Leibniz Universität Hannover

Mensch-Computer-Interaktion 2

Data Analysis

Human-Computer Interaction Group

Prof. Dr. Michael Rohs
michael.rohs@hci.uni-hannover.de

Lectures

Session	Date	Topic	
1	6.4.	Introduction	
2	13.4.	Interaction elements	GUI toolkits, interaction techniques
3	20.4.	Event handling	
4	27.4.	Scene graphs	
5	4.5.	Interaction techniques	
	11.5.	no class (CHI)	
	18.5.	no class (spring break)	
6	25.5.	Experiments	design and analysis of experiments
7	1.6.	Data Analysis	
8	8.6.	Data Analysis	
9	15.6.	Visualization	
10	22.6.	Visualization	current topics beyond-desktop UIs
11	29.6.	Modeling interaction	
12	6.7.	Computer vision for interaction	
13	13.7.	Computer vision for interaction	

Klausur:
28.7.2016
8-11 Uhr
HG E214

Review

- What is a null hypothesis?
- What is a significance level (α)?
- What is a p-value?
- What is a randomization test?
- Basic idea of ANOVA?
- What is a type I error (false positive)?
- What is a type II error (false negative)?
- What is an effect size?
- What should be reported as the result of an experiment?

Between-Subjects Designs

- Research question:

  Do left-handed users and right-handed users differ in the time to complete an interaction task?

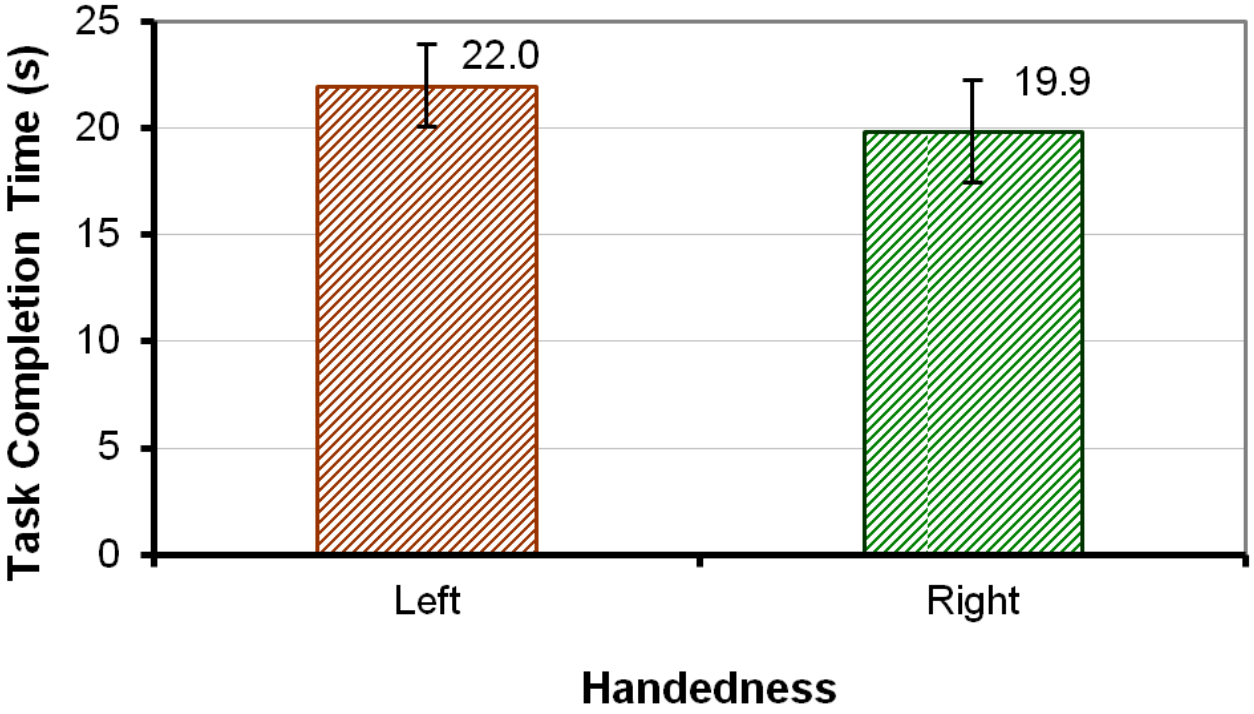
- The independent variable (handedness) must be assigned between-subjects
- Example data set →

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	

MacKenzie: Human-Computer Interaction - An Empirical Research Perspective.

Summary Data and Chart

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



MacKenzie: Human-Computer Interaction - An Empirical Research Perspective.

ANOVA (Single Factor, Between Subjects)

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				

(permutation test yielded p = 0.0939)

- Single-factor, between-subjects design
- The difference was not statistically significant ($F_{1,14} = 3.78$, $p > .05$)
- Degrees of freedom (k=2 groups, n=8 observations per group)
  - Condition, effect → (k – 1)
  - Error, residual → k (n - 1)

MacKenzie: Human-Computer Interaction - An Empirical Research Perspective.

Two-Way Repeated Measures ANOVA

- Experiment with two independent variables (a "two-way design")
- ANOVA tests for
  - Two main effects + one interaction effect
- Interaction effect: Level of IV_i may influence effect of IV_j $(i \neq j)$
- 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
  - Both IVs assigned within-subjects
  - Participants: 12

MacKenzie: Human-Computer Interaction - An Empirical Research Perspective.

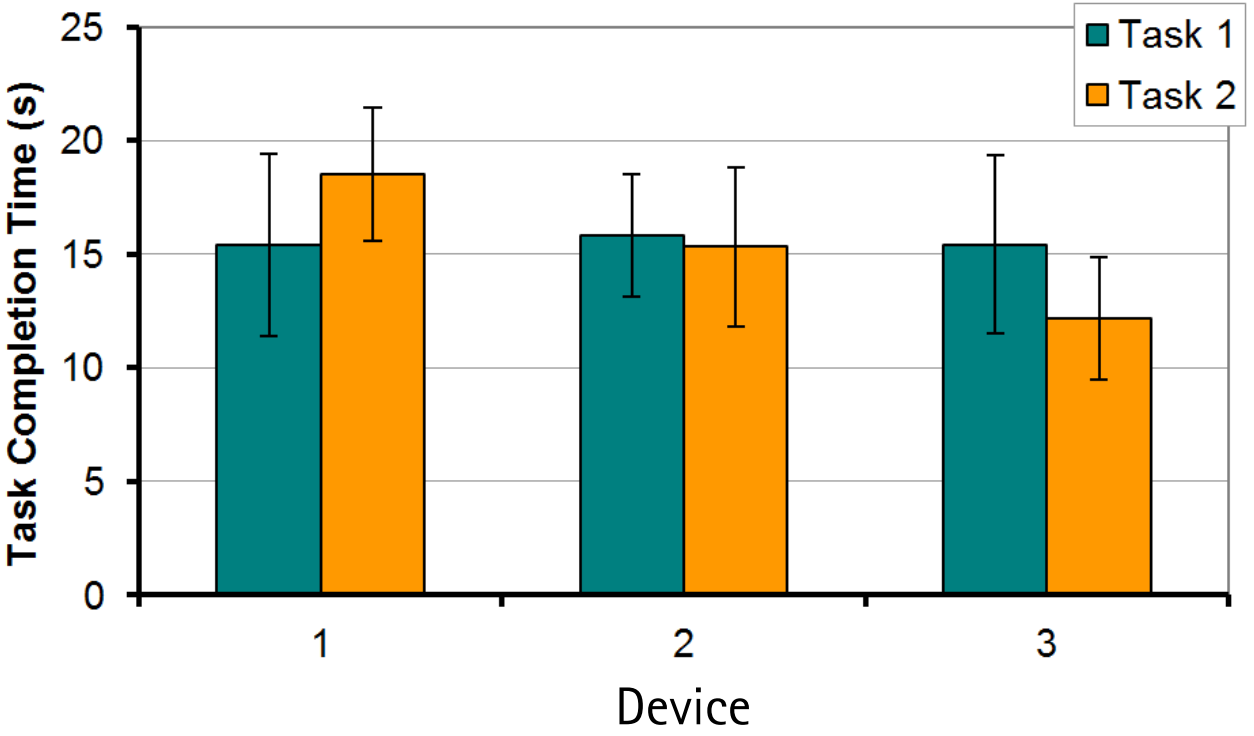
Data Set (Two-Way Repeated Measures ANOVA)

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

MacKenzie: Human-Computer Interaction - An Empirical Research Perspective.

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



MacKenzie: Human-Computer Interaction - An Empirical Research Perspective.

Results (Two-Way Repeated Measures 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				

main effect

interaction effect

How to report these results?

MacKenzie: Human-Computer Interaction - An Empirical Research Perspective.

Reporting (Two-Way Repeated Measures ANOVA)

The grand mean for task completion time was 15.4 s. Device 1 was the slowest at 17.0 s. At 15.6 s Device 2 was 8% faster and at 13.8 s Device 3 was 19% faster than device 1. The main effect of device on task completion time was statistically significant ($F_{2,22} = 5.865$, $p < .01$). The effect of task was modest. Task completion time was 15.6 s for task 1. Task 2 was slightly faster at 15.3 s, but the difference was not statistically significant ($F_{1,11} = 0.076$, ns). The results by device and task are shown in Table X and Figure Y. There was a significant Device x Task interaction effect ($F_{2,22} = 5.435$, $p < .05$).

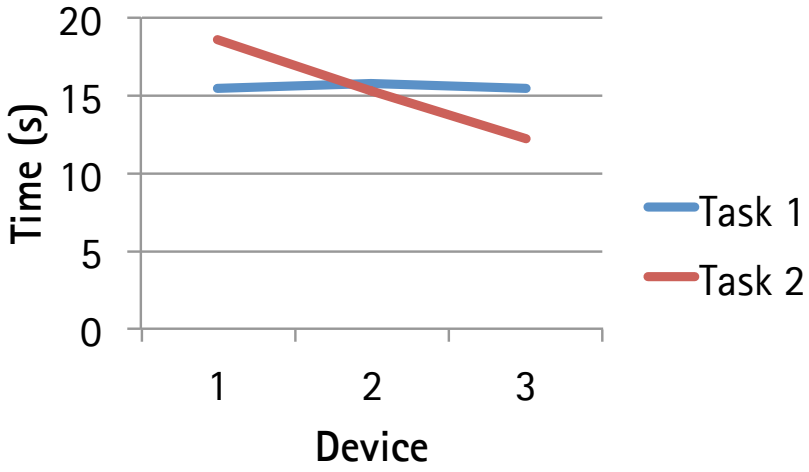
	Task 1	Task 2	*Mean*
Device 1	15.4	18.5	17.0
Device 2	15.8	15.3	15.6
Device 3	15.4	12.2	13.8
Mean	15.6	15.3	15.4

grand mean

MacKenzie: Human-Computer Interaction - An Empirical Research Perspective.

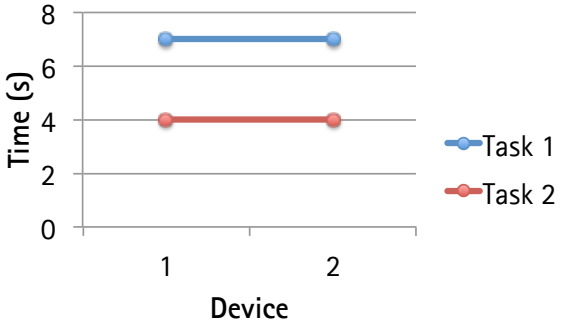
Interaction Effect

- The effects of factor 1 (on the dependent variable) differ, depending on factor 2
  - The levels of one factor affect the levels of another factor in different ways
- Example: Interaction between Device and Task
  - Task 1 is faster with device 1, but slower with device 3

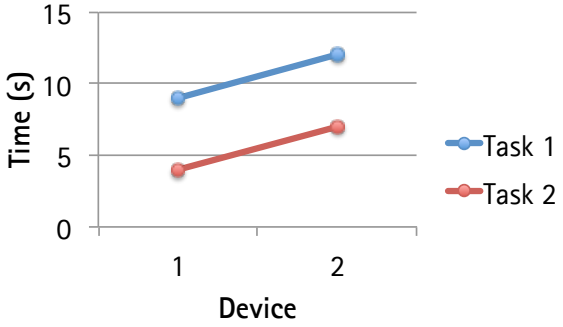


Main Effects and Interaction Effect

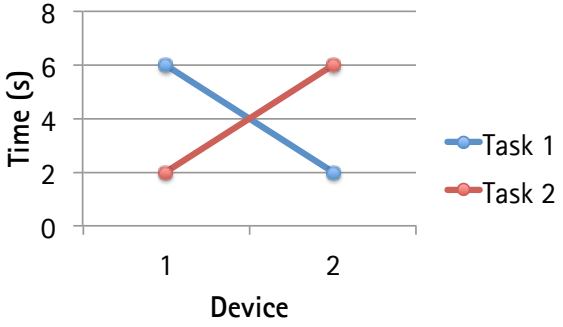
main effect of task: yes
main effect of device: no
interaction effect: no



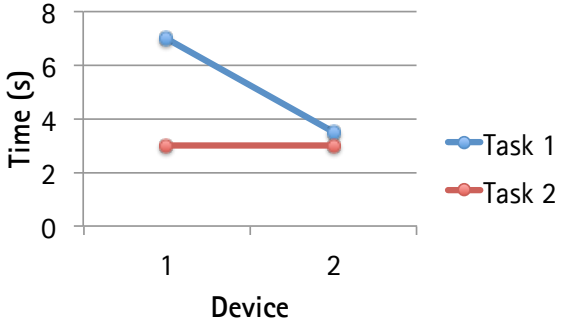
main effect of task: yes
main effect of device: yes
interaction effect: no



main effect of task: no
main effect of device: no
interaction effect: yes



main effect of task: yes
main effect of device: yes
interaction effect: yes



Partitioning Sums of Squares for Two-Way Repeated Measures ANOVA

(optional slide)

- Independent variables: Device (d devices), Task (t tasks)
- Sums of squares
  - $SS_{total} = SS_{device} + SS_{task} + SS_{participant} +$ main effects
    $SS_{device*task} + SS_{device*participant} + SS_{task*participant} +$ interaction effects
    SS_{error}
- How to compute these? $ss_{total} = \sum_{i=1}^{p}\sum_{j=1}^{d}\sum_{k=1}^{t}\left(y_{i,j,k} - \bar{y}\right)^2$

$$ss_p = \sum_{i=1}^{p}\sum_{j=1}^{d}\sum_{k=1}^{t}\left(y_{i,j,k} - \bar{y}_{i,\_,\_}\right)^2 \qquad ss_d = \sum_{i=1}^{p}\sum_{j=1}^{d}\sum_{k=1}^{t}\left(y_{i,j,k} - \bar{y}_{\_,j,\_}\right)^2 \qquad ss_t = \sum_{i=1}^{p}\sum_{j=1}^{d}\sum_{k=1}^{t}\left(y_{i,j,k} - \bar{y}_{\_,\_,k}\right)^2$$

$$ss_{p,d} = \sum_{i=1}^{p}\sum_{j=1}^{d}\sum_{k=1}^{t}\left(y_{i,j,k} - \bar{y}_{i,j,\_}\right)^2 \qquad ss_{p,t} = \sum_{i=1}^{p}\sum_{j=1}^{d}\sum_{k=1}^{t}\left(y_{i,j,k} - \bar{y}_{i,\_,k}\right)^2 \qquad ss_{d,t} = \sum_{i=1}^{p}\sum_{j=1}^{d}\sum_{k=1}^{t}\left(y_{i,j,k} - \bar{y}_{\_,j,k}\right)^2$$

Partitioning Sums of Squares for Two-Way Repeated Measures ANOVA

(optional slide)

$SS_{total} = SS_{device} + SS_{task} + SS_{participant} + SS_{device*task} + SS_{device*participant} + SS_{task*participant} + SS_{error}$

$SS_{p} = SS_{device} + SS_{task} + SS_{device*task} + SS_{device*participant} + SS_{task*participant} + SS_{error}$

$SS_{participant} = SS_{total} - SS_{p}$

$SS_{device} = SS_{total} - SS_{d}$

$SS_{task} = SS_{total} - SS_{t}$

$SS_{p,d} = SS_{task} + SS_{device*task} + SS_{task*participant} + SS_{error}$

$SS_{device*participant} = SS_{total} - SS_{p,d} - SS_{participant} - SS_{device}$

$SS_{p,t} = SS_{device} + SS_{device*task} + SS_{device*participant} + SS_{error}$

$SS_{task*participant} = SS_{total} - SS_{p,t} - SS_{participant} - SS_{task}$

$SS_{d,t} = SS_{participant} + SS_{device*task} + SS_{device*participant} + SS_{error}$

$SS_{device*task} = SS_{total} - SS_{d,t} - SS_{task} - SS_{device}$

Partitioning Sums of Squares for Two-Way Repeated Measures ANOVA

(optional slide)

SS_{total} =	977.778	df = p*d*t-1 = 71	p=12, d=3, t=2
SS_{p} =	843.000		
SS_{d} =	856.750		
SS_{t} =	976.889		
$SS_{participant}$ =	134.778	df = p-1 = 11	
SS_{device} =	121.028	df = d-1 = 2	
SS_{task} =	0.889	df = t-1 = 1	
$SS_{p,d}$ =	495.000		
$SS_{device*participant}$ =	226.972	df = (d-1)*(p-1) = 2*11 = 22	
$SS_{p,t}$ =	714.000		
$SS_{task*participant}$ =	128.111	df = (t-1)*(p-1) = 1*11 = 11	
$SS_{d,t}$ =	734.833		
$SS_{device*task}$ =	121.028	df = (d-1)*(t-1) = 2*1 = 2	
SS_{error} =	244.972	df = (d-1)*(t-1)*(p-1) = 11*2*1 = 22	

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				

Partitioning Sums of Squares for Two-Way Repeated Measures ANOVA

(optional slide)

ss_{device}	= 121.028	df_{device} = 2	ms_{device} = 60.514
ss_{task}	= 0.889	df_{task} = 1	ms_{task} = 0.889
$ss_{device*participant}$	= 226.972	$df_{device*participant}$ = 22	$ms_{device*participant}$ = 10.317
$ss_{task*participant}$	= 128.111	$df_{task*participant}$ = 11	$ms_{task*participant}$ = 11.646
$ss_{device*task}$	= 121.028	$df_{device*task}$ = 2	$ms_{device*task}$ = 60.514
ss_{error}	= 244.972	df_{error} = 22	ms_{error} = 11.135

$$F_{device} = ms_{device} / ms_{device*participant} = 5.865$$
$$F_{task} = ms_{task} / ms_{task*participant} = 0.076$$
$$F_{device*task} = ms_{device*task} / ms_{error} = 5.435$$

Anova2 Java Tool

- Analysis of variance Java tool: Anova2
  - http://www.yorku.ca/mack/HCIbook/Anova2.zip
- Operates from command line on data in a text file

```
$ java Anova2
Usage: java Anova2 file p f1 f2 f3 [-a] [-d] [-m] [-h]
    file = data file (comma or space delimited)
    p = # of rows 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)
```

MacKenzie: Human-Computer Interaction - An Empirical Research Perspective.

Dix et al. Example[1]

- Single-factor, within-subjects design

```
CMD

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>
```

```
CMD

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 between-subjects factor (order "NA", order "AN")
- Includes header lines

```
CMD

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>
```

```
CMD

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>
```

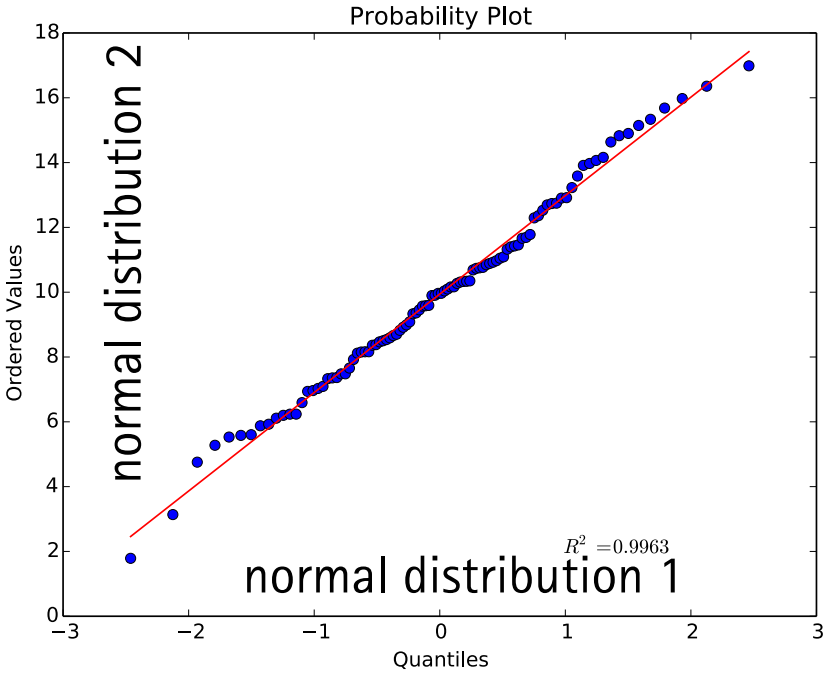
MacKenzie: Human-Computer Interaction - An Empirical Research Perspective.

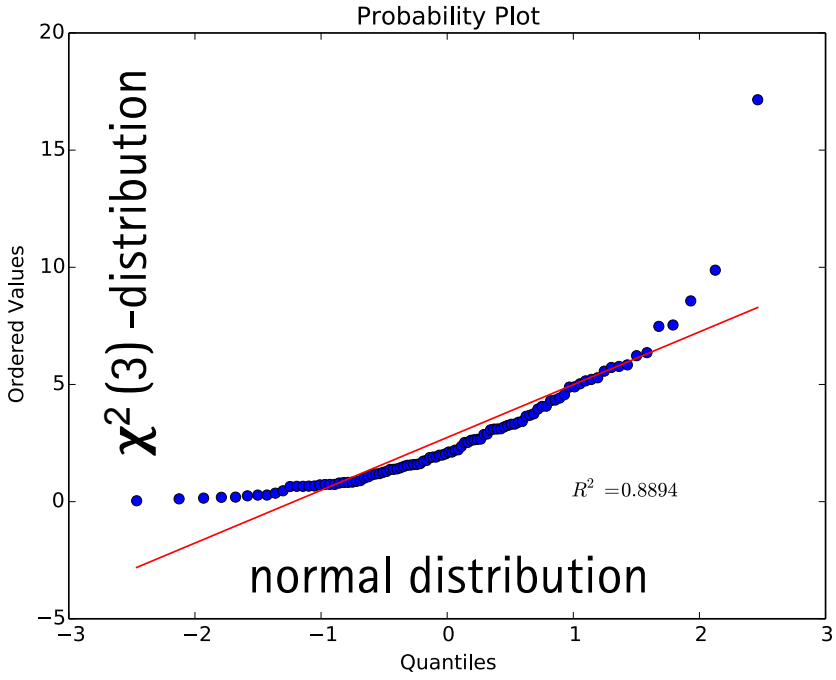
Parametric vs. Non-Parametric Tests

- Violations of normality
  - Normality tests often fail for small sample sizes
  - QQ-plot allows for visual inspection of normality
- Non-parametric tests have lower statistical power
  - Loss of information
  - Example: Resort from interval/ratio data to ranks (1st, 2nd, 3rd, etc.)
- Assumptions of Analysis of Variance (ANOVA)
  - DV is normally distributed
  - Equal variances in groups
  - Independent observations
- ANOVA is reasonably robust against violations of normality
- May run a parametric and a non-parametric test for validation

Quantile-Quantile-Plots (QQ-Plots)

- Visual assessment of distributions
- Plot the quantiles of two distributions against each other
- Similar distributions will produce a straight line





Leibniz Universität Hannover

CHI-SQUARE TEST

Chi-Square Test (for Nominal Data)

- Dependent variable is nominal / categorical
- Data is organized in a contingency table
  - Table of the number of observations in each category
- A chi-square test compares observed against expected counts
- Example: Preferred device: mouse, touch, stylus
  - Does device preference depend on age?
- χ^2-statistic measures deviation of observation from expectation:

$$\chi^2 = \sum_{i=1}^{k} \frac{(observed_i - expected_i)^2}{expected_i}$$

k observations

  - Follows χ^2-distribution if null hypothesis is true

Chi-Square Test – Example #1

Preferred device

		Mouse	Touch	Stylus	row sums	probabilities
Age	< 18	4	9	7	20	20/60=1/3
	18-65	12	6	2	20	20/60=1/3
	≥65	4	15	1	20	20/60=1/3
column sums		20	30	10	N=60	
probabilities		20/60=1/3	30/60=1/2	10/60=1/6		

If statistical independence:
P(<18 ∧ Mouse) = P(<18) * P(Mouse)

Expected values:

Preferred device

		Mouse	Touch	Stylus
Age	< 18	60*1/3*1/3=6.7	60*1/3*1/2=10	60*1/3*1/6=3.3
	18-65	60*1/3*1/3=6.7	60*1/3*1/2=10	60*1/3*1/6=3.3
	≥65	60*1/3*1/3=6.7	60*1/3*1/2=10	60*1/3*1/6=3.3

N * pRow * pCol

Chi-Square Test – Example #1

Observed

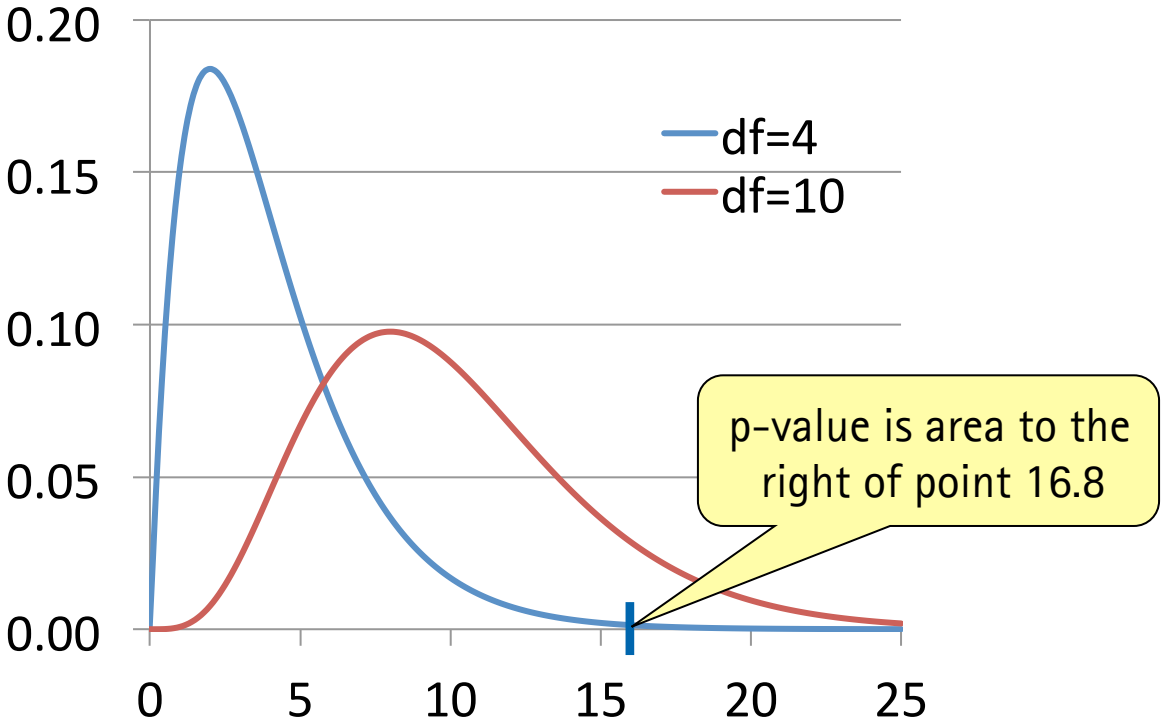
		Preferred device		
		Mouse	Touch	Stylus
Age	< 18	4	9	7
	18-65	12	6	2
	≥65	4	15	1

Expected

		Preferred device		
		Mouse	Touch	Stylus
Age	< 18	6.7	10.0	3.3
	18-65	6.7	10.0	3.3
	≥65	6.7	10.0	3.3

- χ^2-statistic: $\chi^2 = \sum_{i=1}^{k} \frac{(observed_i - \exp ected_i)^2}{\exp ected_i} = 16.8$
- Degrees of freedom = (rows-1) * (columns-1) = (3-1) * (3-1) = 4

Chi-Square Distribution



$p = P(\chi^2 \geq 16.8)$

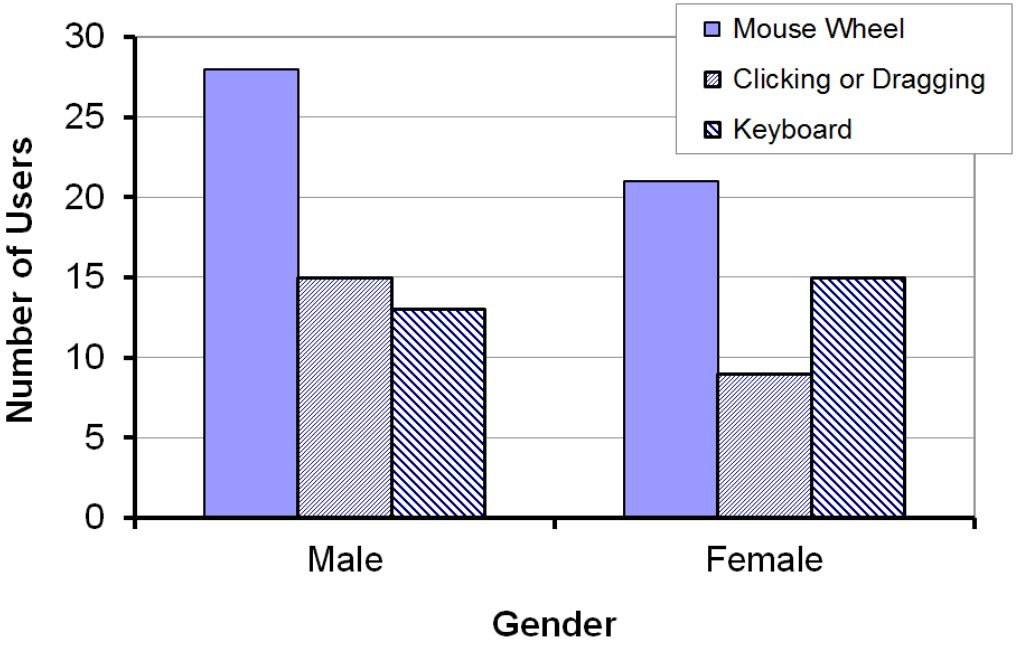
- $\chi^2(4) = 16.8, p < 0.005$
- Degrees of freedom = (rows-1) * (columns-1) = (3-1) * (3-1) = 4

Chi-Square – Example #2

Do males and females differ in their method of scrolling on desktop systems?

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



MacKenzie: Human-Computer Interaction - An Empirical Research Perspective.

Chi-Square – Example #2

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**

$\chi^2 = 1.462$, significant if exceeds critical value (next slide)

MacKenzie: Human-Computer Interaction - An Empirical Research Perspective.

Chi-Square Critical Values

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

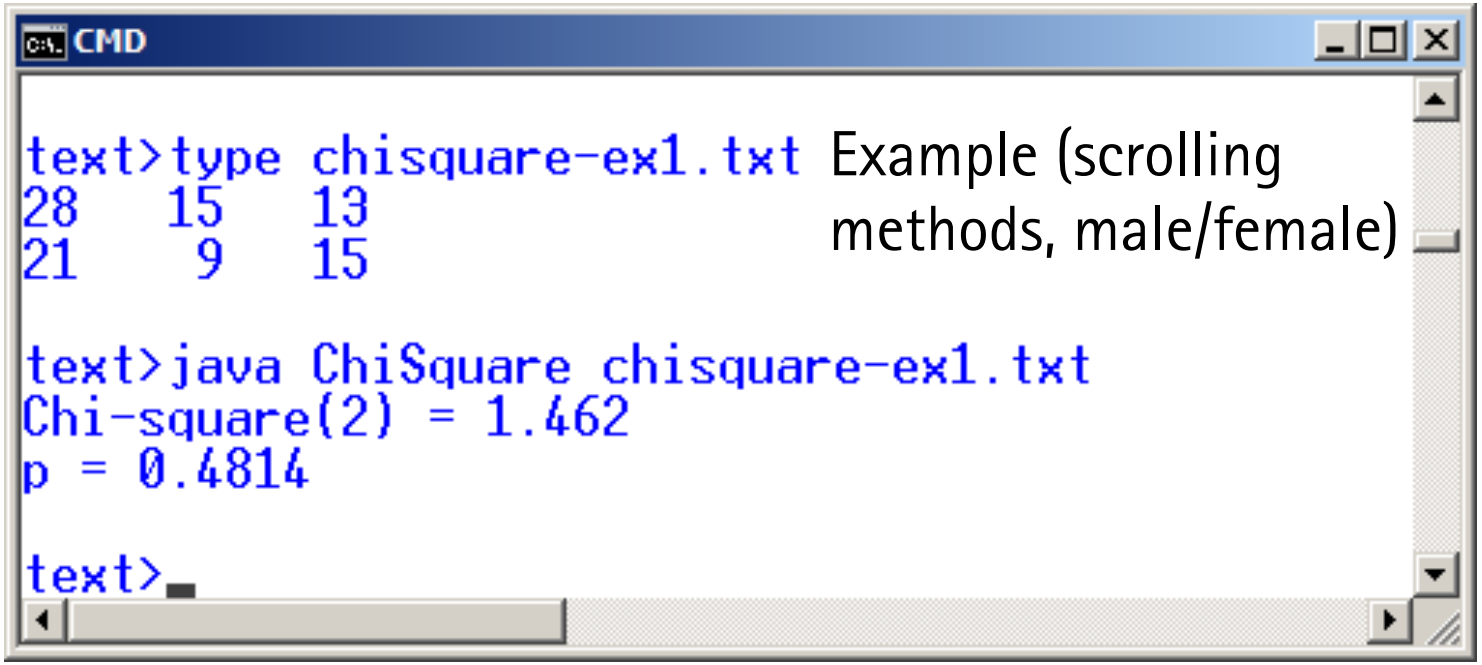
Significance Threshold (α)	Degrees of Freedom							
	1	2	3	4	5	6	7	8
.1	2.71	4.61	6.25	7.78	9.24	10.65	12.02	13.36
.05	3.84	5.99	7.82	9.49	11.07	12.59	14.07	15.51
.01	6.64	9.21	11.35	13.28	15.09	16.81	18.48	20.09
.001	10.83	13.82	16.27	18.47	20.52	22.46	24.32	26.13

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

MacKenzie: Human-Computer Interaction - An Empirical Research Perspective.

ChiSquare Java Tool

http://www.yorku.ca/mack/HCIbook/ChiSquare.zip



- Note: calculates p (assuming $\alpha = .05$)
- ph option for post hoc multiple comparisons across columns

MacKenzie: Human-Computer Interaction - An Empirical Research Perspective.

Chi-Square – Example #3

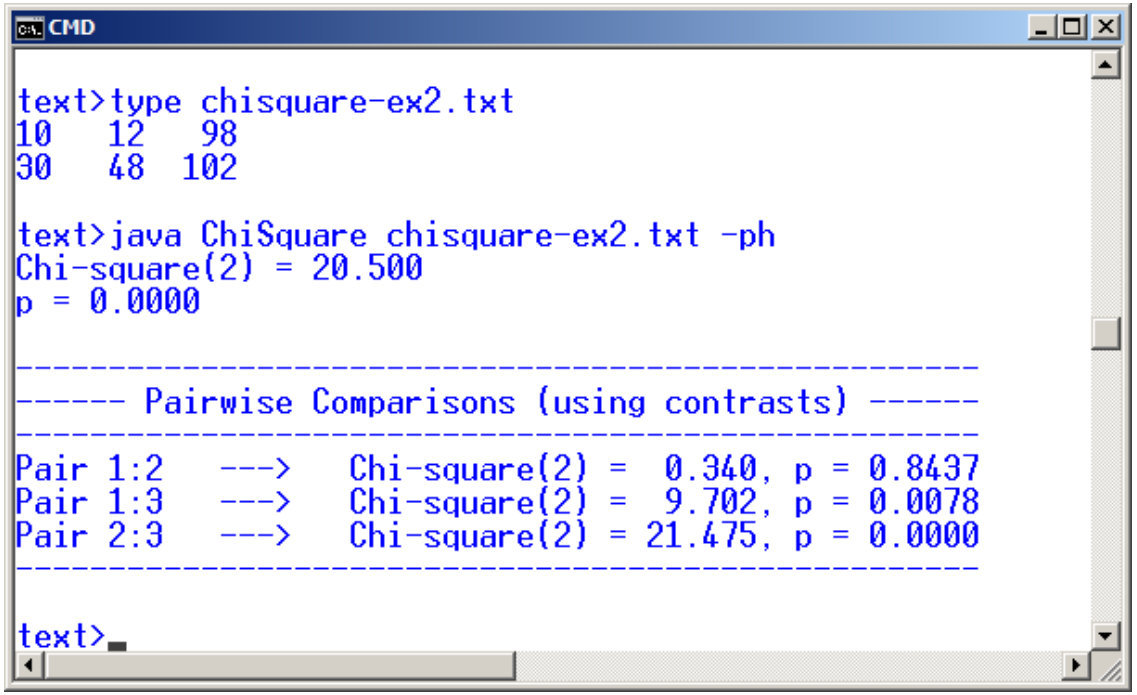
- Research question
  - Do students, professors, and parents differ in their responses to the question: Students should be allowed to use mobile phones during classroom lectures?
- Data

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

MacKenzie: Human-Computer Interaction - An Empirical Research Perspective.

Chi-Square – Example #3

- Result: Significant difference in responses ($\chi 2 = 20.5$, $p < .0001$)
- Post hoc comparisons reveal that opinions differ between students:parents and professors:parents



1 = students
2 = professors
3 = parents

MacKenzie: Human-Computer Interaction - An Empirical Research Perspective.

NON-PARAMETRIC TESTS (FOR ORDINAL DATA)

Non-Parametric Tests for Ordinal Data

- Non-parametric tests used most commonly on ordinal data (ranks)
- 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

MacKenzie: Human-Computer Interaction - An Empirical Research Perspective.

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)

MacKenzie: Human-Computer Interaction - An Empirical Research Perspective.

Data (Example #1)

- Means
  - 3.7 (Mac users)
  - 4.5 (PC users)
- Difference statistically significant?
- Data are ordinal (at least), a non-parametric test is used
- Which test?

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**

MacKenzie: Human-Computer Interaction - An Empirical Research Perspective.

Mann Whitney U Test

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

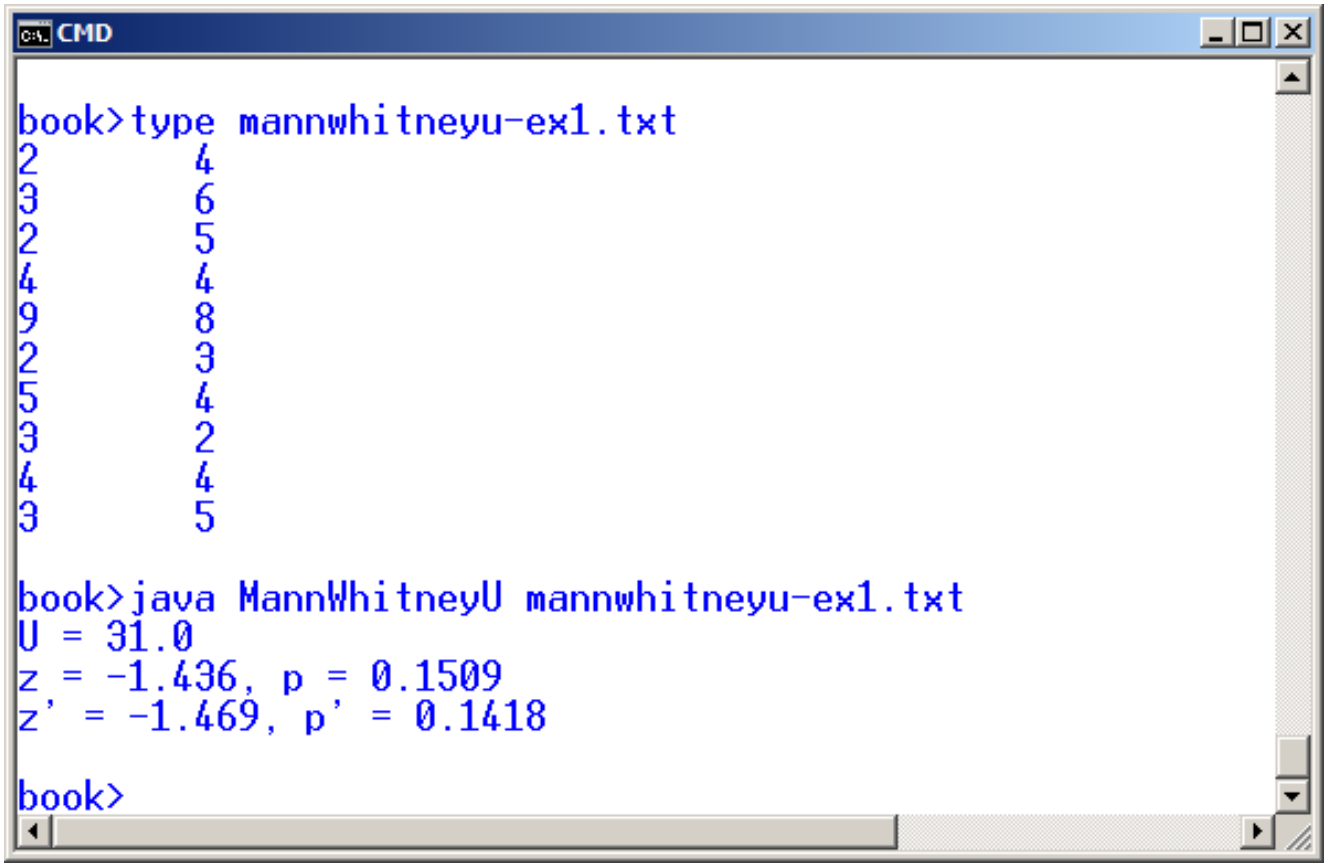
	Count	Sum Ranks	Mean Rank
MAC	10	86.000	8.600
PC	10	124.000	12.400

Conclusion:
The null hypothesis remains tenable: Did not find a difference in the political leaning of *Mac* users and *PC* users ($U = 31.0$, $p > .05$).

MacKenzie: Human-Computer Interaction - An Empirical Research Perspective.

MannWhitneyU Java Tool

http://www.yorku.ca/mack/HCIbook/MannWhitneyU.zip



MacKenzie: Human-Computer Interaction - An Empirical Research Perspective.

Non-Parametric – Example #2

- Research question
  - Do two new designs for media players differ in "cool appeal" for young users?
- Method
  - 10 young 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)

MacKenzie: Human-Computer Interaction - An Empirical Research Perspective.

Data (Example #2)

- Means
  - 6.4 (media player A)
  - 3.7 (media player B)
- 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

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**

MacKenzie: Human-Computer Interaction - An Empirical Research Perspective.

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

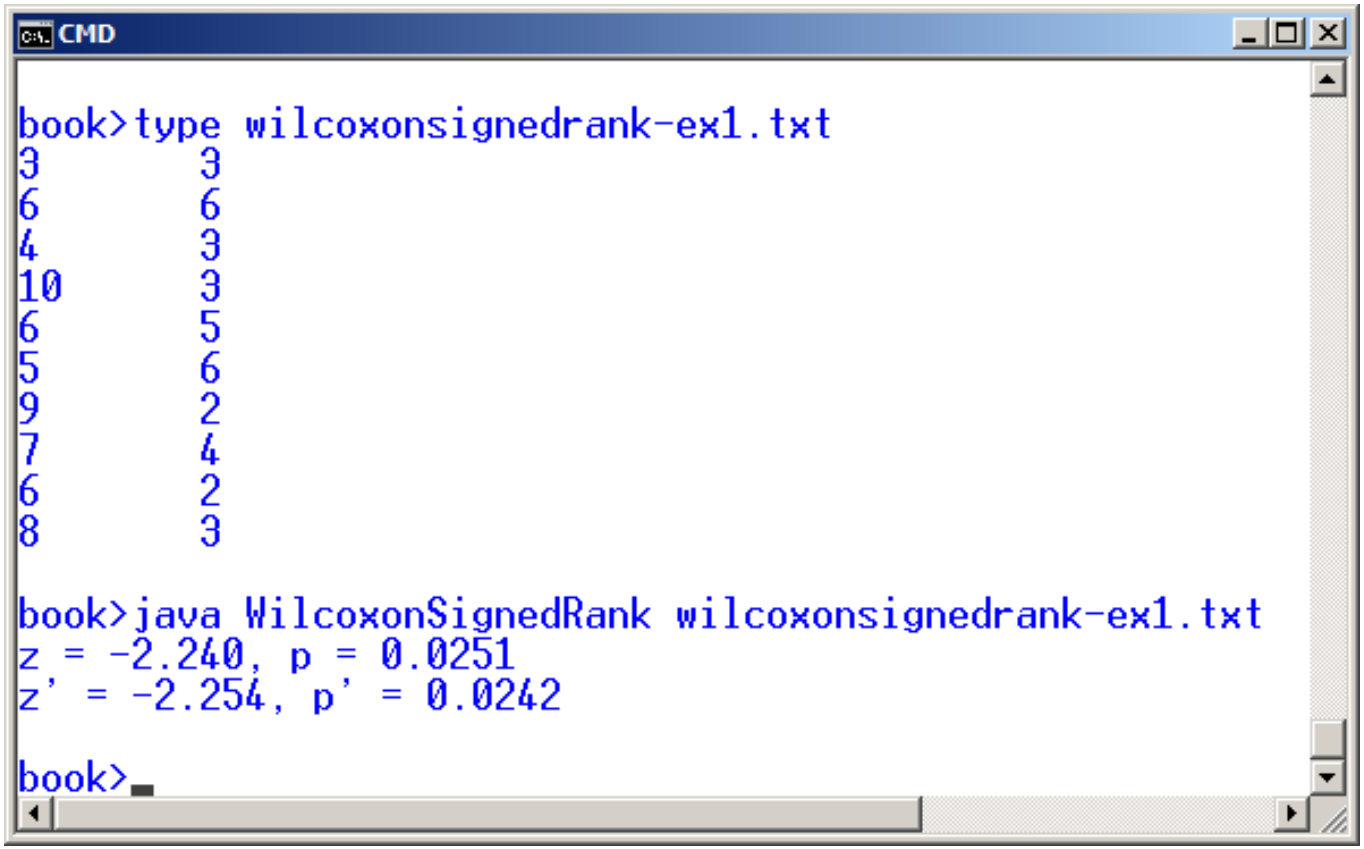
	Count	Sum Ranks	Mean Rank
# Ranks < 0	1	2.000	2.000
# Ranks > 0	7	34.000	4.857

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

MacKenzie: Human-Computer Interaction - An Empirical Research Perspective.

WilcoxonSignedRank Java Tool

http://www.yorku.ca/mack/HCIbook/WilcoxonSignedRank.zip



MacKenzie: Human-Computer Interaction - An Empirical Research Perspective.

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 tried 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)

MacKenzie: Human-Computer Interaction - An Empirical Research Perspective.

Data (Example #3)

- Means
  - 7.1 (20-29)
  - 4.0 (30-39)
  - 2.9 (40-49)
- Differences statistically significant?
- Data are ordinal (at least), a non-parametric is used
- Which test?

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

MacKenzie: Human-Computer Interaction - An Empirical Research Perspective.

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)

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

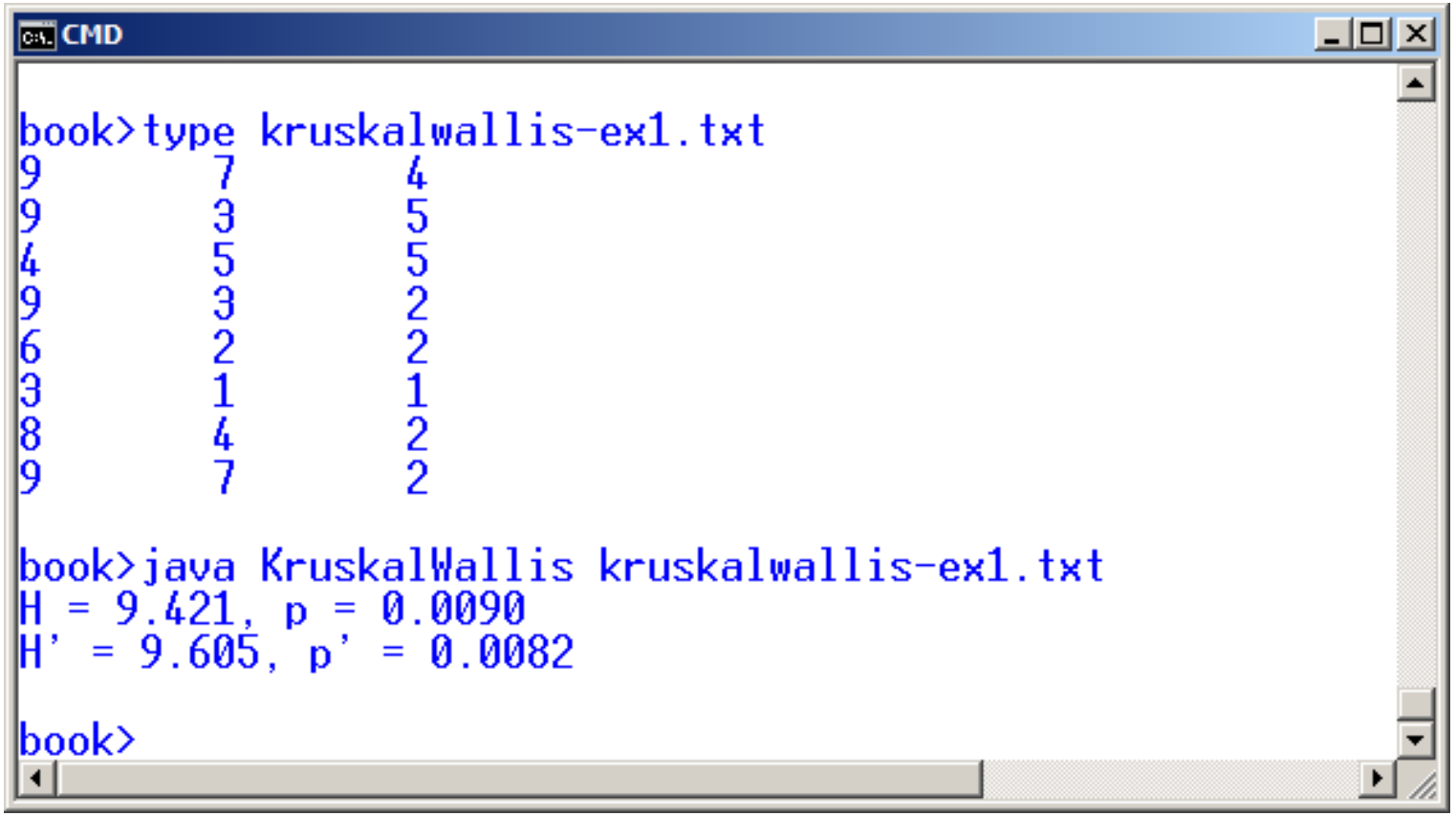
	Count	Sum Ranks	Mean Rank
A	8	148.000	18.500
B	8	88.500	11.063
C	8	63.500	7.938

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

MacKenzie: Human-Computer Interaction - An Empirical Research Perspective.

KruskalWallis Java Tool

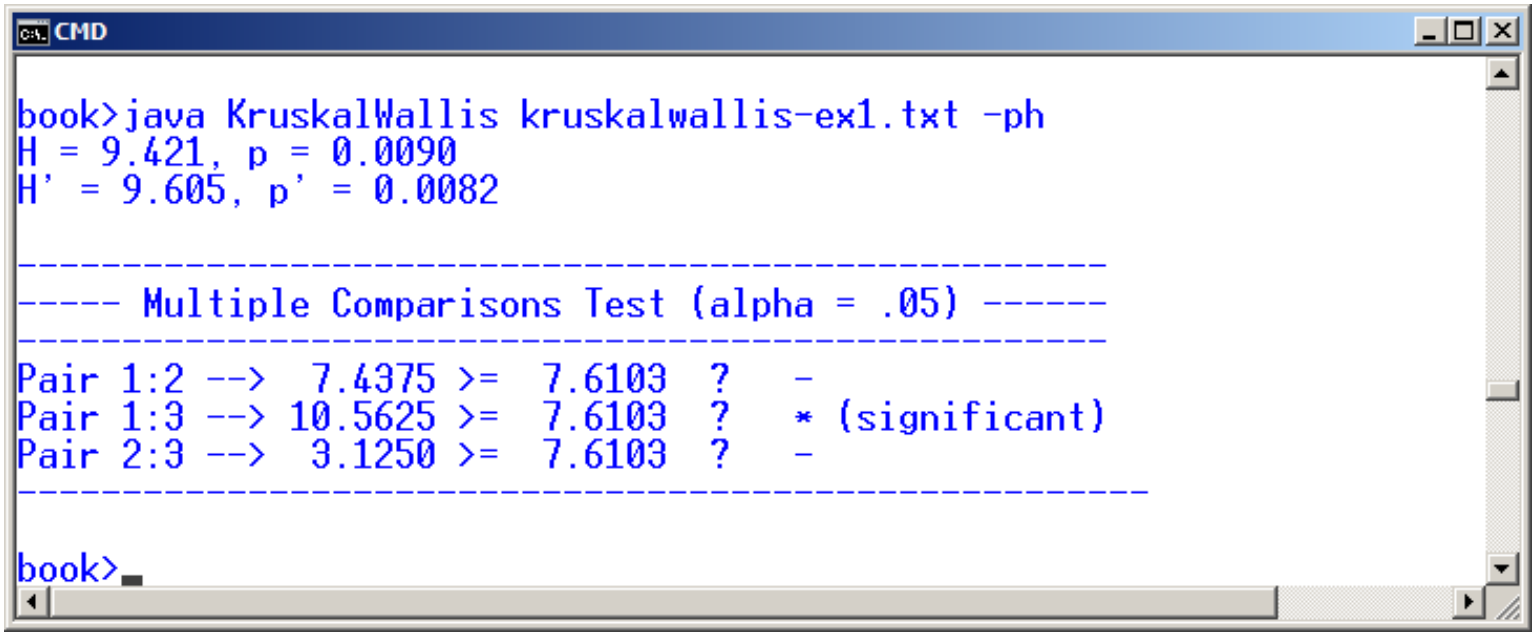
http://www.yorku.ca/mack/HCIbook/KruskalWallis.zip



MacKenzie: Human-Computer Interaction - An Empirical Research Perspective.

KruskalWallis Tool: Post Hoc Comparisons

- As with ANOVA, a significant result only indicates that at least one condition differs significantly from one other condition
  - Post hoc comparison test determines which pairs of conditions differ
- Available using –ph option



MacKenzie: Human-Computer Interaction - An Empirical Research Perspective.

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 tried 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)

MacKenzie: Human-Computer Interaction - An Empirical Research Perspective.

Data (Example #4)

- Means
  - 71.0 (A), 68.1 (B), 60.9 (C), 69.8 (D)
- Differences statistically significant?
- Data are ordinal (at least), non-parametric test is used
- Which test?

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

MacKenzie: Human-Computer Interaction - An Empirical Research Perspective.

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

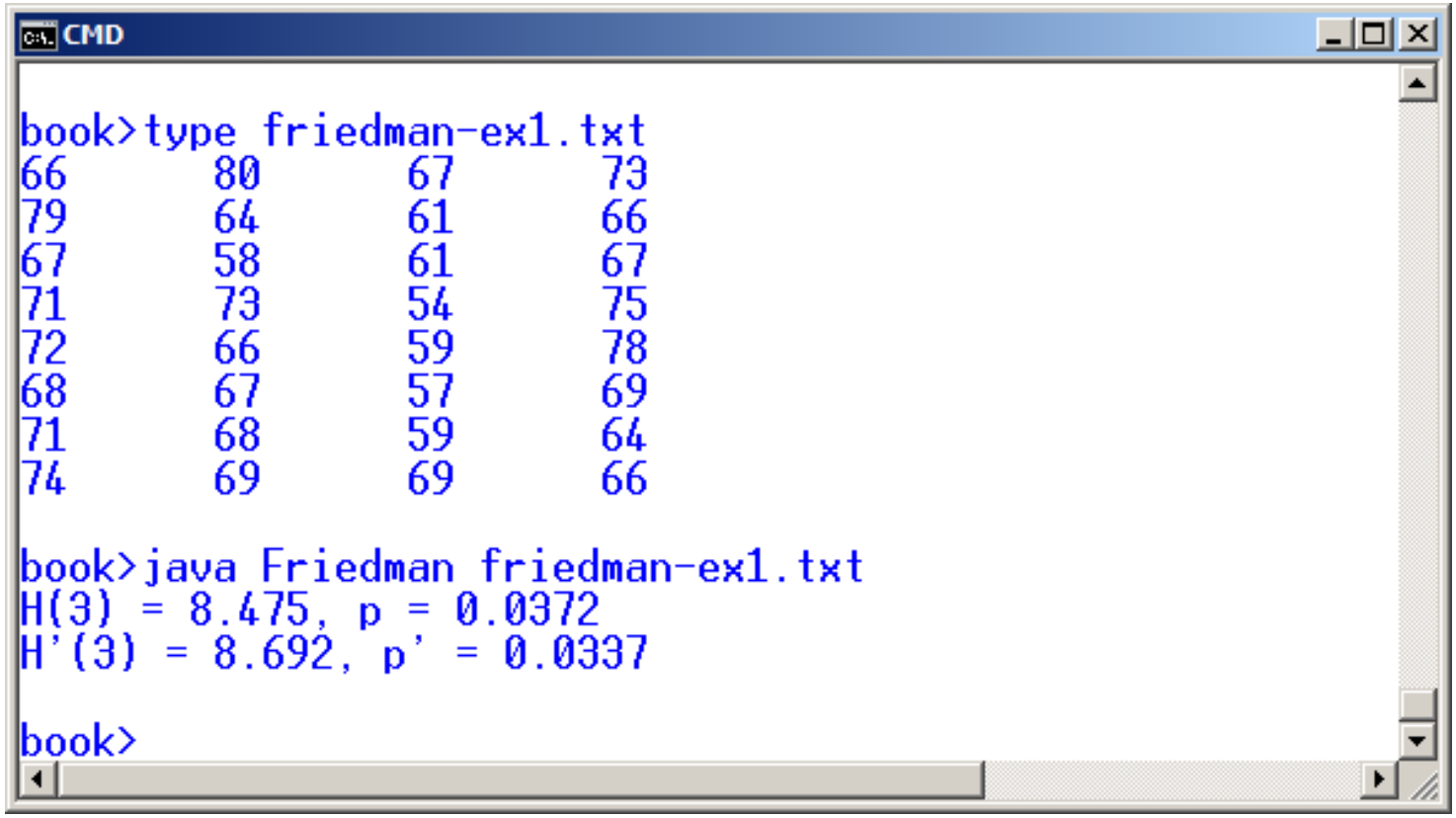
	Count	Sum Ranks	Mean Rank
A	8	24.500	3.063
B	8	19.500	2.438
C	8	11.500	1.438
D	8	24.500	3.063

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

MacKenzie: Human-Computer Interaction - An Empirical Research Perspective.

Friedman Java Tool

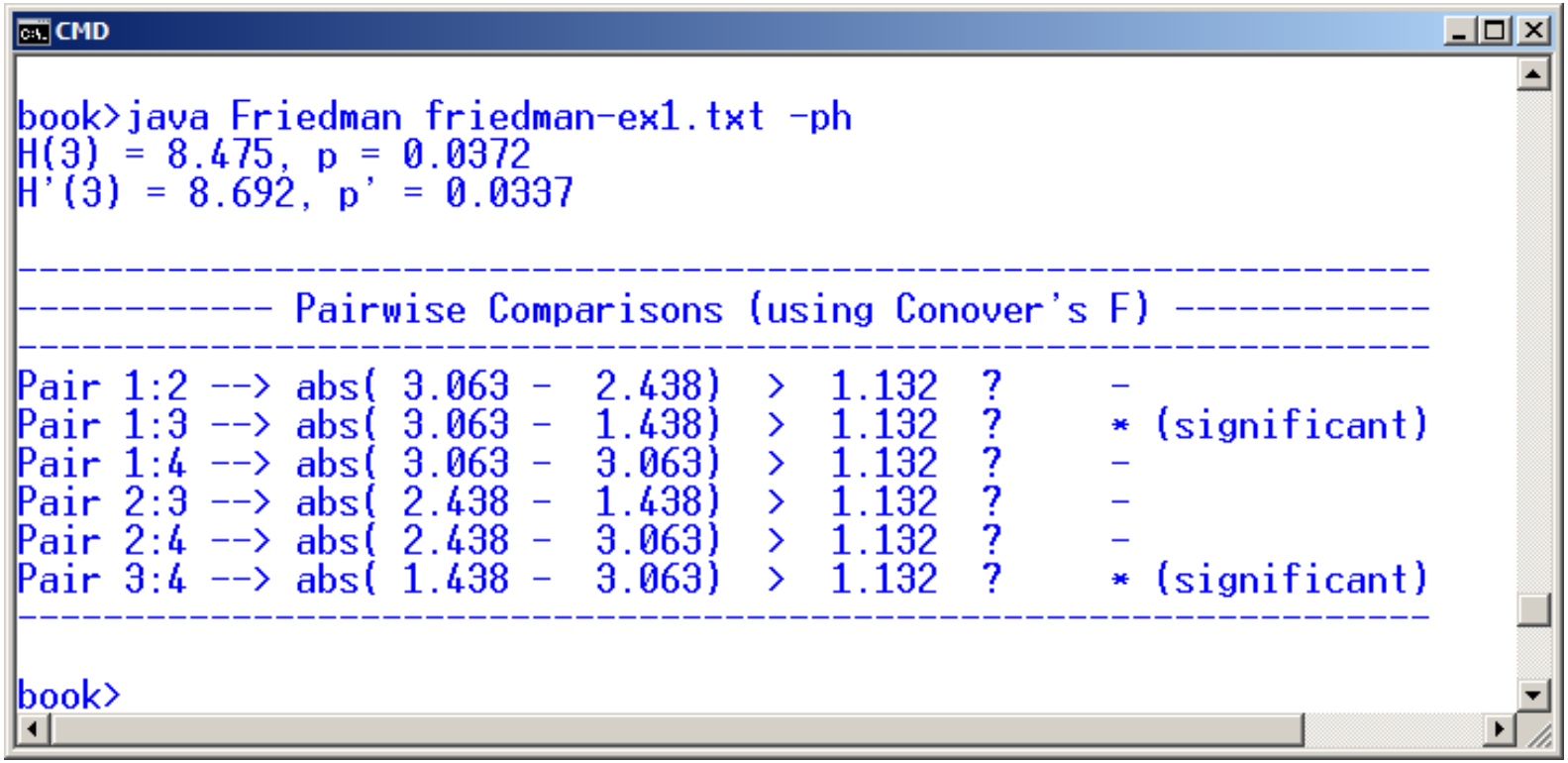
http://www.yorku.ca/mack/HCIbook/Friedman.zip



MacKenzie: Human-Computer Interaction - An Empirical Research Perspective.

Post Hoc Comparisons

- As with KruskalWallis, available using the –ph option…



MacKenzie: Human-Computer Interaction - An Empirical Research Perspective.

Leibniz Universität Hannover

STANDARD ERROR AND CONFIDENCE INTERVALS

Population Parameters and Sample Statistics

- Population mean and standard deviation
  - μ = population mean
  - σ = population standard deviation
  - Typically unknown, but can be estimated
  - Example: Parameters of normal distribution $N\left(\mu,\sigma^2\right)$
- Sample mean and standard deviation
  - $\overline{x}$ = sample mean
  - s = sample standard deviation
  - Used for estimating population parameters
  - Example: t-statistic $t = \frac{\overline{x} - \mu}{s / \sqrt{n}}$

Standard Error = Standard Deviation of the Mean

- Mean of sample size n
  - Take sample of n values, compute the mean
- Standard error of the mean is the standard deviation of the mean of samples of size n
  - Repeat the process of computing means of sample size n
  - Compute the standard deviation of these means

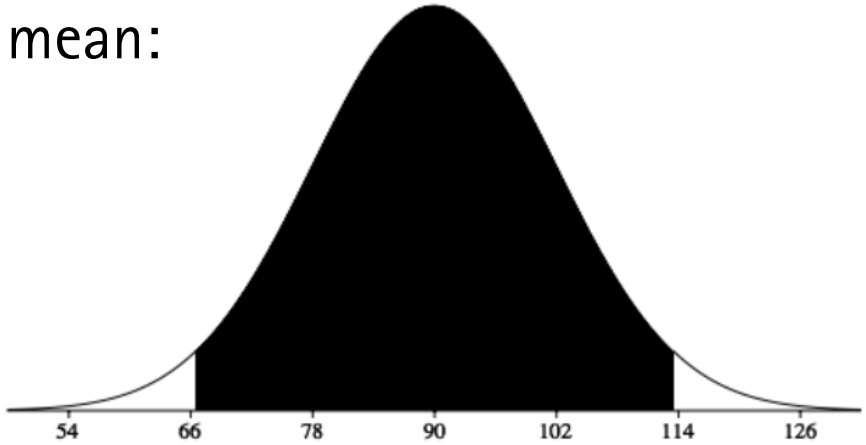
$$\sigma_M = \frac{\sigma}{\sqrt{n}}$$

where σ is the standard deviation of the population

- Example: Normally distributed data with $\sigma = 36$, n = 9:
  standard deviation of the mean is $\sigma_M = 36/3 = 12$
- Caveat: Population standard deviation is typically not known!

Confidence Interval of the Mean (if μ and σ known)

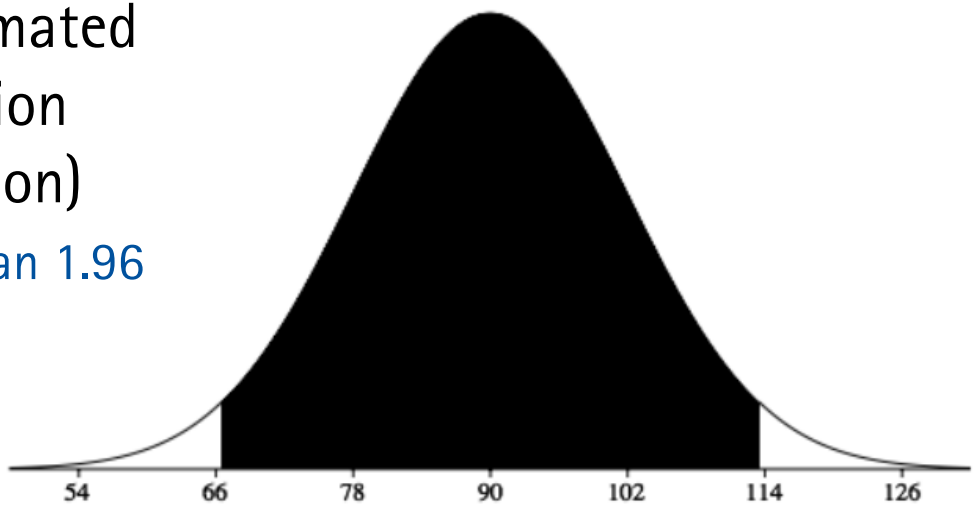
- Assume that the population parameters μ and σ are known
- If data is normally distributed, then 95% of the means are within $\pm 1.96\ \sigma$ (1.96 times the standard deviation)
- Example (continued from previous slide):
  Normally distributed data with μ = 90, σ = 36, n = 9:
  standard deviation of the mean is $\sigma_M = 36/3 = 12$
- 95% confidence interval of the mean:
  - 90 - 1.96 * 12 = 66.48
  - 90 + 1.96 * 12 = 113.52



Example and graph from: http://onlinestatbook.com/2/index.html (David M. Lane, Rice University)

Confidence Interval of the Mean (if σ known)

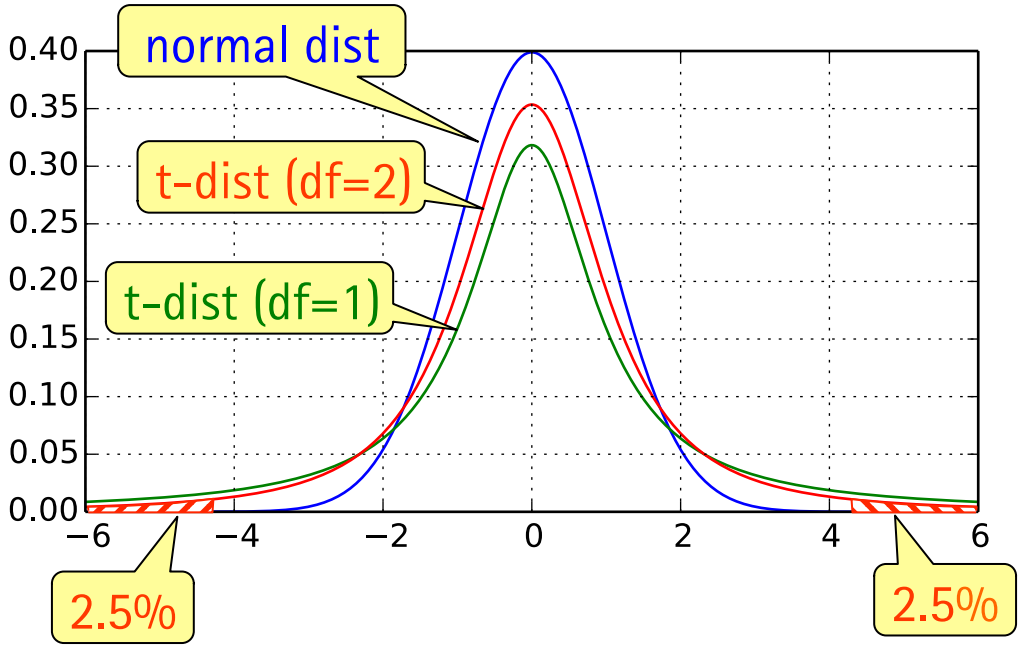
- If computing the empirical mean M for repeated experiments of sample size n, intervals [M - 1.96 σ_M, M + 1.96 σ_M] contain the population mean 95% of the time
- Caveat: Population standard deviation is typically not known!
- If standard deviation is estimated from sample, use t distribution (instead of normal distribution)
  - Yields slightly larger value than 1.96



Example and graph from: http://onlinestatbook.com/2/index.html (David M. Lane, Rice University)

2.5% and 97.5% Quantiles of the t-Distribution

df	0.975
1	12.706
2	4.303
3	3.182
4	2.776
5	2.571
8	2.306
10	2.228
20	2.086
50	2.009
100	1.984
∞	1.96



for large n, the t-distribution approaches the standard normal distribution

```
import scipy.stats as st
print st.t.ppf(0.025, df=2)
-4.30265272991
print st.t.ppf(0.975, df=2)
4.30265272991
```

Confidence Interval (if σ estimated from sample)

- Standard error of the mean is the standard deviation of the mean of samples of size n

$$s_M = \frac{s}{\sqrt{n}}$$ where s is the estimated standard deviation from the sample

- Confidence interval: $[M + t_{0.025}(df) * s_M, M + t_{0.975}(df) * s_M]$
- Example: data = {2, 3, 5, 6, 9}
  - n = 5, M = 5.0, s = 2.739, s_M = 1.225
  - degrees of freedom: df = n-1 = 4
  - $t_{0.025}(4) = -2.776$, $t_{0.975}(4) = 2.776$
  - 95% confidence interval: [5.0 - 2.776*1.225, 5.0 + 2.776*1.225] = [1.6, 8.4]
- If repeating the experiment many times, the population mean will be contained in 95% of the computed confidence intervals

Example and graph from: http://onlinestatbook.com/2/index.html (David M. Lane, Rice University)

Confidence Intervals when Reporting Results

- Confidence intervals help judge the variability of the data
- When means are shown (e.g., in bar chars) confidence intervals should be shown as well (as error bars)
  - The caption should say what the error bars mean
- Confidence intervals are easier interpret than standard error or standard deviation
  - But standard error looks better, because only (about) half the size!

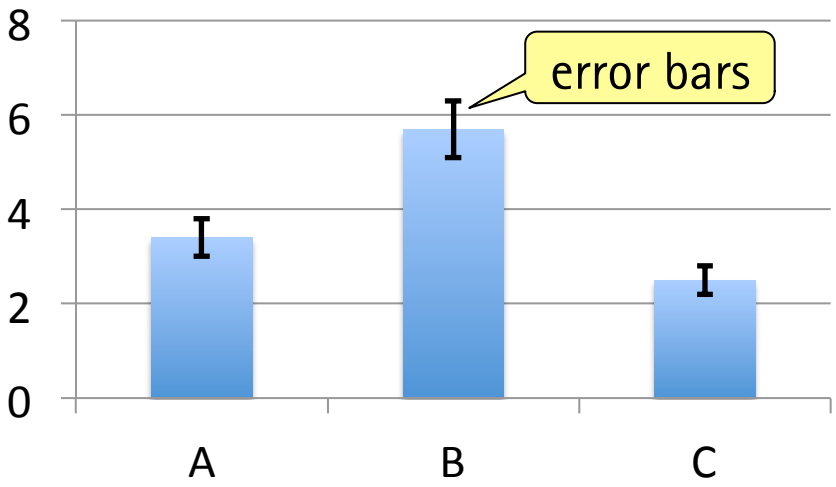


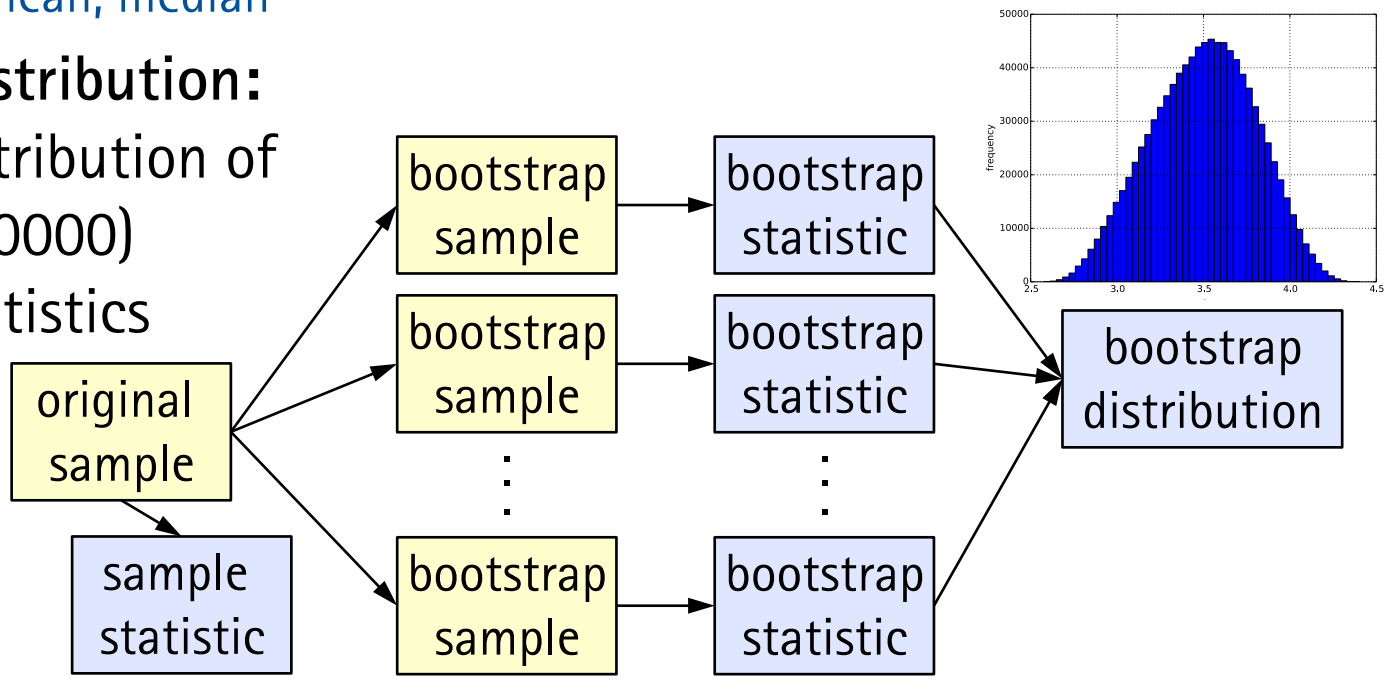
Figure X. Task completion times (in s) for methods A, B, and C. Error bars show 95% confidence intervals.

Bootstrapping Confidence Intervals for Means

- Confidence intervals for means can also be estimated using randomization techniques
  - Useful if distribution of data is unknown
- Bootstrapping
  - Assume that data is a representative sample (size n) of the population
    - Imagine population as a large number of copies of original sample
  - Take repeated samples (size n) from the original sample with replacement
    - i.e. treat original sample as "population"
    - the repeated samples are called bootstrap samples
  - Compute the mean of each bootstrap sample
  - Create a histogram of the means
  - Take the $\alpha/2$ and the $1-\alpha/2$ percentiles as the confidence limits

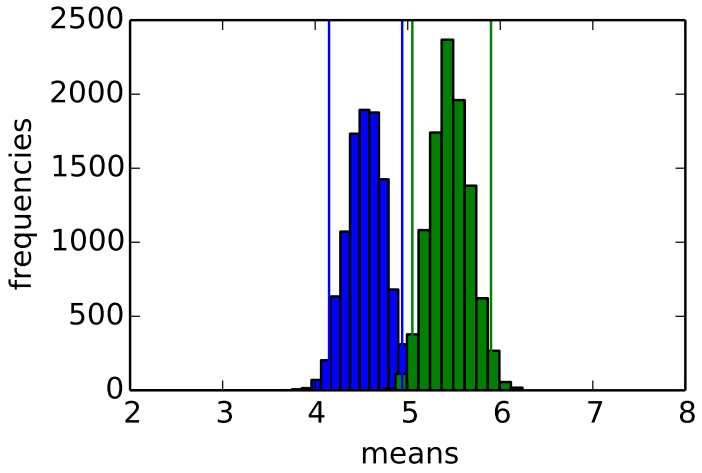
Bootstrap Terms

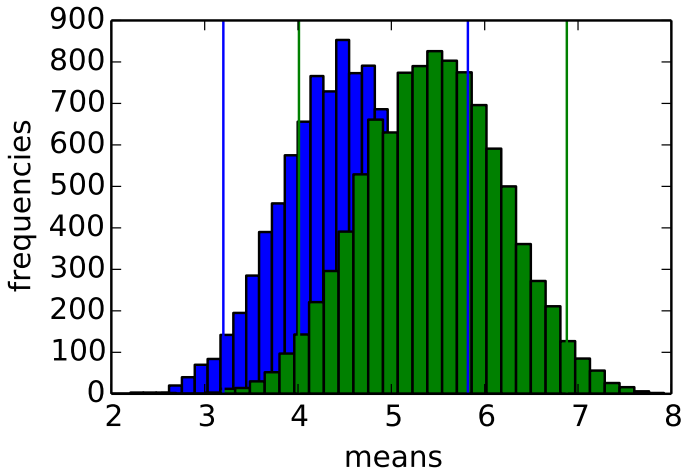
- **Bootstrap sample:** Random sample (size n) taken with replacement from original sample (size n)
- **Bootstrap statistic:** Statistic computed on bootstrap sample
  - Examples: mean, median
- **Bootstrap distribution:** Empirical distribution of many (e.g., 10000) bootstrap statistics



Percentile Method for Computing Confidence Intervals

- Percentile method
  - For 95% confidence interval, keep middle 95% of bootstrap distribution
  - Remove 2.5% in each tail (2.5% percentile and 97.5% percentile)
- Requirements for using percentile method with bootstrap distribution
  - (approximately) Continuous (no gaps)
  - (approximately) Symmetric (not skewed)
- There are more advanced methods than the percentile method





LINEAR RELATIONSHIPS

Correlation

- Correlations describe the linear association between two variables
- Does one variable vary in a similar manner as another variable?
- Correlation coefficient

$$r = \frac{s_{xy}}{s_x \cdot s_y} \quad \text{with} \quad s_{xy} = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{n-1} \qquad s_x = \sqrt{\frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{n-1}} \qquad s_y = \sqrt{\frac{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}{n-1}}$$

- $r \in [-1,1]$
- See link to scalar product?

$$r = \frac{u^T \cdot v}{\|u\| \cdot \|v\|} \qquad u = \begin{pmatrix} X_1 - \bar{X} \\ \dots \\ X_n - \bar{X} \end{pmatrix} \qquad v = \begin{pmatrix} Y_1 - \bar{Y} \\ \dots \\ Y_n - \bar{Y} \end{pmatrix}$$

Correlation Coefficients (between x and y)

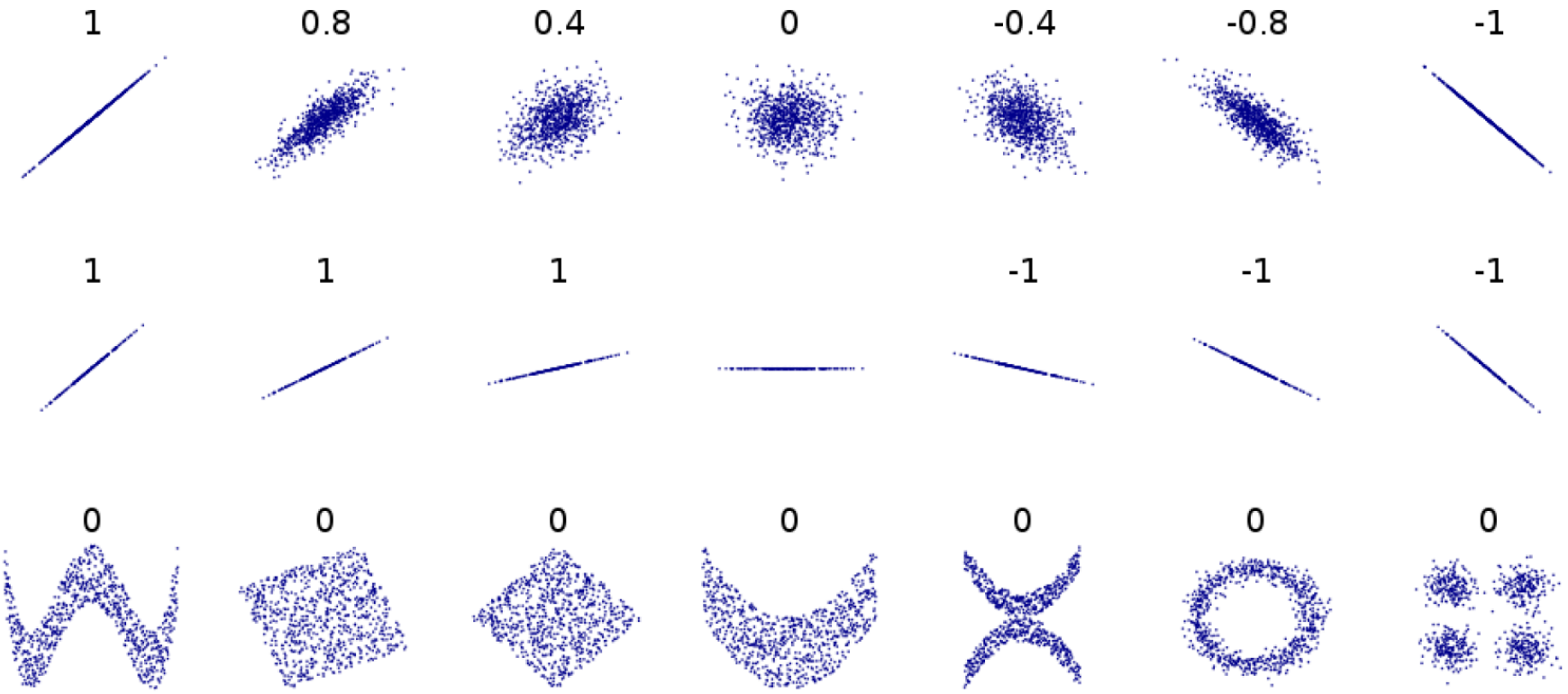
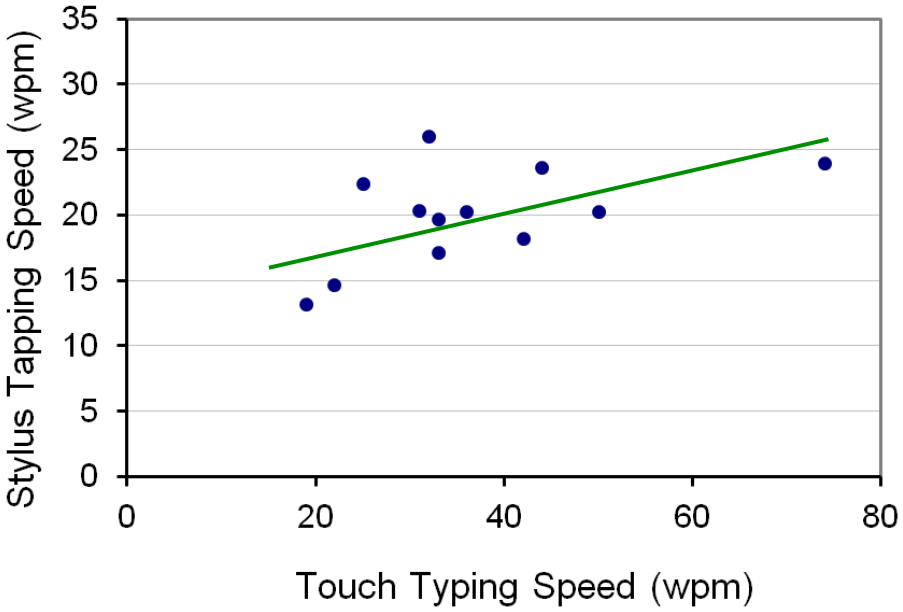


Figure from Wikipedia

Correlation between Stylus Tapping Speed and Touch Typing Speed



Participant	Stylus Tapping Speed (wpm)	Touch Typing Speed (wpm)
P1	18.2	42
P2	23.6	44
P3	26.0	32
P4	20.3	50
P5	20.3	36
P6	17.1	33
P7	24.0	74
P8	14.7	22
P9	20.3	31
P10	19.7	33
P11	22.4	25
P12	13.1	19

- There seems to be a relationship
  - Faster touch typists seem to be faster at stylus tapping
- Correlation coefficient
  - $r = 0.5250$ ← but is this r significantly different from 0?

MacKenzie: Human-Computer Interaction - An Empirical Research Perspective.

Randomization Test for Correlation

- Null hypothesis tests can be applied to correlations
  - Null hypothesis: correlation is zero
- Randomization tests can be applied to correlations as well
- Algorithm idea
  - Randomly rearrange associations of values (if the null hypothesis is true, then each rearrangement is as likely as each other)
  - Compute correlation for each rearrangement
  - Compute proportion of rearrangements for which correlation is more extreme than observed correlation (in the original sample)
  - If a correlation as extreme as the observed correlation only happens for less than 5% of the rearrangements (unlikely under the null hypothesis), call the correlation significant, otherwise call it non-significant

Randomization Test for Correlation

$A = \{t_{11}, t_{12}, ..., t_{1n}\}$ // variable 1

$B = \{t_{2,(n+1)}, t_{2,(n+2)}, ..., t_{2,2n}\}$ // variable 2

obtainedR = corr(A, B) // original correlation

absR = abs(obtainedR)

rep = 10000 // repetitions

R = zeros(rep)

for i = 1..rep: // repetitions

    shuffledB = shuffle(B) // randomly shuffle one of the two variables

    R_i = corr(A, shuffledB)) // correlation of rearrangement

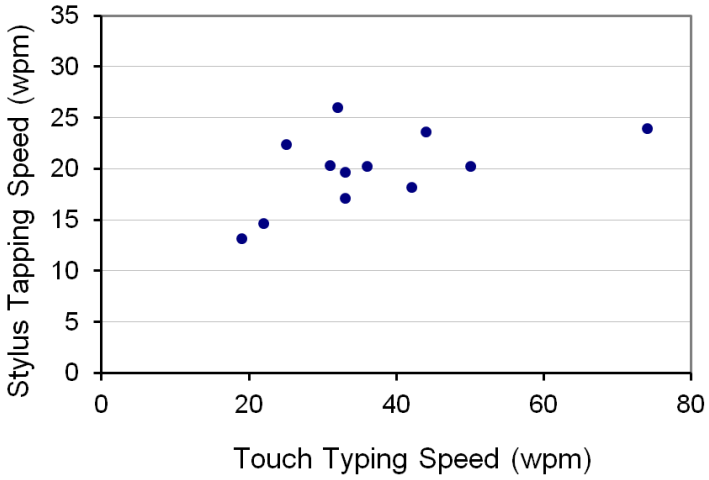
leftTail = count($R_i \leq$ -absR) // number of more extreme values in left tail

rightTail = count($R_i \geq$ absR) // number of more extreme values in right tail

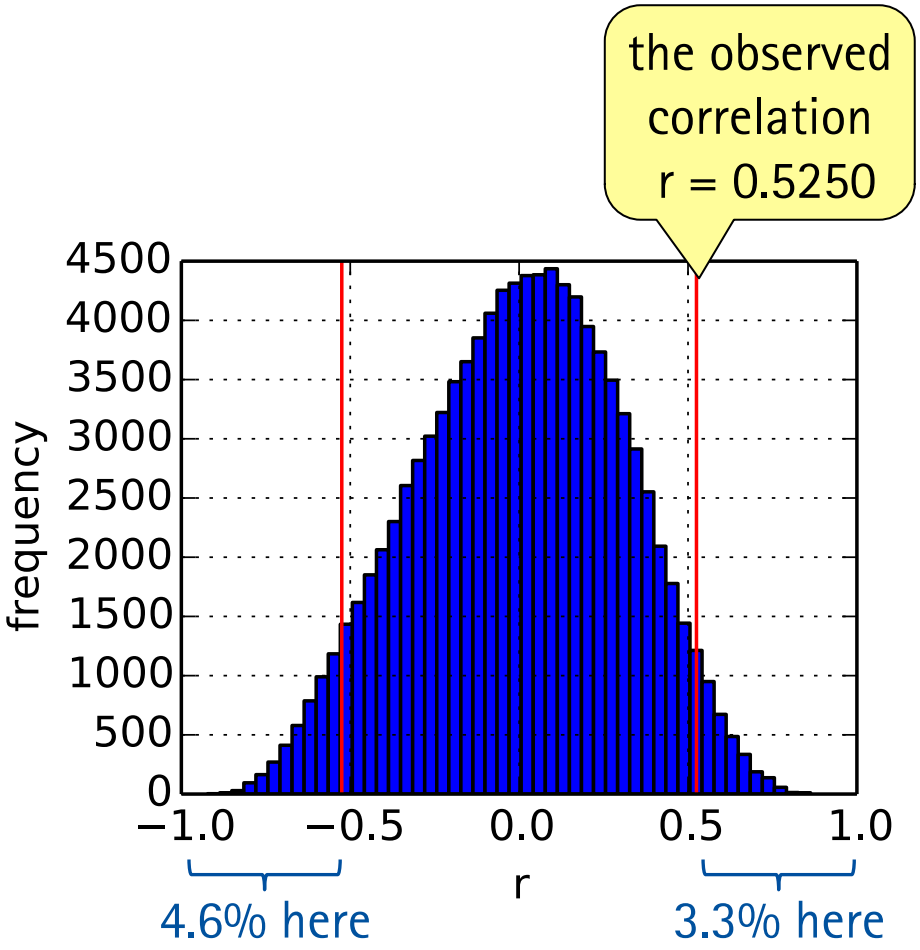
p = (lefTail + rightTail) / rep // empirical probability

if $p \leq 0.05$ then "significant" else "not significant"

Correlation between Stylus Tapping Speed and Touch Typing Speed

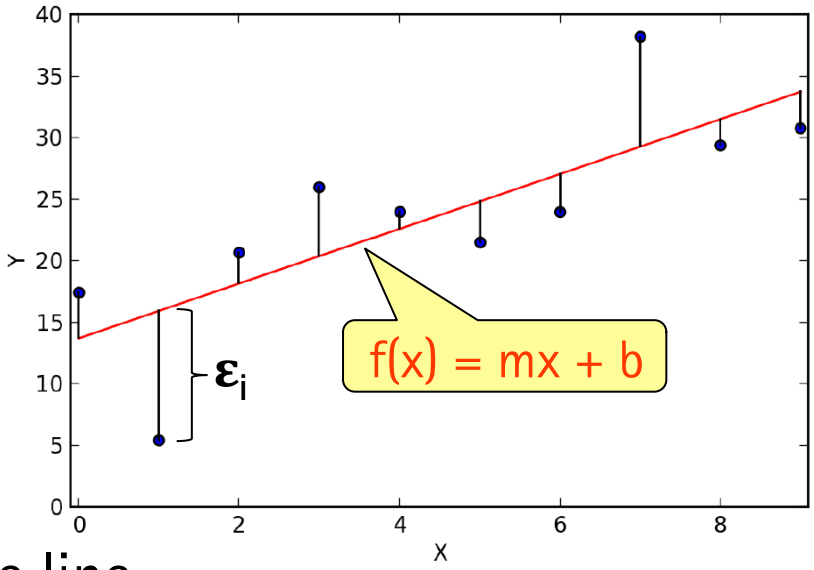


- p = 8.0% → not significant
- We cannot reject the null hypothesis that there is no linear relationship between the two speeds (ρ=0)



MacKnzie: Human-Computer Interaction - An Empirical Research Perspective.

Linear Regression



- Given data $\{(x_1, y_1), \ldots, (x_n, y_n)\}$
- Find slope m and intercept b of the best-fitting line
  - $f(x) = mx + b$
- Measured data y_i not always on the line
  - $y_i = f(x_i) + \boldsymbol{\varepsilon}_i = m\, x_i + b + \boldsymbol{\varepsilon}_i$
  - $\boldsymbol{\varepsilon}_i$: errors (aka. residuals)
- Measured and predicted values
  - $m\, x_i + b$ is the prediction at point x_i, at which we measured y_i
  - Measured values: $y_1, \ldots, y_n$
  - Predicted values: $f(x_1), \ldots, f(x_n)$

Figure: Haslwanter: An Introduction to Statistics

Linear Regression

- Find slope m and intercept b of the best-fitting line
- Best-fitting line minimizes squared errors
  - Errors: $\boldsymbol{\varepsilon}_i = y_i - (m\, x_i + b)$
  - Minimize $e(m, b) = \text{sum}(\boldsymbol{\varepsilon}_i^2) = \text{sum}((y_i - m\, x_i - b)^2)$
  - Differentiate e w.r.t m and b, set to 0, solve linear system
- In general: Predict y_i from Vector $(x_{i1}, \ldots, x_{ip})$
  - data = $(x_{11}, \ldots, x_{1p}; y_1), \quad \ldots, \quad (x_{n1}, \ldots, x_{np}; y_n)$
  - $y = X\beta + \boldsymbol{\varepsilon}$
    - y and $\boldsymbol{\varepsilon}$ are n-Vectors
    - β is a (p+1)-Vector
    - X is a nx(p+1)-Matrix (first column all 1s)

Figure: Haslwanter: An Introduction to Statistics

Linear Regression

- $y = X\beta + \boldsymbol{\varepsilon}$
- Example: 4 observations: (x_i, y_i), i=1..4
- Equation of regression model

$$\begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \end{bmatrix} = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ 1 & x_3 \\ 1 & x_4 \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \varepsilon_3 \\ \varepsilon_4 \end{bmatrix}$$

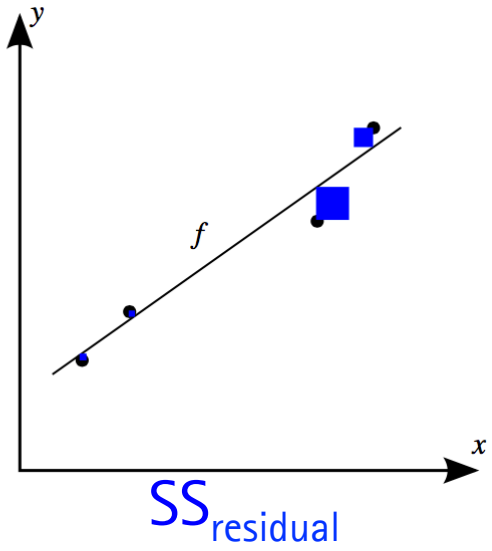
- Fitting this model with Scipy

```
import scipy.stats as st
slope, intercept, r, p, e = st.linregress(x, y)
print slope, intercept
```

Haslwanter: An Introduction to Statistics

Coefficient of Determination

- Coefficient of determination
  - $R^2 = 1 - SS_{residual} / SS_{total}$
  - $SS_{residual} = sum((y_i - f_i)^2)$ (unexplained variability)
  - $SS_{total} = sum((y_i - y_{mean})^2)$ (total variability)
  - y_i: the data
  - f_i: predicted values on the regression line
- Coefficient of determination
  - $R^2 = r^2$, $R^2 \in [0,1]$
  - Squared correlation coefficient between observed (y_i) and predicted (f_i) values
  - R close to 1: high determination
  - R close to 0: low determination



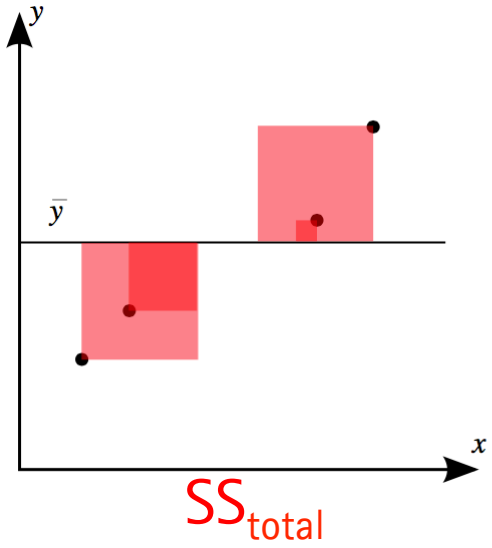


Figure from Wikipedia

TOOLS

Data Analysis Tools, Python Libraries

- Python: Base language
- NumPy: Array and matrix operations
- SciPy: Scientific computing
- Matplotlib: Creating graphs and figures
- Pandas: Data manipulation and analysis
- Statsmodels: Statistics...
- iPython: Notebook for programming (and graphs and and notes)

- Anaconda: Package of these (and other libraries)
  - https://store.continuum.io/cshop/anaconda/

iPython Notebook

- Local server
- Browser-based UI
- Cells with code snippets
- Autocompletion
- Documentation
- Graphics inline
- Formatted text inline

IP[y]: Notebook ReactionTimesBlocksANOVA (autosaved)

File Edit View Insert Cell Kernel Help

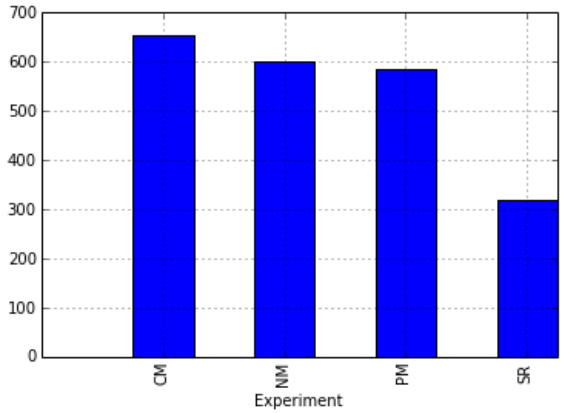
Markdown Cell Toolbar: None

Summary statistics:

In [17]:
```
pet = df[['Experiment', 'Time']].groupby(['Experiment']).mean()
print pet
pet.Time.plot(kind="bar")
```

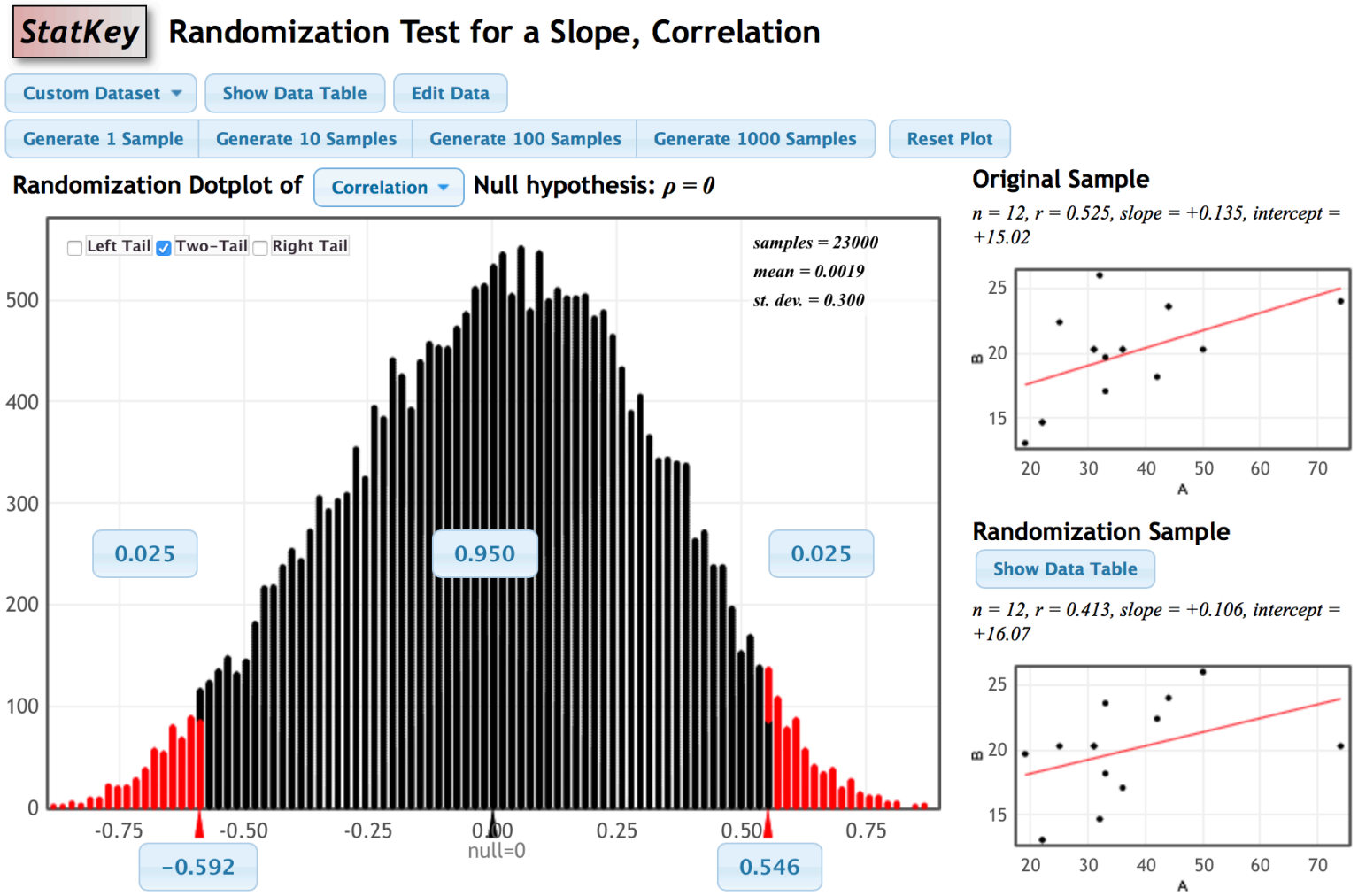
```
                Time
Experiment
CM         652.644000
NM         599.398667
PM         586.261333
SR         318.540000
```

Out[17]: <matplotlib.axes.AxesSubplot at 0x109756a90>



In [18]:
```
#print df.head()
pet = df[df['Block']>1][['Experiment', 'Time']]
pet = pet.groupby(['Experiment']).mean()
print pet
pet.Time.plot(kind="bar")
```

Online Tool for Randomization Tests: lock5stat.com/statkey



NEXT ASSIGNMENTS…

Experiment Exercises

- Find a small interaction to improve, create an experiment
- 14.6.: Experiment planning
  - Review literature, or find example from your own experience
  - Define independent variable, define dependent variable
  - Write up experimental design
- 21.6.: Prototype development
  - Implement test prototype in JavaFX, should log and save data as csv
- 28.6.: Experiment execution
  - Perform your experiment yourself and find 4 other participants
  - Perform 4 experiments from others yourself (remotely, email prototype)
  - Collect data (email data)
- 8.7.: Data analysis
  - Analyze data, write up findings

Example Study

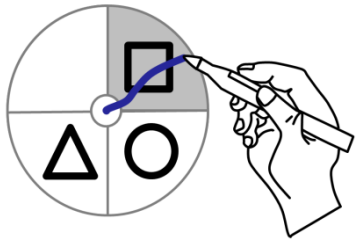
- Check whether progress bar duration is perceived differently depending on how progress evolves
- Equal duration, but: linear, fast-then-slow, slow-then-fast
- Let participants rank which is faster
- Replication of (a small part of):
  - Harrison, Yeo, Hudson: Faster Progress Bars: Manipulating Perceived Duration with Visual Augmentations. CHI 2010.
  - http://www.chrisharrison.net/index.php/Research/ProgressBars2
- Other possibilities
  - Time for searching an item depending on its visual salience
  - Time for selecting an item depending on target size or cursor technique
  - Preference for a layout

Leibniz Universität Hannover

RESEARCH HINTS

Research Topics

- Finding a research topic is a challenge (for students… and for seasoned researchers too!)



- Four tips:
  - 1. Think small
  - 2. Replicate
  - 3. Know the literature
  - 4. Think inside the box

MacKenzie: Human-Computer Interaction - An Empirical Research Perspective.

Tip #1 – Think Small

- Looking for that big idea?
- Advice: Forget it (besides, it isn't necessary)
- Research questions are small, narrowly focused
  - Do not try to develop and evaluate a UI for a complete system
  - Focus on a small aspect of the interface or interaction
- Pursue several small, related research topics, and before you know it, a broader topic is formed
  - Focus on one aspect often yields further ideas for improvement

MacKenzie: Human-Computer Interaction - An Empirical Research Perspective.

Tip #2 – Replicate

- Start by replicating research that was done before
  - Of course, there is no research in bare replication
- Replicating prior research is a lot of work
  - Studying a paper, implementing the test prototype, running and analyzing the experiment, comparing the results
- Along the way, you will discover small and novel improvements – things to try
- From what is learned in this process
  - A little tweak here, a small modification there
- You might not find a novel idea until well into the process

MacKenzie: Human-Computer Interaction - An Empirical Research Perspective.

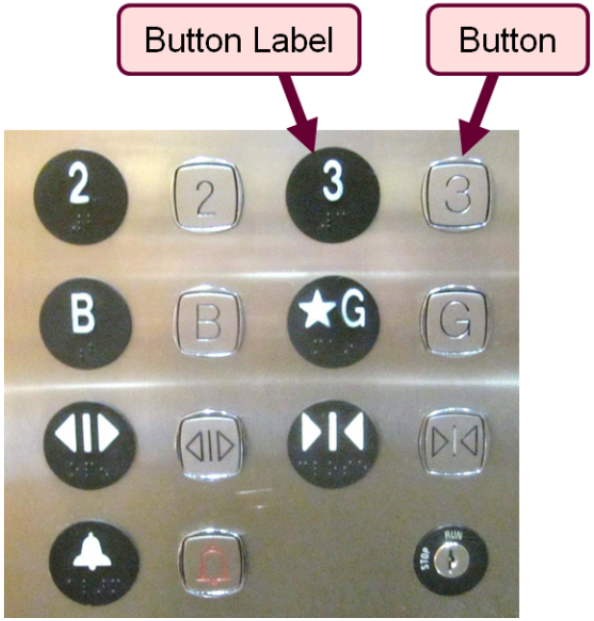
Tip #3 – Know The Literature

- Whatever topic interests you, read the literature
  - E.g., social networking, gaming
- If too broad, narrow
  - E.g., privacy settings in social networking, avatars in gaming
- Read papers, open a spreadsheet, tabulate variables in the methodology and the findings
- Chaotic at first, order and shape will emerge (eventually)
- With some luck (and further study) a research topic will emerge

MacKenzie: Human-Computer Interaction - An Empirical Research Perspective.

Tip #4 – Think Inside The Box

- Think outside the box → dispense with accepted beliefs and assumptions (in the box), and think in a way that assumes nothing and challenges everything
- No need
- Think inside the box: just get on with your day; but at every juncture, every interaction, think and question
- Our everyday foibles are fertile ground for research topics



MacKenzie: Human-Computer Interaction - An Empirical Research Perspective.