Mensch-Maschine-Interaktion 1

Chapter 4 continued (June 10, 2010, 9am-12pm):

User Study Statistics

Looking Back: User Study Design

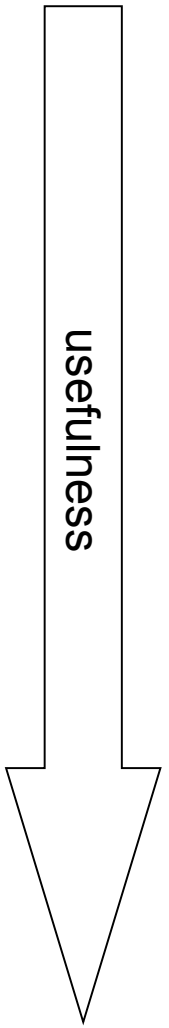
- Purpose of user studies
- Placement within the development process
- Types of user studies
  - Observational, experimental
  - Within subjects, between groups
- Independent vs. dependent variables
- Setup process
  - Form hypotheses → design the study → run a pilot study → recruit participants → run the study → analyze the data
  - Results must be valid, reliable, generalisable, important

User Study Design

- The Purpose of User Studies
- Research Aims: Reliability, Validity and Generalizability
- Research Methods and Experimental Designs
- Ethical Considerations
- HCI-related and practical information for your own studies
- Interpretation of Data and Presentation of Results

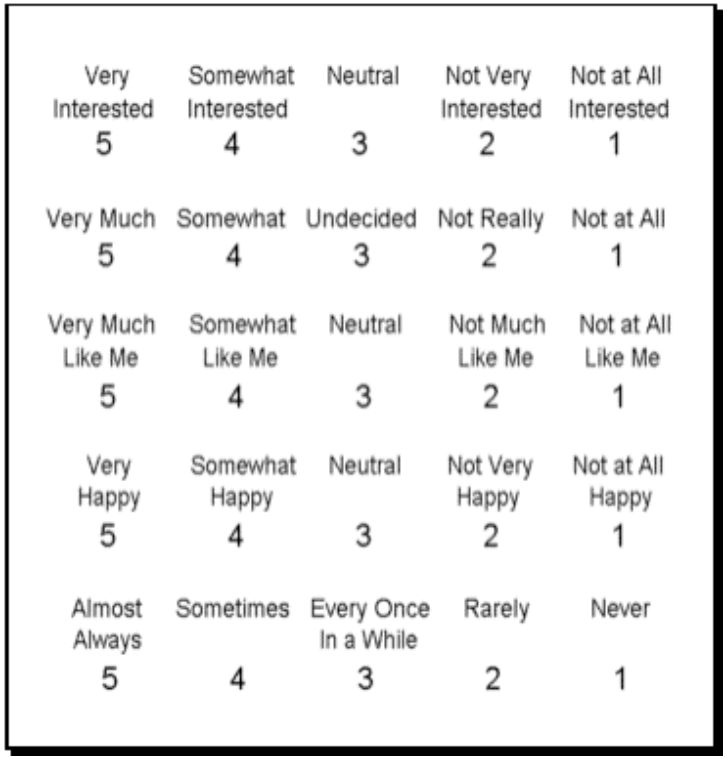
Types of Data

- Nominal (categorical) data
  - No relationship between the size of the number
  - Operations: A=B, A!=B
  - E.g. numbers in a football team
- Ordinal Data
  - Order / ranking
  - Operations: A>B, A<B, A=B
  - E.g. marks in school: 1, 2, 3, 4, 5, 6
- Interval scale data
  - Equal intervals = equal differences in the measured property
  - Zero point is arbitrary
  - E.g. temperature (°C/°F)
- Ratio scale data
  - Fixed zero point
  - E.g. wpm, error rates



Types of Variables

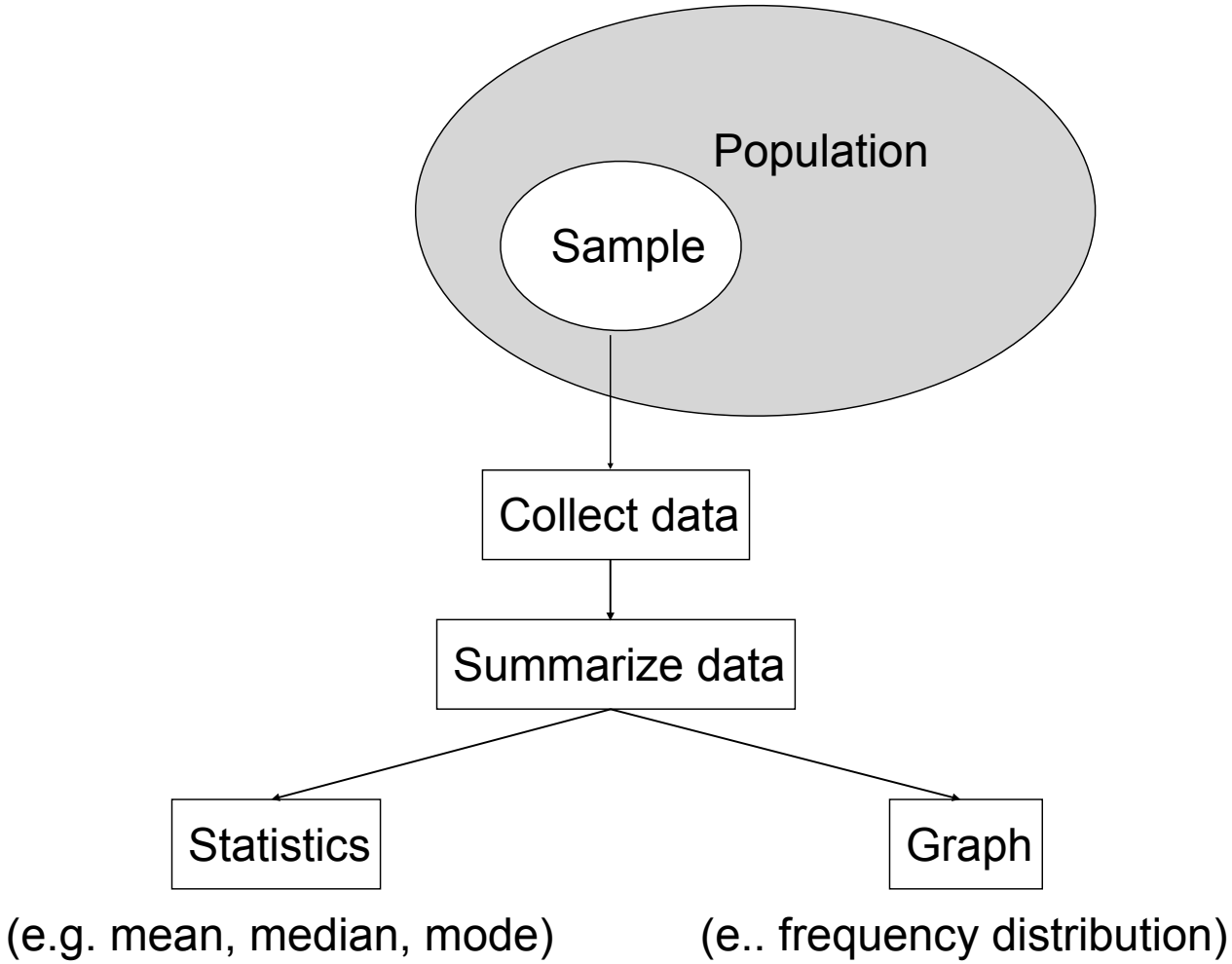
- Discrete Data
  - Distinct and separate
  - Can be counted
  - E.g. Likert scales, preferences from a list, ...
- Continuous Data
  - Any value within a finite or infinite interval
  - Always have a order
  - E.g. weight, length, task completion time, ...



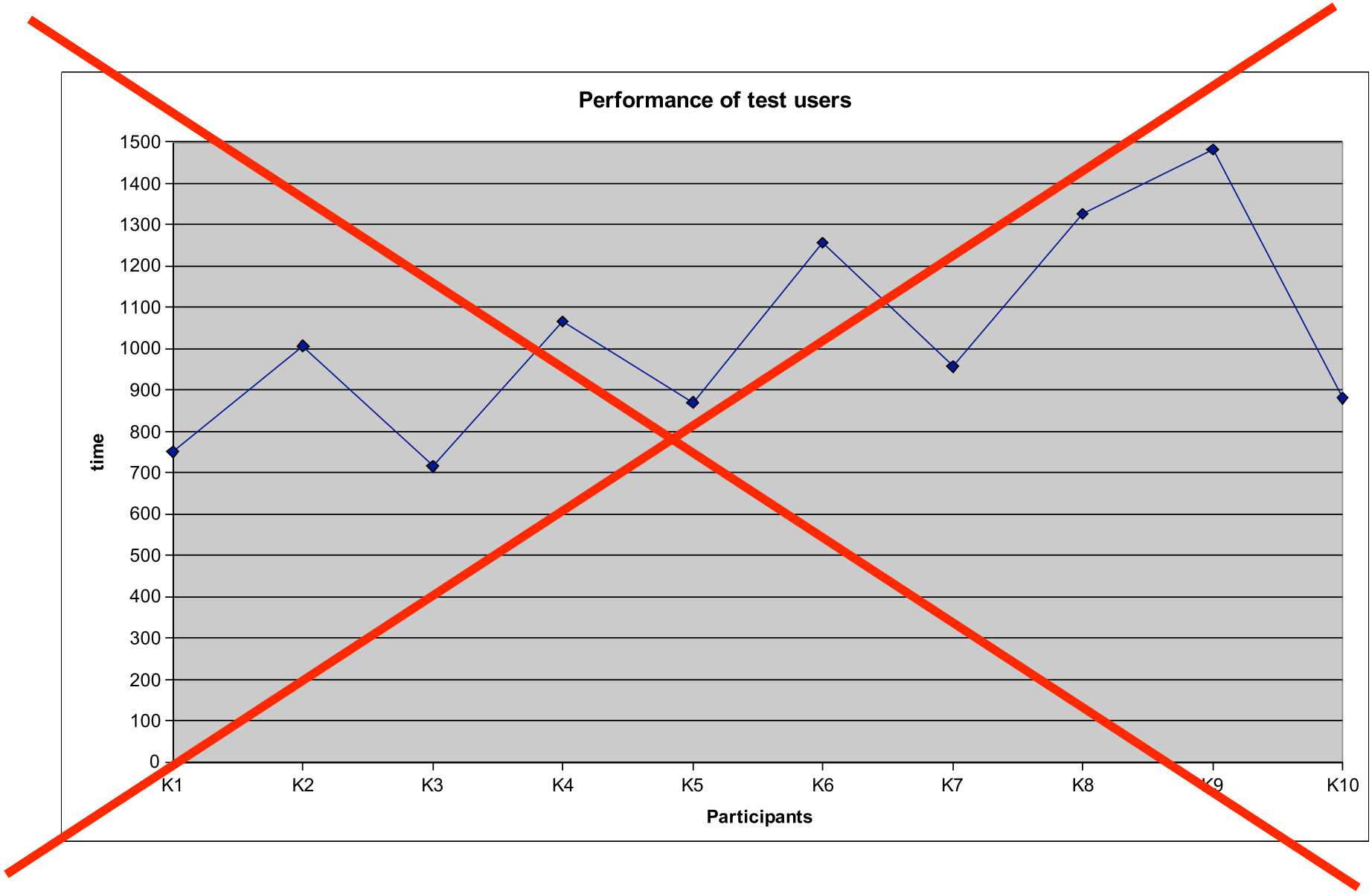
5-point Likert Scales
http://allpsych.com/researchmethods/images/likertscales.gif

Summarizing Data

- Collected data needs to be summarized
  - Recognize patterns
  - Aggregate data
- Two ways:
  - Statistics
  - Graph



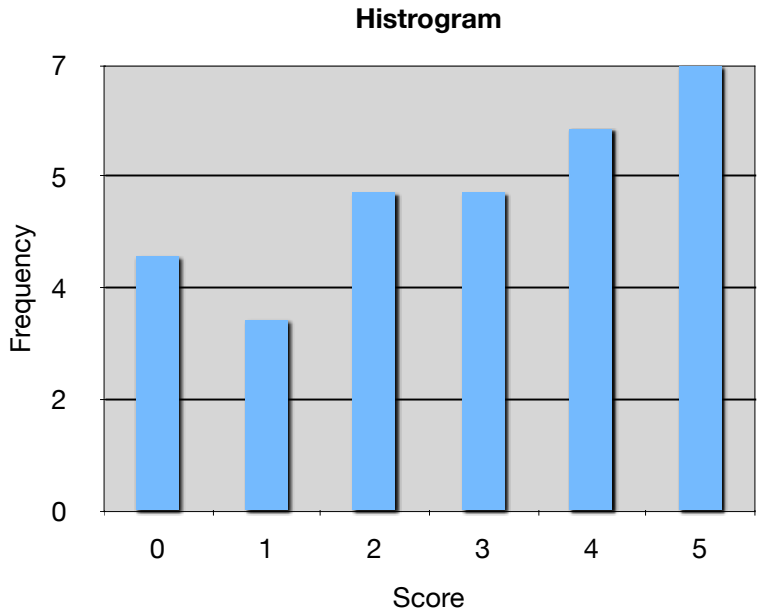
Don’t Do This


Performance of test users
time
1500
1400
1300
1200
1100
1000
900
800
700
600
500
400
300
200
100
0
K1
K2
K3
K4
K5
K6
K7
K8
K9
K10
Participants

Frequency Distributions (Histograms)

- Example: days needed to answer my email
  Data: 5 2 2 3 4 4 3 2 0 3 0 3 2 1 5 1 3 1 5 5 2 4 0 0 4 5 4 4 5 5
- Count the number of times each score occurs
  ⇨ Frequency table:

Days	Frequency	Frequency (%)
0	4	13%
1	3	10%
2	5	17%
3	5	17%
4	6	20%
5	7	23%



Averages: Mode, Median, Mean

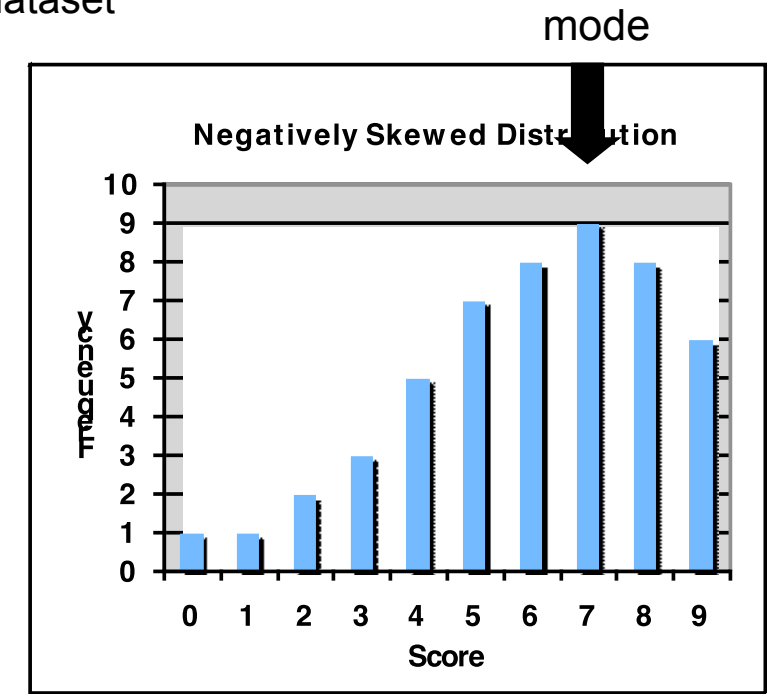
- How can the data be summed up in a single value?
- Idea: get the centric point

- Three ways:
  - Mode
    - The most frequent score
  - Median
    - Middle score
  - Mean
    - Average

Mode

- The most frequent score
- Describes how most people behave

- Pros:
  - Easy to calculate and understand
  - Can be used with nominal data
- Cons:
  - There can be more than one modes
  - Mode can change dramatically by adding only one dataset
  - Independent of all other data in the set



Median (Mdn)

- Middle score of the distribution

  Example data: 1 7 3 9 6 9 2

- Sorted by magnitude: 9 9 7 6 3 2 1 ⇨ median = 6
- If #scores even ⇨ average two middle scores

  Example data: 1 7 3 9 4 6 9 2

- Sorted by magnitude: 9 9 7 6 4 3 2 1 ⇨ median = 5
- Pros:
  - Relatively unaffected by outliers (very low or high scores) and skewed distributions
  - Can be used with ordinal, interval and ratio data
- Cons:
  - Does not consider all scores of the data set
  - Not very stable

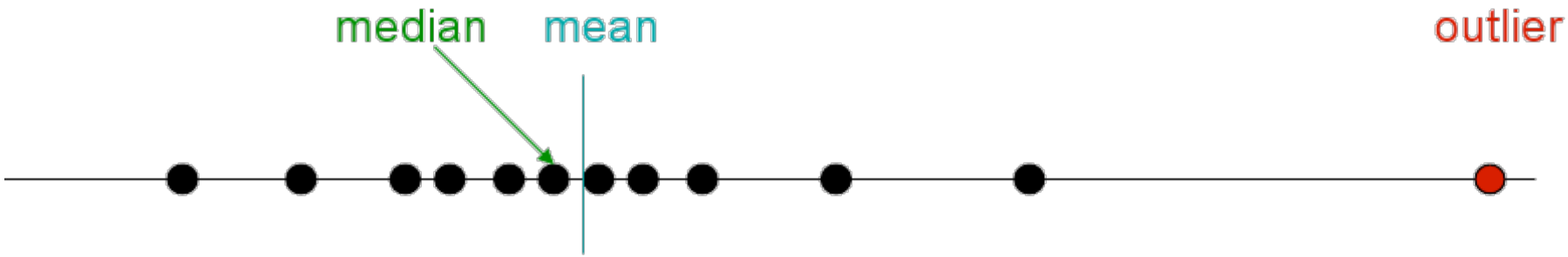
if n is odd: $x_{(n+1)/2}$

if n is even: $(x_{n/2} + x_{n/2+1}) / 2$

Mean (M)

- Sum of all scores divided by #scores:
- Most often used if ‘average’ is mentioned
- Pros:
  - Considers every score
    - ⇨ most accurate summary of the data
  - Resistant to sampling variation: removing one sample changes the mean far less than mode or median
- Cons:
  - Heavily affected by extreme scores and skewed distributions
  - Can only be used with interval and ratio data

$$m = \frac{1}{n} \sum_{i=1}^{n} X_i$$



Averages for Likert-Scales?

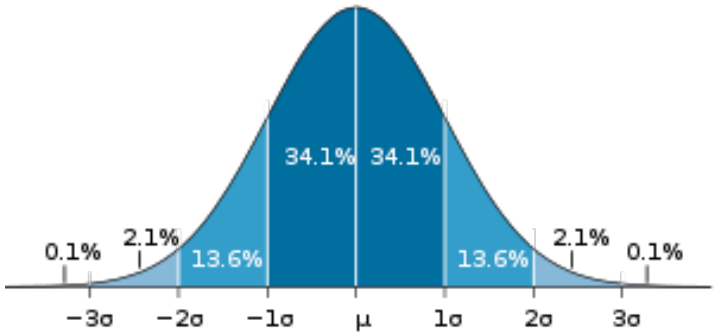
- Average: what does 2.5 mean?!
  - Distances between each item on the scale might be different
    e.g. between ‘neutral’ and ‘agree’ vs. ‘agree’ and ‘totally agree’
  - Does not show the distribution (half disagree, half agree vs. all neutral)
    - This could be done with standard deviation
- Mode:
  - Shows the most frequent opinion
  - ... but not whether this was the majority
  - ... but not the distribution (half disagree,
    half agree vs. all neutral)
- Mean:
  - Gives some indication about the overall distribution
  - ... but not about outliers
- => report frequencies of all items
- => otherwise, if it must be one value,
  mode is most often used

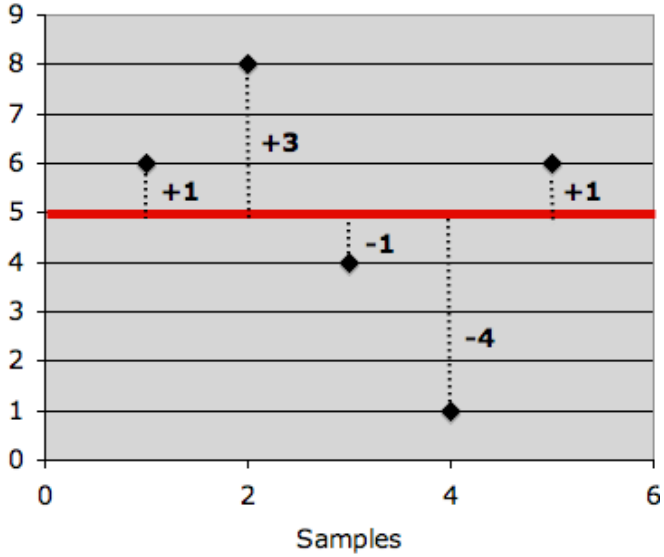
Standard Deviation and Variance

- How do you measure the accuracy of the mean?
- Example data set 1: 5 5 5 5 5 ⇨ mean = 5
- Example data set 2: 6 8 4 1 6 ⇨ mean = 5
- Which of the data sets is better reflected by the mean?

- If $x_1, x_2, \ldots x_n$ are the data in a sample with mean m
  - **Deviation** = difference between mean and scores $= \sum (x_i - m)$
  - **Variance** $s^2 = \frac{\sum(x_i - m)^2}{n}$ $(= E(X^2) - m^2)$
  - **Standard deviation (SD)** $s = \sqrt{Var(X)}$

- Both variance and standard deviations measure the
  - Accuracy of the data set
  - Variability of the data

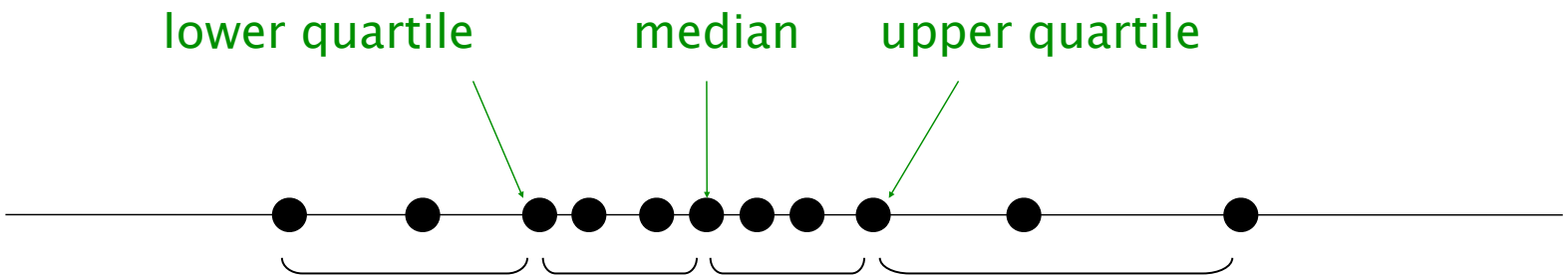




http://en.wikipedia.org/wiki/Normal_distribution

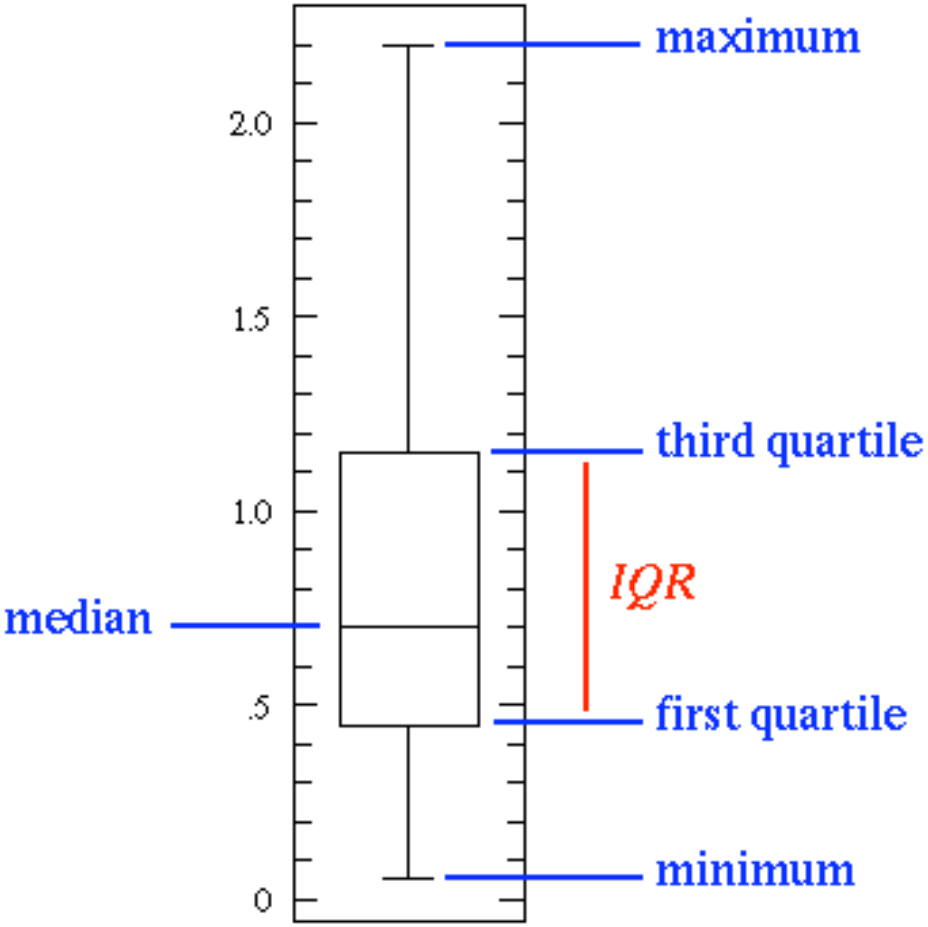
Quantile, Quartile and Percentile

- Quantile
  - ‘Cut points' that divide a sample of data into groups containing (as far as possible) equal numbers of observations.
- Quartile (Quantile of 4)
  - Values that divide a sample of data into 4 groups containing (as far as possible) equal numbers of observations
- Percentile (Quantile of 100)
  - Values that divide a sample of data into 100 groups containing (as far as possible) equal numbers of observations



Boxplots

- Also known as
  - box-and-whisker diagram
  - candlestick chart
- Quick overview of the most important values

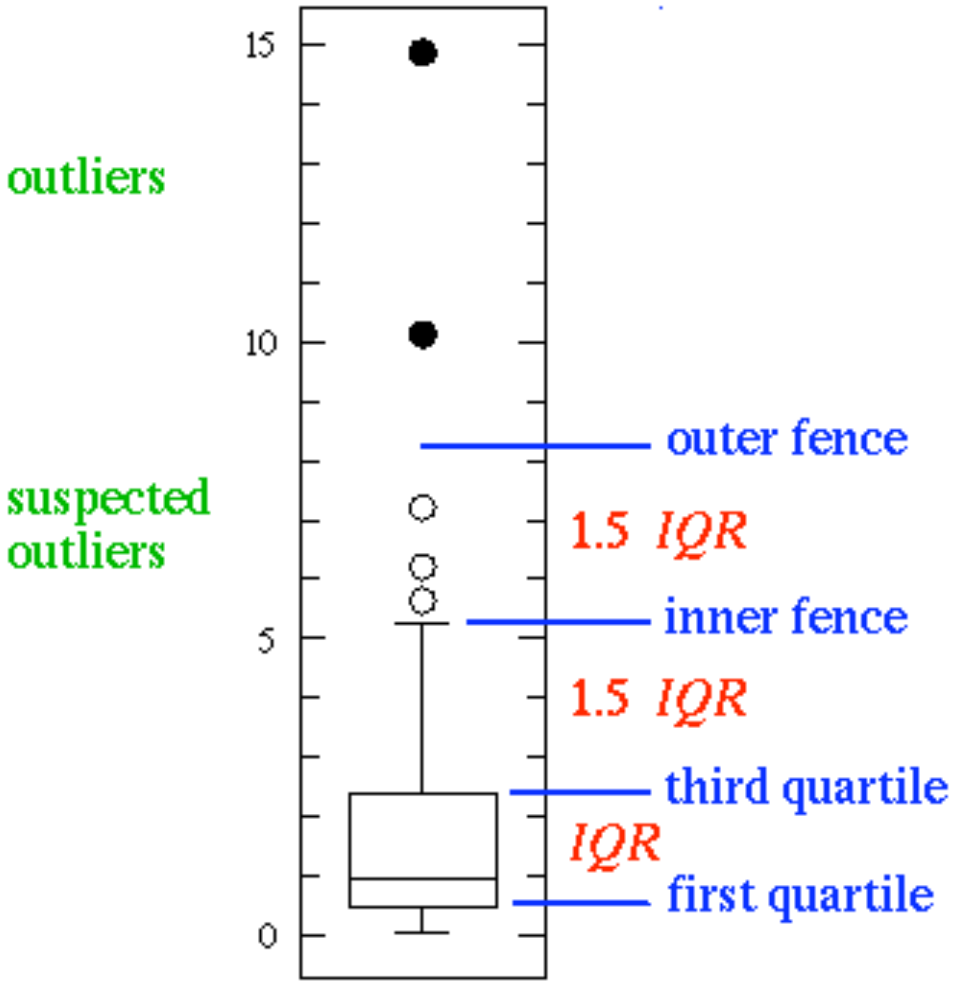


Source: http://www.physics.csbsju.edu/stats/box2.html

Outliers

- Try to avoid outliers!
  - Improve your test equipment
  - Eliminate sources of disturbances
  - Repeat parts of your experiment in case of disturbance

- Outliers are not generally bad – they give valuable information
- With large data sets outliers can often not be avoided



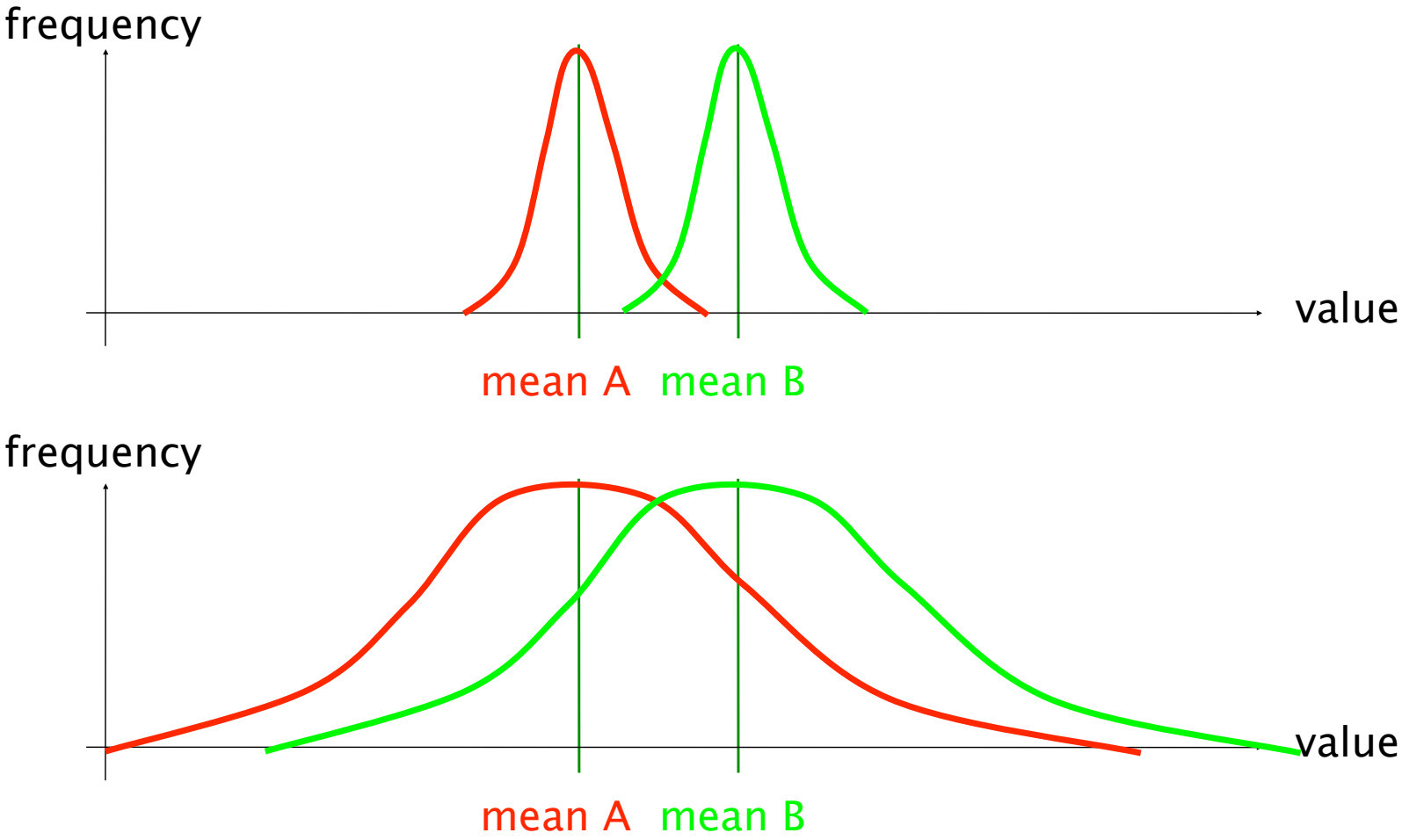
Creating Boxplots with Excel

- Useful functions in Excel (and many other applications)
  - MIN, MAX
  - MEDIAN
  - AVERAGE
  - QUARTILE
  - PERCENTILE

- Box Plots with Excel 2007
  - http://blog.immeria.net/2007/01/box-plot-and-whisker-plots-in-excel.html
  - http://www.bloggpro.com/box-plot-for-excel-2007/

Comparing Values

- Significant differences between measurements?



Example: Pepsi Challenge

- The Pepsi Challenge
  - Let participants „blindly“ taste glasses of Pepsi/Coca Cola and identify it
  - Half the glasses are filled with Pepsi, half with Coca Cola
  - 2 glasses ⇨ chance of guessing correct = (1:2)
  - 4 glasses ⇨ chance of guessing correct = (1:6)
  - 6 glasses ⇨ chance of guessing correct = (1:20)
  - 8 glasses ⇨ chance of guessing correct = (1:70)

  ⇨ More choices means less probable that the result occurred by chance

- Differences can be due to
  - The manipulation caused a real difference
  - The difference occurred by chance
- Appropriate level of confidence: 95%
- **Significance**: A difference is „significant“ if the probability of the result occurring by chance ≤ 5%

Significance

- In statistics, a result is called significant if it is unlikely (probability p ≤ 5%) to have occurred by chance.
- **Never use the word significant if you don‘t mean statistically significant!**
- It does not necessarily mean that the result is of practical significance!

- T-Test can be used to calculate the probability p
  - The t-test gives the probability that both populations have the same mean (and thus their differences are due to random noise)
- A result of 0.05 from a t-test is a 5% chance for the same mean

T-Test in Excel

- Mean and T-Test can be calculated using MS Excel
  - AVERAGE
  - TTEST
- TTEST(…) Parameters:
  1. Data row 1
  2. Data row 2
  3. Ends / Tails (e.g. A higher B => 1-tailed; A different from B => 2-tailed)
  4. Type (use ‘paired’ for within-subjects tests)

	A	B
K1	751	1097
K2	1007	971,5
K3	716	1121
K4	1066,5	1096,5
K5	871	932
K6	1256,5	926,5
K7	957	1111
K8	1327	1211,5
K9	1482	1062
K10	881	976
Mean	**1031,5**	**1050,5**
T-test	**0,8236863**	

	A	B
K1	826,5	1382
K2	806	1066
K3	791	1276,5
K4	896,5	1352
K5	696	1191
K6	1121	1066
K7	891	1217
K8	1327	1412
K9	1277	1266,5
K10	656	1101
Mean	**928,8**	**1233**
T-test	**0,0020363**	

Analysis of Variance (ANOVA)

- Generalisation of the t-test
- Can cope with more than 2 data sets
- For 2 sets, basically the same as t-test => use t-test
- Can cope with more independent variables with multiple levels
- Multivariate ANOVA for more than one dependent variable
- Excel: http://office.microsoft.com/en-au/excel/HP100908421033.aspx

> "The experiment used a repeated measures within-participant factorial design 3 x 2 x 3 (interaction technique x transfer type x task type)."
>
> "The independent variable interaction technique consisted of three levels: standard Bluetooth, touch & connect and touch & select."
>
> Khooviraj, Rukzio, Hardy, Holleis. MobileHCI'09

For Researchers / the Geeks ...

ANOVA: ANALYSIS OF VALUE

IS YOUR RESEARCH WORTH ANYTHING?

Developed in 1912 by geneticist R.A. Fisher, the Analysis of Value is a powerful statistical tool designed to test the significance of one's work.

am i wasting my time?

Significance is determined by comparing one's research with the **Dull Hypothesis**:

$$H_0 : \mu_1 = \mu_2 \ ?$$

where,

- H_0 : the Dull Hypothesis
- μ_1 : significance of your research
- μ_2 : significance of a monkey typing randomly on a typewriter in a forest where no one hears it.

WWW.PHDCOMICS.COM

JORGE CHAM © 2007

The test involves computation of the *F'd* ratio:

$$F'd = \frac{\text{sum(people who care about your research)}}{\text{world population}}$$

This ratio is compared to the F distribution with I-1, N_T degrees of freedom to determine a *p(in your pants)* value. A low *p(in your pants)* value means you're on to something good (though statistically improbable).

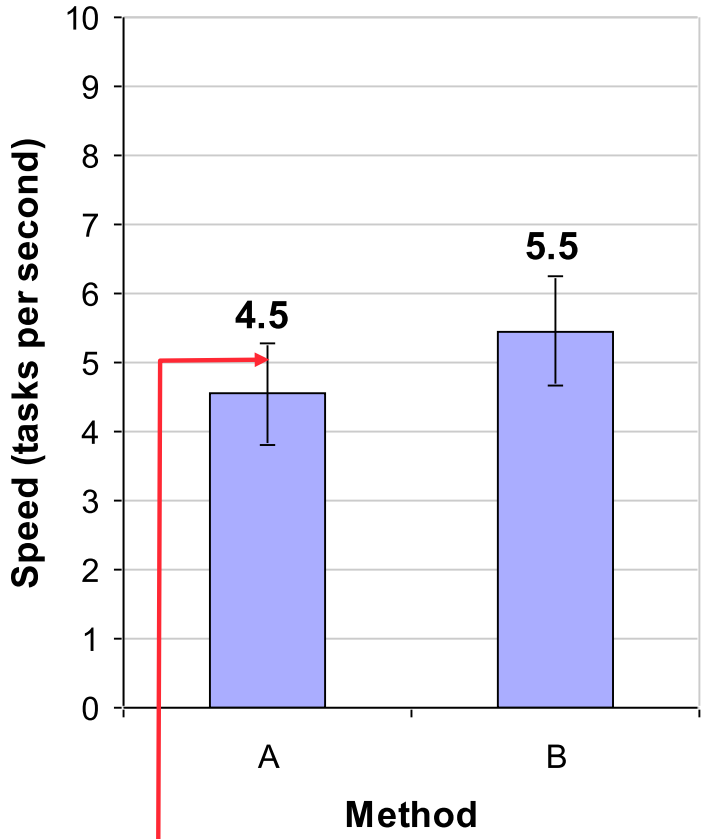
Type I/II Errors

The Analysis of Value must be used carefully to avoid the following two types of errors:

- Type I: You incorrectly believe your research is not Dull.
- Type II: No conclusions can be made. Good luck graduating.

Of course, this test assumes both Independence and Normality on your part, neither of which is likely true, which means *it's not your problem.*

Significant Example



Example #1		
Participant	Method	
	A	B
1	5,3	5,7
2	3,6	4,6
3	5,2	5,1
4	3,3	4,5
5	4,6	6,0
6	4,1	7,0
7	4,0	6,0
8	5,0	4,6
9	5,2	5,5
10	5,1	5,6
Mean	4,5	5,5
SD	0,73	0,78

Error bars show
±1 standard deviation

Source: MacKenzie, Empirical Research in HCI:What? Why? How?

Significant Example - Anova

ANOVA Table for Speed

	DF	Sum of Squares	Mean Square	F-Value	P-Value	Lambda	Power
Subject	9	5.839	.649				
Method	1	4.161	4.161	8.443	.0174	8.443	.741
Method * Subject	9	4.435	.493				

Probability that the difference in the means is due to chance

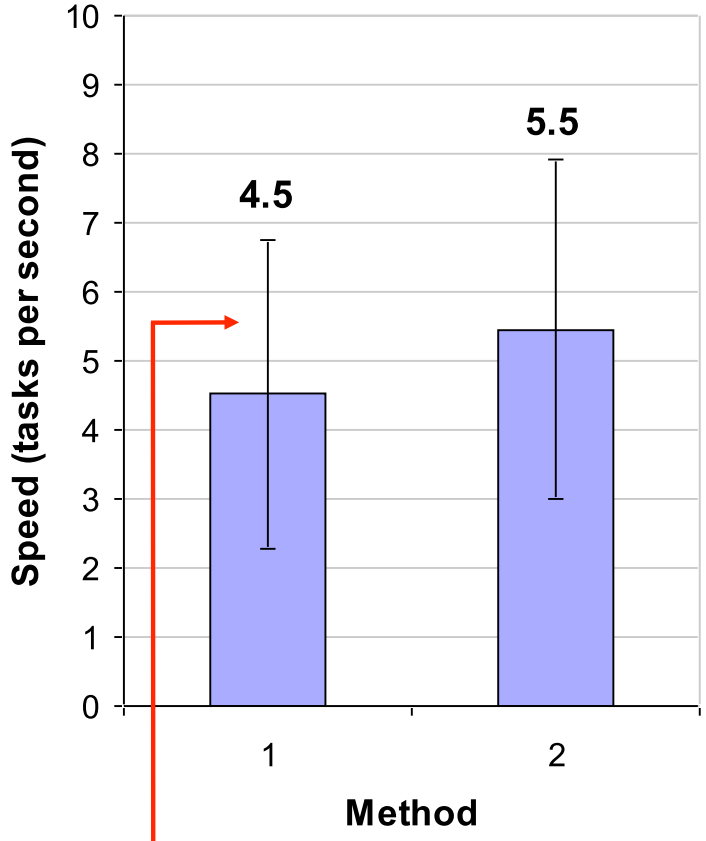
Reported as…

$F_{1,9} = 8.443, p < .05$

Thresholds for "p"
- **.05**
- .01
- .005
- .001
- .0005
- .0001

Source: MacKenzie, Empirical Research in HCI:What? Why? How?

Not Significant Example



Error bars show
±1 standard deviation

Example #2		
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

Source: MacKenzie, Empirical Research in HCI:What? Why? How?

Not Significant Example - Anova

ANOVA Table for Speed

	DF	Sum of Squares	Mean Square	F-Value	P-Value	Lambda	Power
Subject	9	37.017	4.113				
Method	1	4.376	4.376	.634	.4462	.634	.107
Method * Subject	9	62.079	6.898				

Probability that the difference in the means is due to chance

Reported as…

$F_{1,9} = 0.634$, ns

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

Source: MacKenzie, Empirical Research in HCI:What? Why? How?

ANOVA in Excel

http://office.microsoft.com/en-au/excel/HP100908421033.aspx: One-Way ANOVA

Anova: Single Factor

Which Bowler is Best?

SUMMARY

Groups	*Count*	*Sum*	***Average***	***Variance***
Pat	6	922	153.6667	92.26667
Mark	6	1070	178.3333	116.6667
Sheri	6	937	156.1667	54.96667

ANOVA

Source of Variation	*SS*	*df*	*MS*	***F***	*P-value*	***F crit***
Between Groups	2212.111	2	1106.056	**12.57358**	0.000621	**3.682317**
Within Groups	1319.5	15	87.96667			
Total	3531.611	17				

ANOVA test online: http://www.physics.csbsju.edu/stats/anova.html

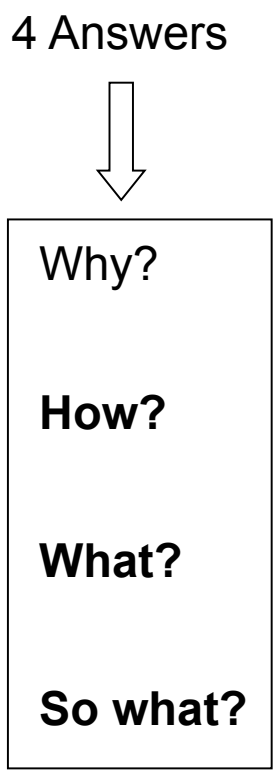
Overview Parametric and Non-Parametric

Experiment Design	Parametric Test	Non-Parametric Test
2 groups with different participants (one indep. variable)	Independent T-Test	Mann-Whitney Test
2 groups with same participants (one indep. variable)	Dependent T-Test	Wilcoxon Signed-Rank Test
≥ 3 levels groups with different participants and one indep. variable	One-way independent ANOVA	Kruskal-Wallis Test
≥ 3 levels groups with same participants and one indep. variable	One-way repeated measures ANOVA	Friedman‘s ANOVA
...	...	...

Reporting Study Results

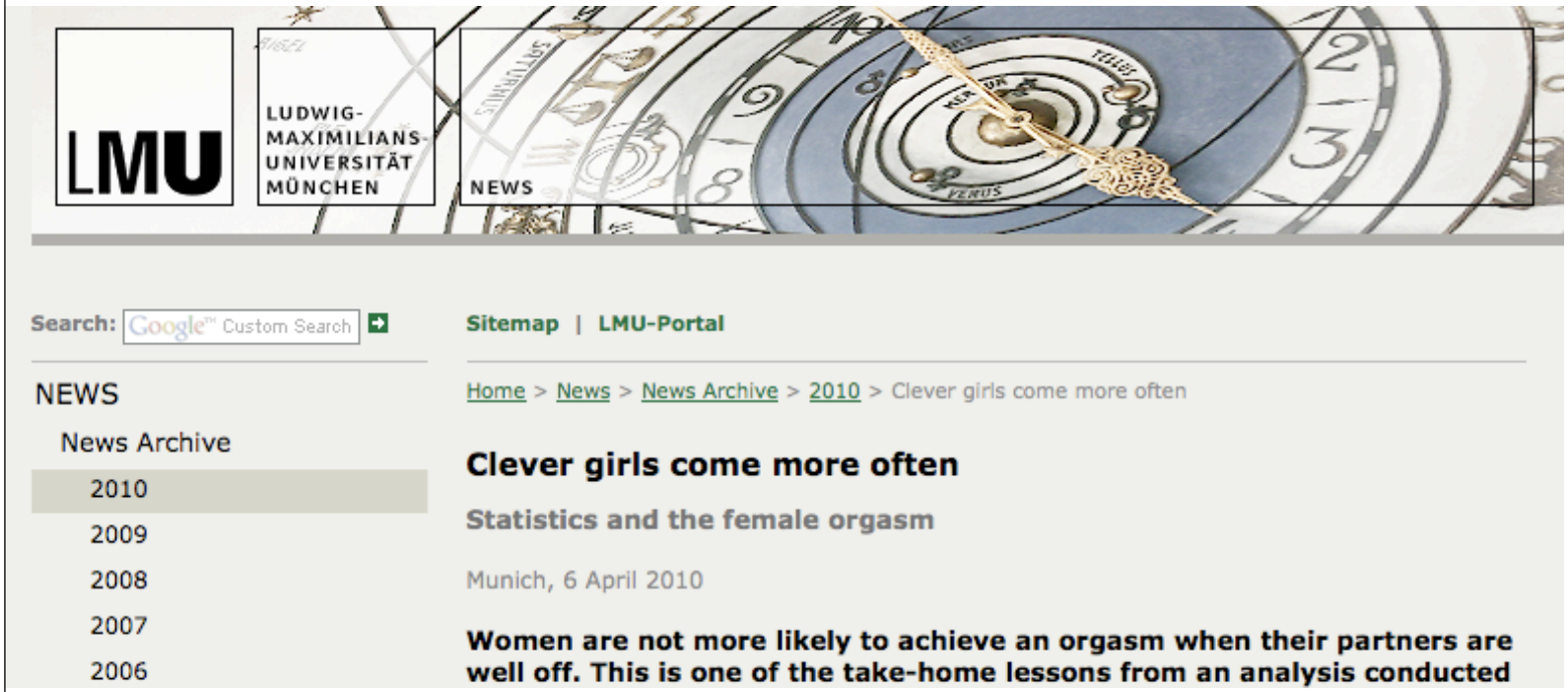
Sections of a report

1. Title
2. Abstract (brief summary of about 150 words)
3. Introduction (motivation)
   - Description of previous research
   - Rationale of your work
4. **Method**
   - **Overview of the study**
   - **Variables, levels, participants, procedure, ...**
5. **Results**
   - **What was scored?**
   - **Descriptive and inferential statistics**
6. **Discussion**
7. References
8. (Appendices)



Reporting Study Results

- Why it is important to tell HOW a conclusion was derived:



http://www.en.uni-muenchen.de/news/newsarchiv/2010/2010-hothorn.html

Women are not more likely to achieve an orgasm when their partners are well off. This is one of the take-home lessons from an analysis conducted by LMU researchers Professor Torsten Hothorn and Esther Herberich. The result clearly refutes the conclusion reached by a study that made headlines last year. Statistical analysis of the responses of more than 1500 Chinese women to a questionnaire on health and family life had led British and Dutch investigators to conclude that women were more likely to have orgasms when their male partners happened to be high earners. When Hothorn and Herberich re-evaluated the original data for teaching purposes, they discovered that the reported effect was actually an artefact caused by an error in the statistical software used to analyse the data. “Our analysis showed that the women's educational level in particular, but also general health and age, were associated with reported frequencies of orgasms” says Herberich. The LMU researchers have now published their results in a paper written together with the authors of the original study. “The primary study was actually based on data that are freely available”, remarks Hothorn. “Its ease of accessibility greatly enhances the scientific value of the original survey, because it allows statistical inferences to be independently checked by other interested groups, and either be confirmed or − as in this case − refuted”. (Evolution and Human Behavior online, March 2010)

This Lecture is not Enough!

- We strongly recommend to teach yourself.
  There is plenty of material on the WWW.

- Further Literature:
  - Andy Field & Graham Hole: How to design and report experiments, Sage
  - Jürgen Bortz: Statistik für Sozialwissenschaftler, Springer
  - Christel Weiß: Basiswissen Medizinische Statistik, Springer
  - Lothar Sachs, Jürgen Hedderich: Angewandte Statistik, Springer
  - Various books by Edward R. Tufte
  - ... and many more ...

References

- Carmines, E. and Zeller, R. (1979). Reliability and Validity Assessment. Newbury Park: Sage Publications
- Colosi, L (1997) The Layman's Guide to Social Research Methods http://www.socialresearchmethods.net/tutorial/Colosi/lcolosi1.htm
- Field, A. and Hole, G. (2003). How to Design and Report Experiments. Sage Publications