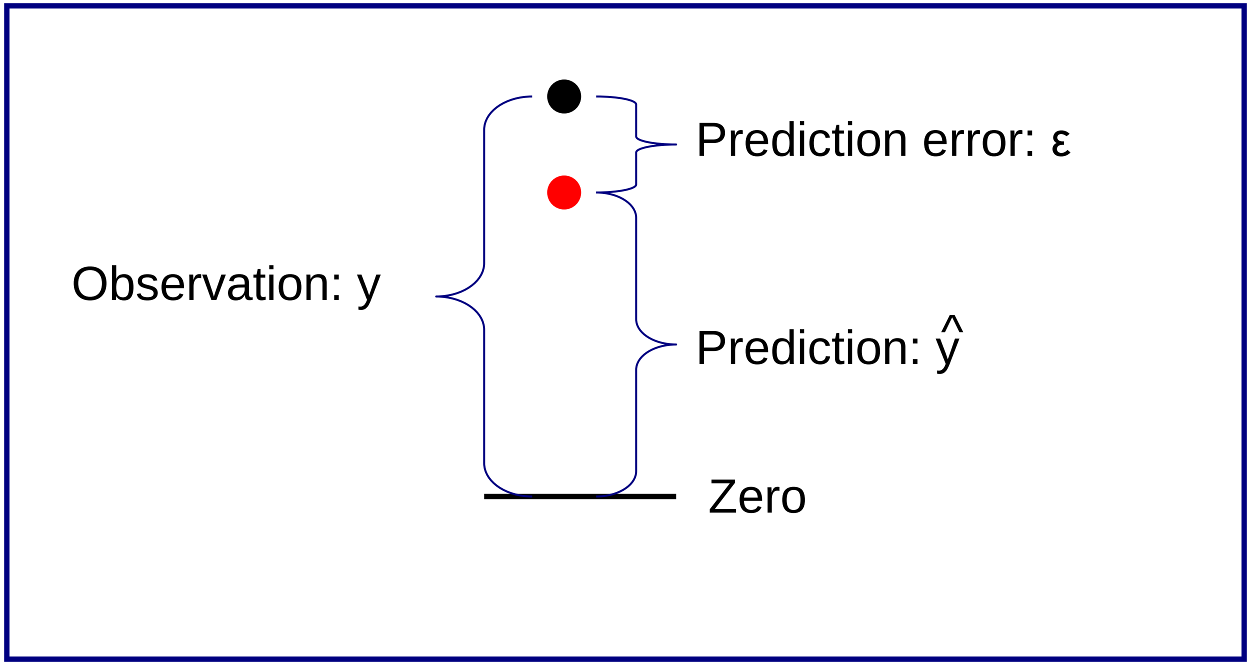
Simple Linear Regression



The function will make a prediction for each observed data point.

The observation is denoted by y and the prediction is denoted by $\hat{y}$.

Simple Linear Regression

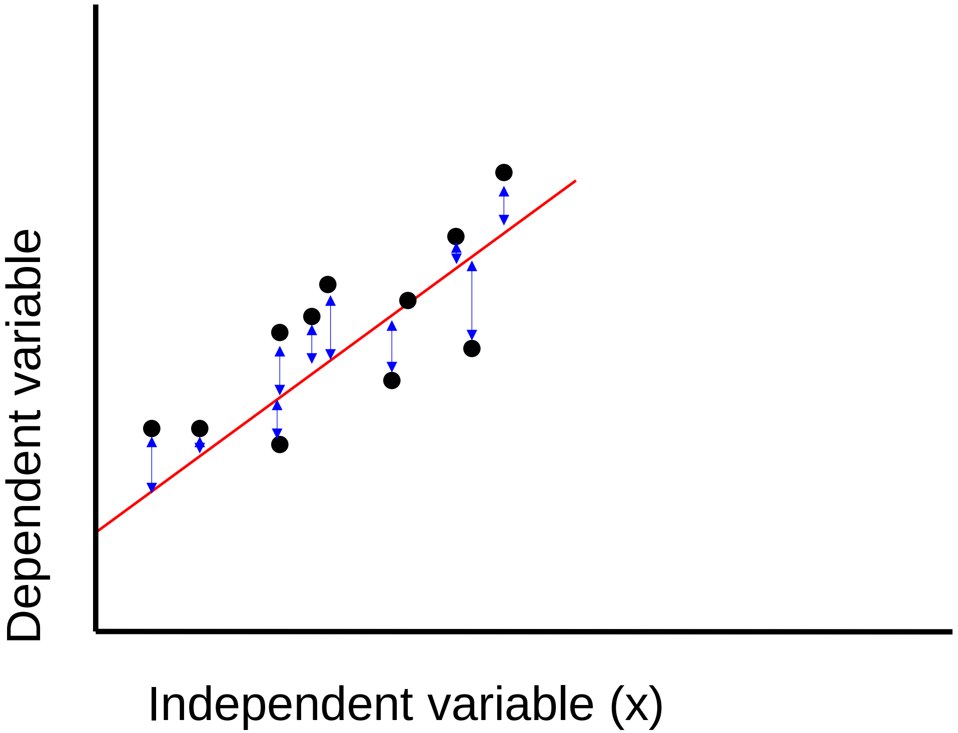


For each observation, the variation can be described as:

$$\mathbf{y = \hat{y} + \varepsilon}$$

Actual = Explained + Error

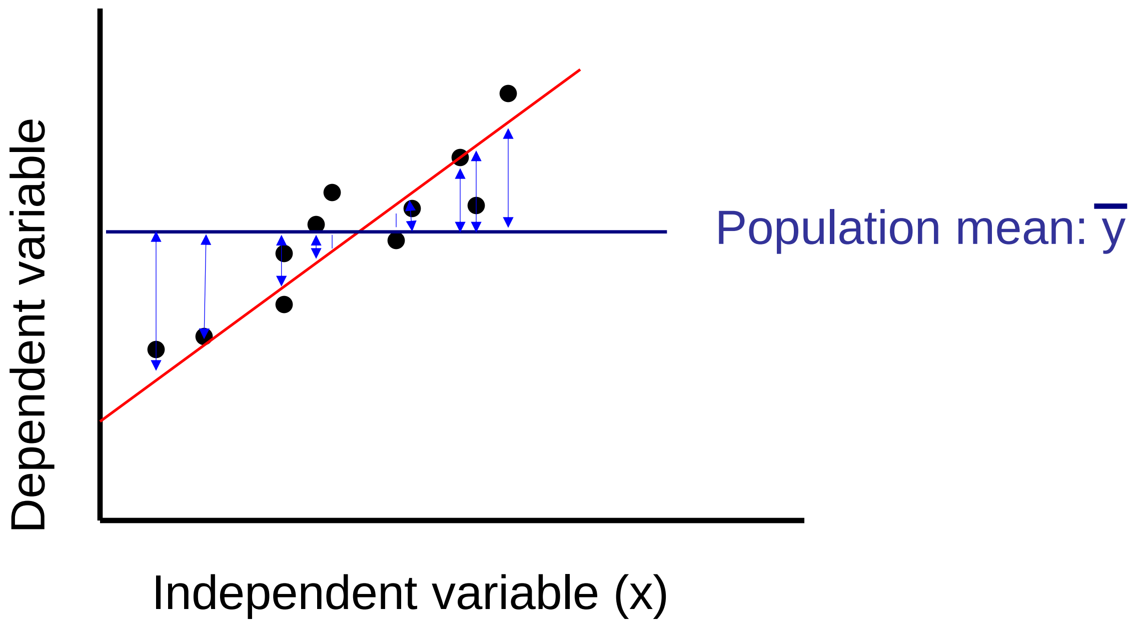
Regression



A least squares regression selects the line with the lowest total sum of squared prediction errors.

This value is called the Sum of Squares of Error, or SSE.

Calculating SSR



The Sum of Squares Regression (SSR) is the sum of the squared differences between the prediction for each observation and the population mean.

Regression Formulas

The Total Sum of Squares (SST) is equal to SSR + SSE.

Mathematically,

SSR = $\sum (\hat{y} - \bar{y})^2$ **(measure of explained variation)**

SSE = $\sum (y - \hat{y})$ **(measure of unexplained variation)**

SST = SSR + SSE = $\sum (y - \bar{y})^2$ **(measure of total variation in y)**

What is Hypothesis Testing?

- … the use of statistical procedures to answer research questions
- Typical research question (generic):

> Is the time to complete a task less using Method A than using Method B?

- For hypothesis testing, research questions are statements:

> There is no difference in the mean time to complete a task using Method A vs. Method B.

- This is the *null hypothesis* (assumption of “no difference”)
- Statistical procedures seek to reject or accept the null hypothesis (details to follow)

Statistical Procedures

- Two types:
  - Parametric
    - Data are assumed to come from a distribution, such as the normal distribution, *t*-distribution, etc.
  - Non-parametric
    - Data are not assumed to come from a distribution
  - Lots of debate on assumptions testing and what to do if assumptions are not met (avoided here, for the most part)
  - A reasonable basis for deciding on the most appropriate test is to match the type of test with the measurement scale of the data (next slide)

Measurement Scales vs. Statistical Tests

Measurement Scale	Defining Relations	Examples of Appropriate Statistics	Appropriate Statistical Tests
Nominal	• Equivalence	• Mode • Frequency	• Non-parametric tests
Ordinal	• Equivalence • Order	• Median • Percentile	
Interval	• Equivalence • Order • Ratio of intervals	• Mean • Standard deviation	• Parametric tests • Non-parametric tests
Ratio	• Equivalence • Order • Ratio of intervals • Ratio of values	• Geometric mean • Coefficient of variation	

- Parametric tests most appropriate for…
  - Ratio data, interval data
- Non-parametric tests most appropriate for…
  - Ordinal data, nominal data (although limited use for ratio and interval data)

Tests Presented Here

- Parametric
  - Analysis of variance (ANOVA)
    - Used for ratio data and interval data
    - Most common statistical procedure in HCI research
- Non-parametric
  - Chi-square test
    - Used for nominal data
  - Mann-Whitney U, Wilcoxon Signed-Rank, Kruskal-Wallis, and Friedman tests
    - Used for ordinal data

Analysis of Variance

- The *analysis of variance* (ANOVA) is the most widely used statistical test for hypothesis testing in factorial experiments
- Goal → determine if an independent variable has a significant effect on a dependent variable
- Remember, an independent variable has at least two levels (test conditions)
- Goal (put another way) → determine if the test conditions yield different outcomes on the dependent variable (e.g., one of the test conditions is faster/slower than the other)

Why Analyse the Variance?

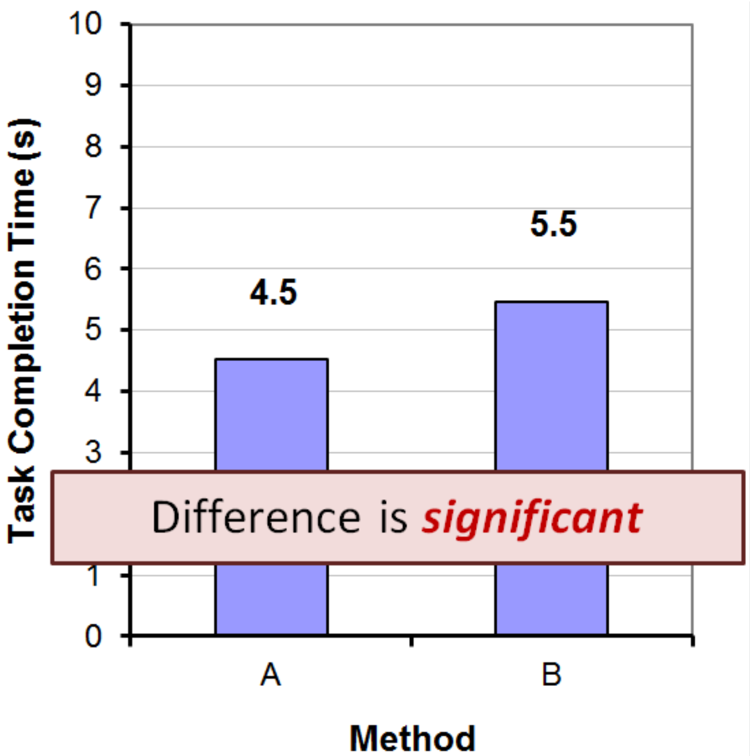
- Seems odd that we analyse the variance, but the research question is concerned with the overall means:

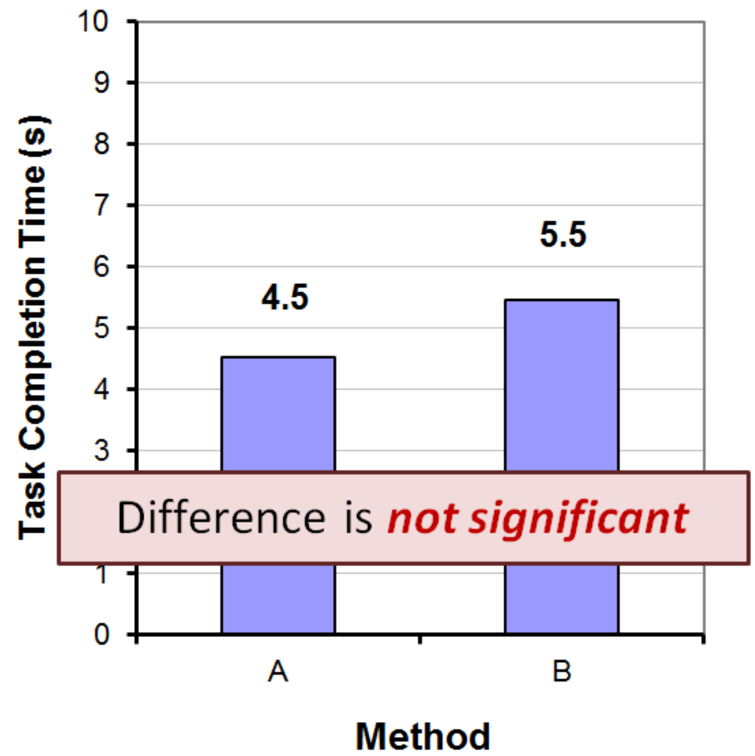
> Is the time to complete a task less using Method A than using Method B?

- Let's explain through two simple examples (next slide)

Example #1

Example #2



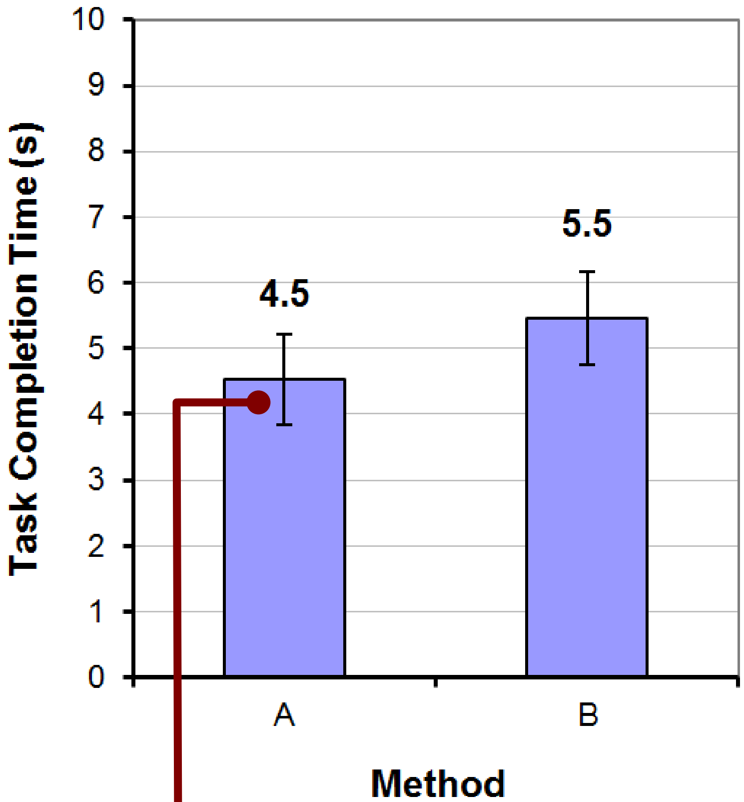


"Significant" implies that in all likelihood the difference observed is due to the test conditions (Method A vs. Method B).

"Not significant" implies that the difference observed is likely due to chance.

File: `06-AnovaDemo.xlsx`

Example #1 - Details



Note: Within-subjects design

Participant	Method	
	A	B
1	5.3	5.7
2	3.6	4.8
3	5.2	5.1
4	3.6	4.5
5	4.6	6.0
6	4.1	6.8
7	4.0	6.0
8	4.8	4.6
9	5.2	5.5
10	5.1	5.6
Mean	4.5	5.5
SD	0.68	0.72

Error bars show
±1 standard deviation

Note: *SD* is the square root of the variance

Example #1 – ANOVA[1]

ANOVA Table for Task Completion Time (s)

	DF	Sum of Squares	Mean Square	F-Value	P-Value	Lambda	Power
Subject	9	5.080	.564				
Method	1	4.232	4.232	9.796	.0121	9.796	.804
Method * Subject	9	3.888	.432				

Probability of obtaining the observed data if the null hypothesis is true

Reported as…

$F_{1,9} = 9.80$, $p < .05$

Thresholds for "p"

- .05
- .01
- .005
- .001
- .0005
- .0001

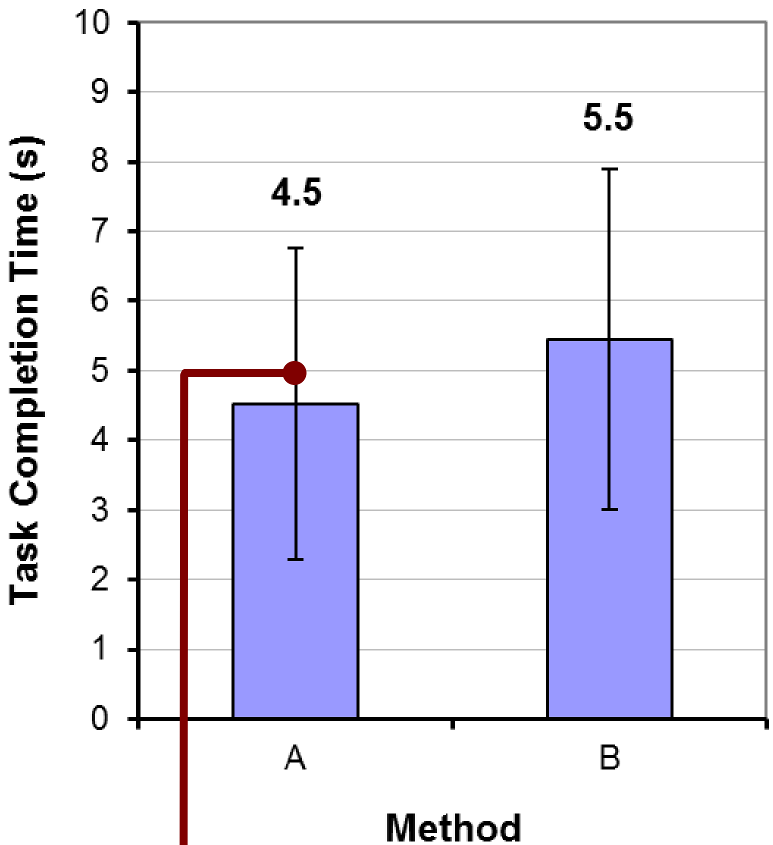
[1] ANOVA table created by *StatView* (now marketed as *JMP*, a product of SAS; www.sas.com)

How to Report an *F*-statistic

The mean task completion time for Method A was 4.5 s. This was 20.1% less than the mean of 5.5 s observed for Method B. The difference was statistically significant ($F_{1,9} = 9.80$, $p < .05$).

- Notice in the parentheses
  - Uppercase for *F*
  - Lowercase for *p*
  - Italics for *F* and *p*
  - Space both sides of equal sign
  - Space after comma
  - Space on both sides of less-than sign
  - Degrees of freedom are subscript, plain, smaller font
  - Three significant figures for *F* statistic
  - No zero before the decimal point in the *p* statistic (except in Europe)

Example #2 - Details



Participant	Method	
	A	B
1	2.4	6.9
2	2.7	7.2
3	3.4	2.6
4	6.1	1.8
5	6.4	7.8
6	5.4	9.2
7	7.9	4.4
8	1.2	6.6
9	3.0	4.8
10	6.6	3.1
Mean	4.5	5.5
SD	2.23	2.45

Error bars show
±1 standard deviation

Example #2 – ANOVA

ANOVA Table for Task Completion Time (s)

	DF	Sum of Squares	Mean Square	F-Value	P-Value	Lambda	Power
Subject	9	37.372	4.152				
Method	1	4.324	4.324	.626	.4491	.626	.107
Method * Subject	9	62.140	6.904				

Probability of obtaining the observed data if the null hypothesis is true

Note: For non-significant effects, use "ns" if $F < 1.0$, or "$p > .05$" if $F > 1.0$.

Reported as…

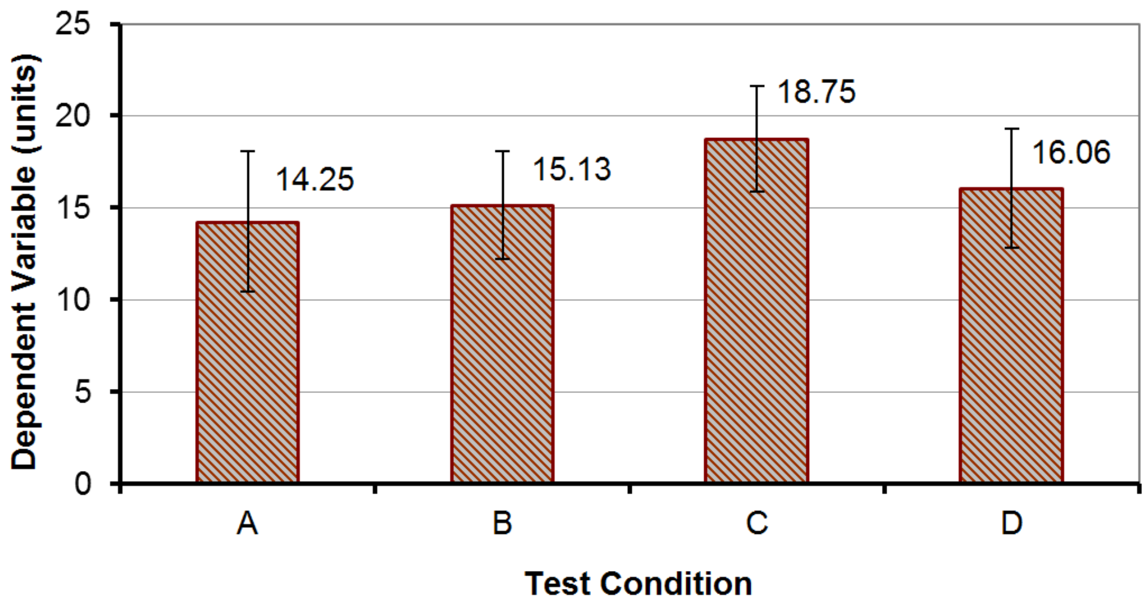
$F_{1,9} = 0.626$, ns

Example #2 - Reporting

The mean task completion times were 4.5 s for Method A and 5.5 s for Method B. As there was substantial variation in the observations across participants, the difference was not statistically significant as revealed in an analysis of variance ($F_{1,9}$ = 0.626, ns).

More Than Two Test Conditions

Participant	Test Condition			
	A	B	C	D
1	11	11	21	16
2	18	11	22	15
3	17	10	18	13
4	19	15	21	20
5	13	17	23	10
6	10	15	15	20
7	14	14	15	13
8	13	14	19	18
9	19	18	16	12
10	10	17	21	18
11	10	19	22	13
12	16	14	18	20
13	10	20	17	19
14	10	13	21	18
15	20	17	14	18
16	18	17	17	14
Mean	14.25	15.13	18.75	16.06
SD	3.84	2.94	2.89	3.23



ANOVA

ANOVA Table for Dependent Variable (units)

	DF	Sum of Squares	Mean Square	F-Value	P-Value	Lambda	Power
Subject	15	81.109	5.407				
Test Condition	3	182.172	60.724	4.954	.0047	14.862	.896
Test Condition * Subject	45	551.578	12.257				

- There was a significant effect of Test Condition on the dependent variable ($F_{3,45} = 4.95$, $p < .005$)
- Degrees of freedom
  - If n is the number of test conditions and m is the number of participants, the degrees of freedom are…
  - Effect → $(n - 1)$
  - Residual → $(n - 1)(m - 1)$
  - Note: single-factor, within-subjects design

Post Hoc Comparisons Tests

- A significant *F*-test means that at least one of the test conditions differed significantly from one other test condition
- Does not indicate which test conditions differed significantly from one another
- To determine which pairs differ significantly, a post hoc comparisons tests is used
- Examples:
  - Fisher PLSD, Bonferroni/Dunn, Dunnett, Tukey/Kramer, Games/Howell, Student-Newman-Keuls, orthogonal contrasts, Scheffé
- Scheffé test on next slide

Scheffé Post Hoc Comparisons

Scheffe for Dependent Variable (units)
Effect: Test Condition
Significance Level: 5 %

	Mean Diff.	Crit. Diff.	P-Value	
A, B	-.875	3.302	.9003	
A, C	-4.500	3.302	.0032	S
A, D	-1.813	3.302	.4822	
B, C	-3.625	3.302	.0256	S
B, D	-.938	3.302	.8806	
C, D	2.688	3.302	.1520	

- Test conditions A:C and B:C differ significantly (see chart three slides back)

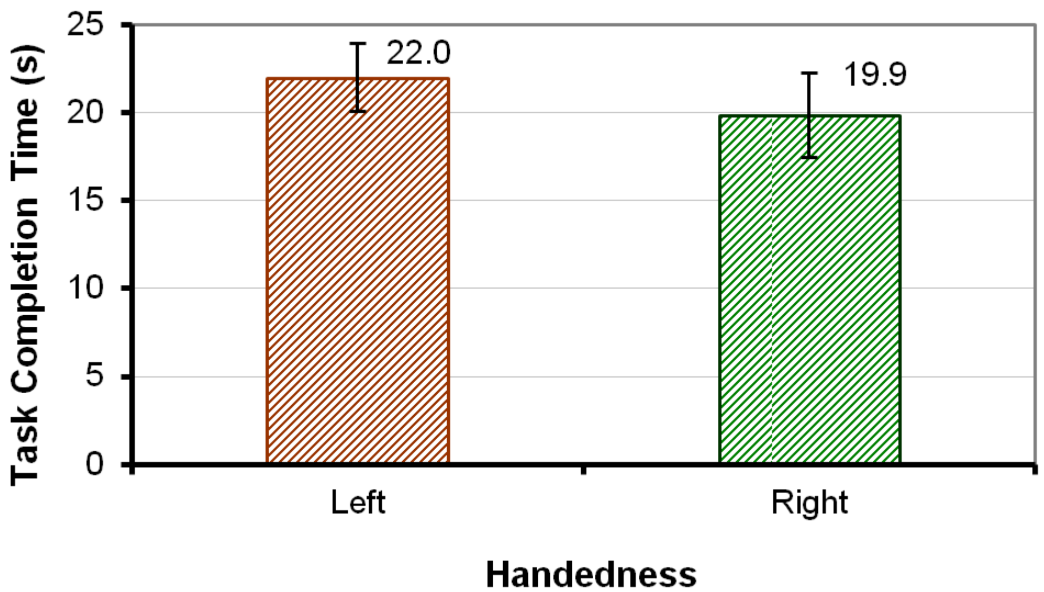
Between-subjects Designs

- Research question:
  - *Do left-handed users and right-handed users differ in the time to complete an interaction task?*
- The independent variable (handedness) must be assigned between-subjects
- Example data set →

Participant	Task Completion Time (s)	Handedness
1	23	L
2	19	L
3	22	L
4	21	L
5	23	L
6	20	L
7	25	L
8	23	L
9	17	R
10	19	R
11	16	R
12	21	R
13	23	R
14	20	R
15	22	R
16	21	R
Mean	20.9	
SD	2.38	

Summary Data and Chart

Handedness	Task Completion Time (s)	
	Mean	*SD*
Left	22.0	1.93
Right	19.9	2.42



ANOVA

ANOVA Table for Task Completion Time (s)

	DF	Sum of Squares	Mean Square	F-Value	P-Value	Lambda	Power
Handedness	1	18.063	18.063	3.781	.0722	3.781	.429
Residual	14	66.875	4.777				

- The difference was not statistically significant ($F_{1,14}$ = 3.78, p > .05)
- Degrees of freedom:
  - Effect → (n – 1)
  - Residual → (m – n)
  - Note: single-factor, between-subjects design

Two-way ANOVA

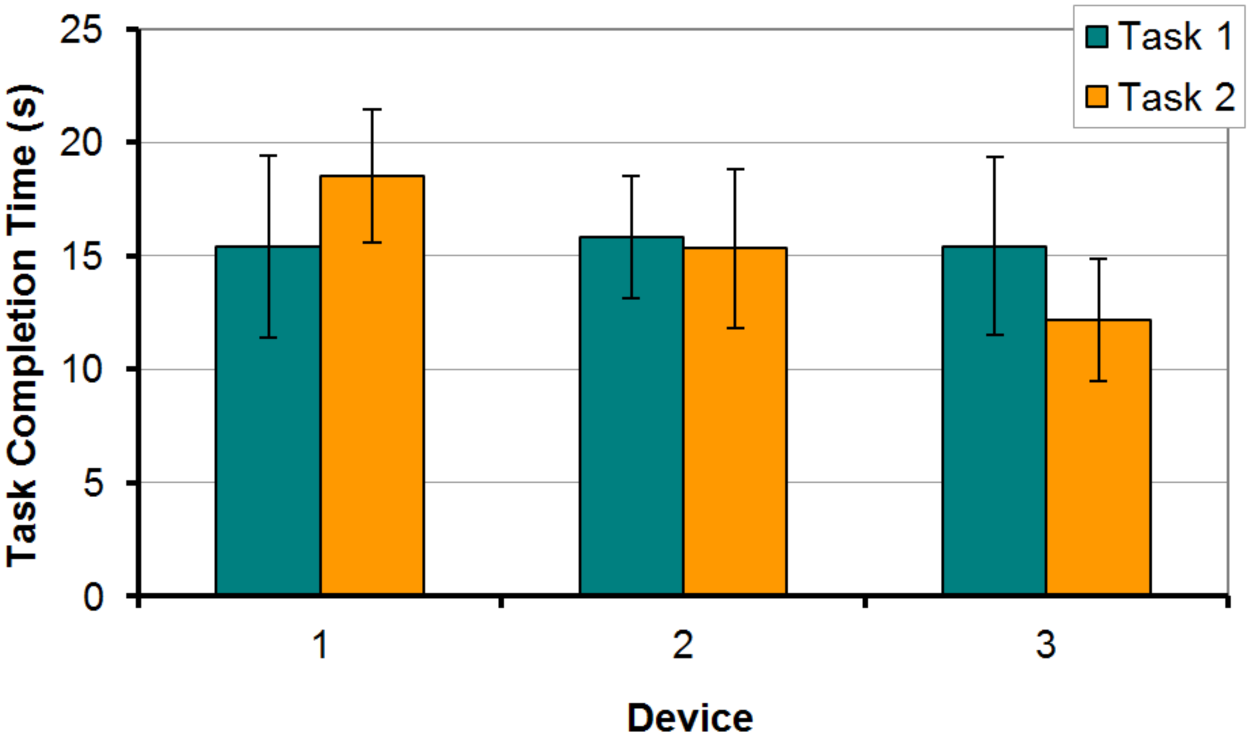
- An experiment with two independent variables is a *two-way design*
- ANOVA tests for
  - Two main effects + one interaction effect
- Example
  - Independent variables
    - Device → D1, D2, D3 (e.g., mouse, stylus, touchpad)
    - Task → T1, T2 (e.g., point-select, drag-select)
  - Dependent variable
    - Task completion time (or something, this isn't important here)
  - Both IVs assigned within-subjects
  - Participants: 12
  - Data set (next slide)

Data Set

Participant	Device 1		Device 2		Device 3	
	Task 1	Task 2	Task 1	Task 2	Task 1	Task 2
1	11	18	15	13	20	14
2	10	14	17	15	11	13
3	10	23	13	20	20	16
4	18	18	11	12	11	10
5	20	21	19	14	19	8
6	14	21	20	11	17	13
7	14	16	15	20	16	12
8	20	21	18	20	14	12
9	14	15	13	17	16	14
10	20	15	18	10	11	16
11	14	20	15	16	10	9
12	20	20	16	16	20	9
Mean	15.4	18.5	15.8	15.3	15.4	12.2
SD	4.01	2.94	2.69	3.50	3.92	2.69

Summary Data and Chart

	Task 1	Task 2	*Mean*
Device 1	15.4	18.5	17.0
Device 2	15.8	15.3	15.6
Device 3	15.4	12.2	13.8
Mean	15.6	15.3	15.4



ANOVA

ANOVA Table for Task Completion Time (s)

	DF	Sum of Squares	Mean Square	F-Value	P-Value	Lambda	Power
Subject	11	134.778	12.253				
Device	2	121.028	60.514	5.865	.0091	11.731	.831
Device * Subject	22	226.972	10.317				
Task	1	.889	.889	.076	.7875	.076	.057
Task * Subject	11	128.111	11.646				
Device * Task	2	121.028	60.514	5.435	.0121	10.869	.798
Device * Task * Subject	22	244.972	11.135				

Can you pull the relevant statistics from this chart and craft statements indicating the outcome of the ANOVA?

ANOVA - Reporting

The grand mean for task completion time was 15.4 seconds. Device 3 was the fastest at 13.8 seconds, while device 1 was the slowest at 17.0 seconds. The main effect of device on task completion time was statistically significant ($F_{2,22} = 5.865$, $p < .01$). The task effect was modest, however. Task completion time was 15.6 seconds for task 1. Task 2 was slightly faster at 15.3 seconds; however, the difference was not statistically significant ($F_{1,11} = 0.076$, ns). The results by device and task are shown in Figure x. There was a significant Device $\times$ Task interaction effect ($F_{2,22} = 5.435$, $p < .05$), which was due solely to the difference between device 1 task 2 and device 3 task 2, as determined by a Scheffé post hoc analysis.

`Anova2` Software

- **HCI:ERP** web site includes analysis of variance Java software: `Anova2`
- Operates from command line on data in a text file
- Extensive API with demos, data files, discussions, etc.
- Download and demonstrate

CMD

```
text>java Anova2
-----------------------------------------------------------------------
Usage: java Anova2 file p f1 f2 f3 [-a] [-d] [-m] [-h]

  file = data file (comma or space delimited)
  p = # of rows (participants) in data file
  f1 = # of levels, 1st within-subjects factor ("." if not used)
  f2 = # of levels, 2nd within-subjects factor ("." if not used)
  f3 = # of levels, between-subjects factor ("." if not used)
  -a = output anova table
  -d = output debug data
  -m = output main effect means
  -h = data file includes header lines (see API for details)
  (Note: default is no output)
-----------------------------------------------------------------------

text>
```

Dix et al. Example[1]

- Single-factor, within-subjects design
- See API for discussion

```
book>type dix-example-10x2.txt
656,702
259,339
612,658
609,645
1049,1129
1135,1179
542,604
495,551
905,893
715,803

book>
```

```
book>java Anova2 dix-example-10x2.txt 10 2 . . -a

============================================================================
Effect                   df        SS              MS            F         p
----------------------------------------------------------------------------
Participant               9   1231492.000     136832.444
F1                        1     13833.800      13833.800     33.359    2.7E-4
F1_x_Par                  9      3732.200        414.689
============================================================================

book>
```

[1] Dix, A., Finlay, J., Abowd, G., & Beale, R. (2004). *Human-computer interaction* (3rd ed.). London: Prentice Hall. (p. 337)

Dix et al. Example

- With counterbalancing
- Treating "Group" as a between-subjects factor[1]
- Includes header lines

```
book>type dix-example-h10x2b.txt
DV: Completion Time (s)
F1: Icon Type, Natural, Abstract
F2: .
F3: Group
656,702,NA
259,339,NA
612,658,NA
609,645,NA
1049,1129,NA
1135,1179,AN
542,604,AN
495,551,AN
905,893,AN
715,803,AN

book>
```

```
book>java Anova2 dix-example-h10x2b.txt 10 2 . 2 -h -a

ANOVA_table_for_Completion Time (s)
============================================================================
Effect                  df          SS            MS           F        p
----------------------------------------------------------------------------
Group                    1     67744.800     67744.800     0.466   0.51424
Participant(Group)       8   1163747.200    145468.400
Icon Type                1     13833.800     13833.800    30.680   3.6E-4
Icon Type_x_Group        1       125.000       125.000     0.277   0.61281
Icon Type_x_P(Group)     8      3607.200       450.900
============================================================================

book>
```

[1] See API and **HCI:ERP** for discussion on "counterbalancing and testing for a group effect".

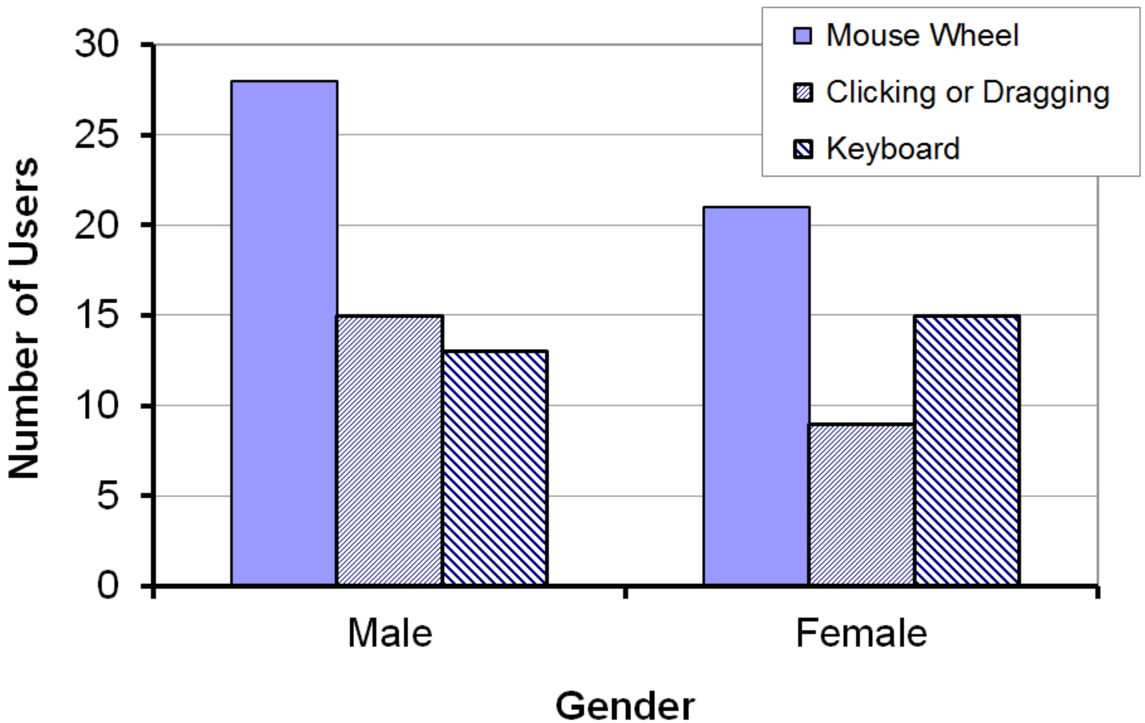
Chi-square Test (Nominal Data)

- A *chi-square test* is used to investigate relationships
- Relationships between categorical, or nominal-scale, variables representing attributes of people, interaction techniques, systems, etc.
- Data organized in a *contingency table* – cross tabulation containing counts (frequency data) for number of observations in each category
- A chi-square test compares the *observed values* against *expected values*
- Expected values assume "no difference"
- Research question:
  - *Do males and females differ in their method of scrolling on desktop systems?* (next slide)

Chi-square – Example #1

Observed Number of Users				
Gender	Scrolling Method			Total
	MW	CD	KB	
Male	28	15	13	56
Female	21	9	15	45
Total	49	24	28	101

MW = mouse wheel
CD = clicking, dragging
KB = keyboard



Chi-square – Example #1

Expected Number of Users				
Gender	Scrolling Method			Total
	MW	CD	KB	
Male	27.2	13.3	15.5	56.0
Female	21.8	10.7	12.5	45.0
Total	49.0	24.0	28.0	101

Chi Squares				
Gender	Scrolling Method			Total
	MW	CD	KB	
Male	0.025	0.215	0.411	0.651
Female	0.032	0.268	0.511	0.811
Total	0.057	0.483	0.922	**1.462**

Significant if it exceeds critical value (next slide)

$$\chi^2 = 1.462$$

(See **HCI:ERP** for calculations)

Chi-square Critical Values

- Decide in advance on *alpha* (typically .05)
- Degrees of freedom
  - $df = (r - 1)(c - 1) = (2 - 1)(3 - 1) = 2$
  - r = number of rows, c = number of columns

Significance Threshold (α)	Degrees of Freedom							
	1	2	3	4	5	6	7	8
.1	2.71	4.61	6.25	7.78	9.24	10.65	12.02	13.36
.05	3.84	$\boxed{5.99}$	7.82	9.49	11.07	12.59	14.07	15.51
.01	6.64	9.21	11.35	13.28	15.09	16.81	18.48	20.09
.001	10.83	13.82	16.27	18.47	20.52	22.46	24.32	26.13

$\chi^2 = 1.462$ (< 5.99 $\therefore$ not significant)

`ChiSquare` Software

- Download `ChiSquare` software from **HCI:ERP**
- Note: calculates *p* (assuming α = .05)

CMD

```
text>type chisquare-ex1.txt
28   15   13
21    9   15

text>java ChiSquare chisquare-ex1.txt
Chi-square(2) = 1.462
p = 0.4814

text>
```

Demo

Chi-square – Example #2

- Research question:
  - *Do students, professors, and parents differ in their responses to the question: Students should be allowed to use mobile phones during classroom lectures?*
- Data:

Observed Number of People				
Opinion	Category			Total
	Student	Professor	Parent	
Agree	10	12	98	120
Disagree	30	48	102	180
Total	40	60	200	300

Chi-square – Example #2

- Result: significant difference in responses ($\chi^2 = 20.5$, $p < .0001$)
- Post hoc comparisons reveal that opinions differ between students:parents and professors:parents (students:professors do not differ significantly in their responses)

```
text>type chisquare-ex2.txt
10   12   98
30   48   102

text>java ChiSquare chisquare-ex2.txt -ph
Chi-square(2) = 20.500
p = 0.0000

----------------------------------------------------
------ Pairwise Comparisons (using contrasts) ------
----------------------------------------------------
Pair 1:2   --->   Chi-square(2) =  0.340, p = 0.8437
Pair 1:3   --->   Chi-square(2) =  9.702, p = 0.0078
Pair 2:3   --->   Chi-square(2) = 21.475, p = 0.0000
----------------------------------------------------

text>
```

1 = students, 2 = professors, 3 = parents

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)

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**

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

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

Test statistic: *U*

Normalized *z* (calculated from *U*)

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

Corrected for ties

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

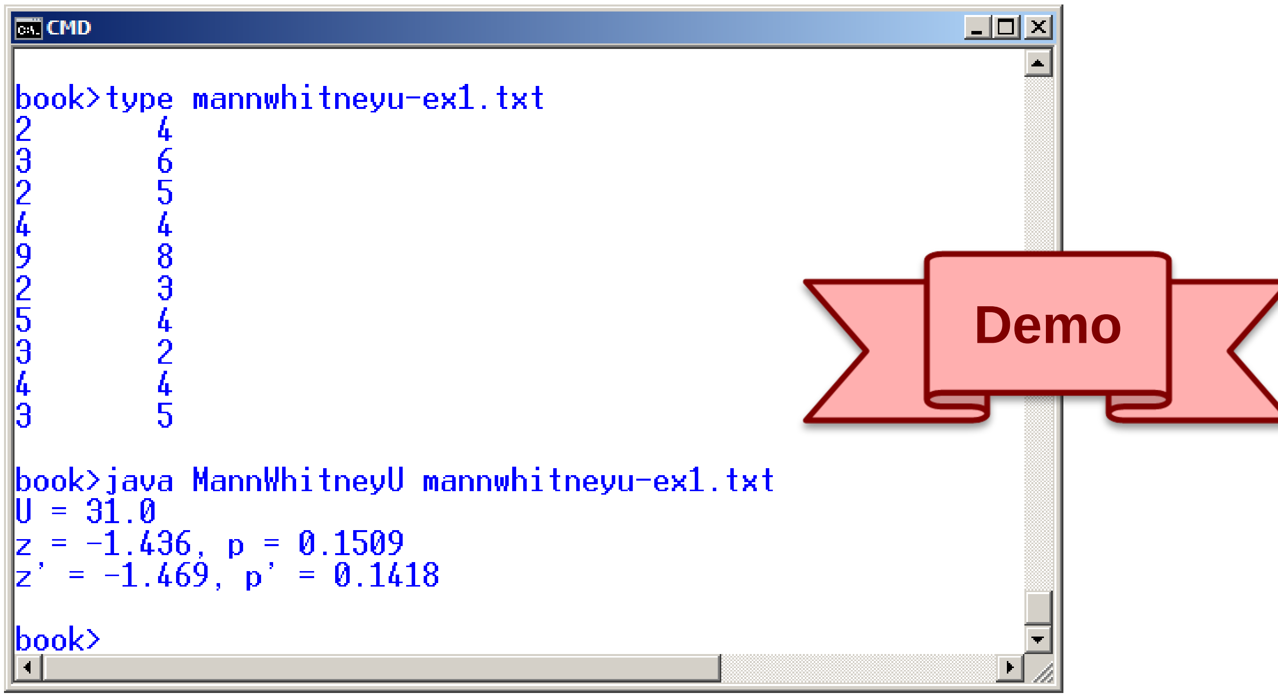
<u>Conclusion</u>:
The null hypothesis remains tenable: No difference in the political leaning of *Mac* users and *PC* users (U = 31.0, p > .05)

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)

`MannWhitneyU` Software

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




```
CMD
book>type mannwhitneyu-ex1.txt
2       4
3       6
2       5
4       4
9       8
2       3
5       4
3       2
4       4
3       5

book>java MannWhitneyU mannwhitneyu-ex1.txt
U = 31.0
z = -1.436, p = 0.1509
z' = -1.469, p' = 0.1418

book>
```

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

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

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

Test statistic: Normalized *z* score

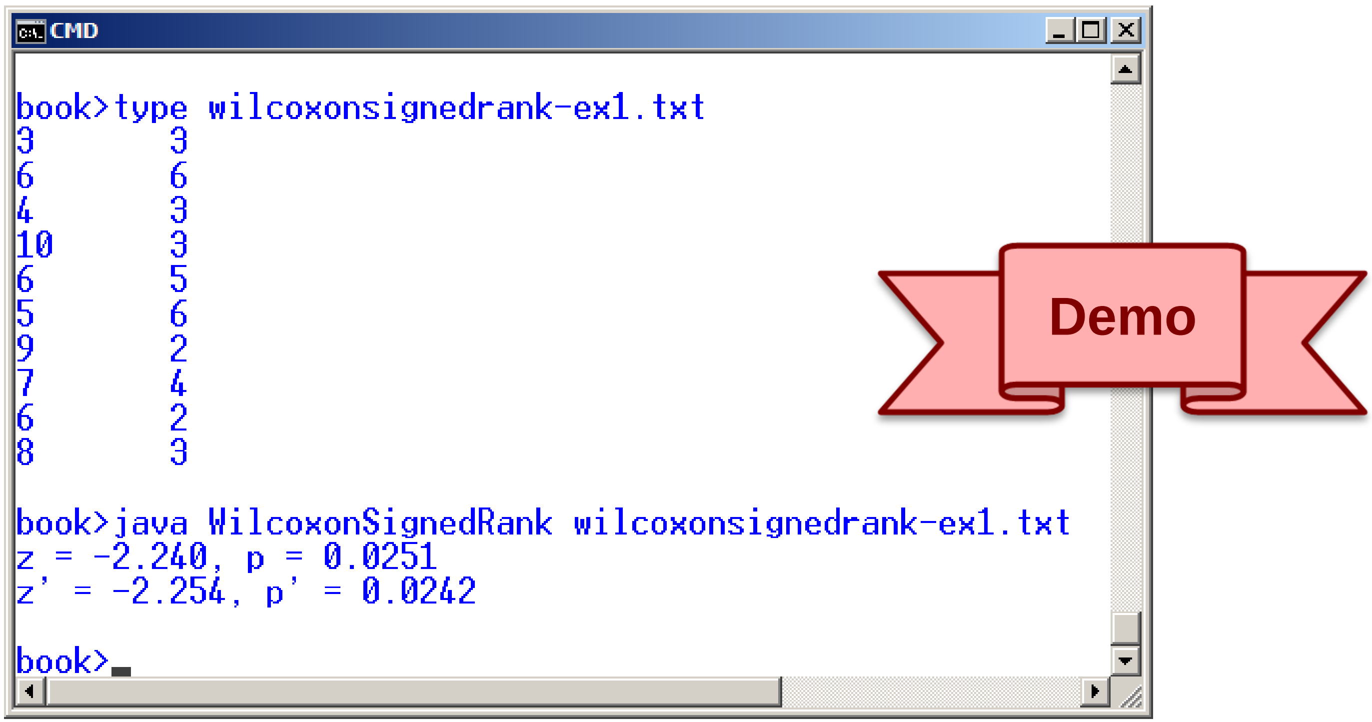
p (probability of the observed data, given the null hypothesis)

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

WilcoxonSignedRank Software

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



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

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

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

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. ($\chi^2 = 9.605$, $p < .01$).

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

`KruskalWallis` Software

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

```
book>type kruskalwallis-ex1.txt
9       7       4
9       3       5
4       5       5
9       3       2
6       2       2
3       1       1
8       4       2
9       7       2

book>java KruskalWallis kruskalwallis-ex1.txt
H = 9.421, p = 0.0090
H' = 9.605, p' = 0.0082

book>
```

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

Post Hoc Comparisons

- As with the analysis of variance, a significant result only indicates that at least one condition differs significantly from one other condition
- To determine which pairs of conditions differ significantly, a post hoc comparisons test is used
- Available using `-ph` option (see below)

CMD

```
book>java KruskalWallis kruskalwallis-ex1.txt -ph
H = 9.421, p = 0.0090
H' = 9.605, p' = 0.0082

----------------------------------------------------
----- Multiple Comparisons Test (alpha = .05) ------
----------------------------------------------------
Pair 1:2 -->  7.4375 >=  7.6103  ?   -
Pair 1:3 --> 10.5625 >=  7.6103  ?   * (significant)
Pair 2:3 -->  3.1250 >=  7.6103  ?   -
------------------------------------------------------

book>
```

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

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

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

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

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]

```
CMD

book>type friedman-ex1.txt
66      80      67      73
79      64      61      66
67      58      61      67
71      73      54      75
72      66      59      78
68      67      57      69
71      68      59      64
74      69      69      66

book>java Friedman friedman-ex1.txt
H(3) = 8.475, p = 0.0372
H'(3) = 8.692, p' = 0.0337

book>
```

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

Post Hoc Comparisons

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

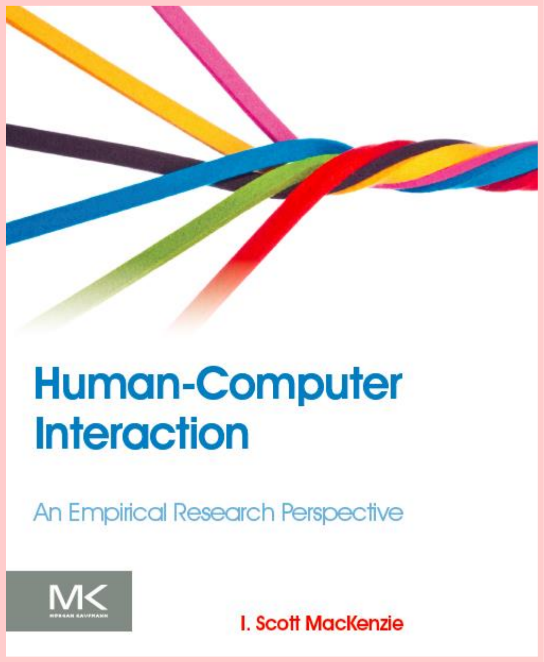
```
book>java Friedman friedman-ex1.txt -ph
H(3) = 8.475, p = 0.0372
H'(3) = 8.692, p' = 0.0337

-----------------------------------------------------------------
------------ Pairwise Comparisons (using Conover's F) ------------
-----------------------------------------------------------------
Pair 1:2 --> abs( 3.063 -  2.438)  >  1.132  ?     -
Pair 1:3 --> abs( 3.063 -  1.438)  >  1.132  ?     * (significant)
Pair 1:4 --> abs( 3.063 -  3.063)  >  1.132  ?     -
Pair 2:3 --> abs( 2.438 -  1.438)  >  1.132  ?     -
Pair 2:4 --> abs( 2.438 -  3.063)  >  1.132  ?     -
Pair 3:4 --> abs( 1.438 -  3.063)  >  1.132  ?     * (significant)
-----------------------------------------------------------------

book>
```

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