# Mensch-Maschine-Interaktion 1

Chapter 5:

User Study Design + Statistics

Slides based on material by Sara Streng + Paul Holleis

# User Study Design + Statistics

- 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

# The Purpose of User Studies

What are user studies needed for?

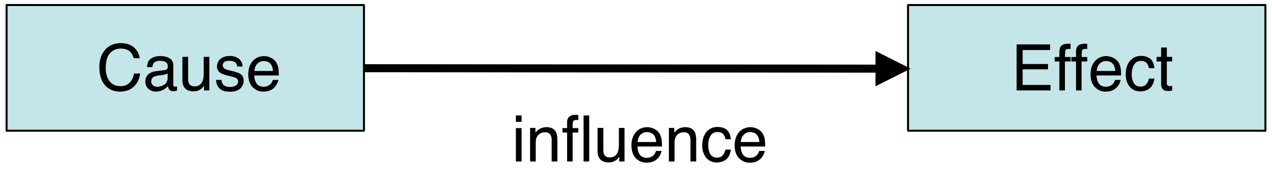
- "To learn more"
- To ensure quality in product development
- To compare solutions
- To provide quantitative figures
- To get a scientific statement (instead of personal opinion)

Examples of scientific statements

- Users are quicker using version A than using version B
- Users make 10% less errors when using version X than when using version Y
- 90% of the users can complete the transaction using version Y in less than 3 minutes
- On average users will be able to buy a ticket using version A in less than 30 seconds

# Cause and Effect

- Why do scientists measure things?
  ⇨ Find causal links between variables, e.g. smoking ⇨ cancer



- Criteria that need to be met to infer cause and effect (Mill):
  1. Cause has to precede effect
  2. Cause and effect should correlate
  3. All other explanations of the cause-effect relationship must be ruled out

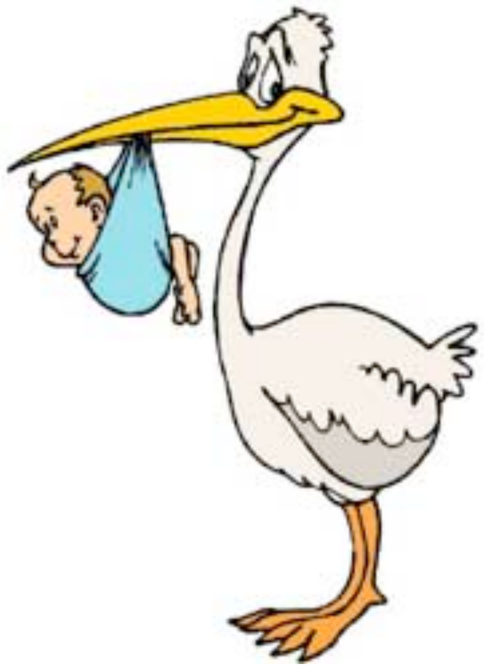
- Only way to infer causality:
  - Two controlled situations
    1. Cause is present (*experimental condition*)
    2. Cause is absent (*control condition*)
  - Otherwise the situations have to be identical!

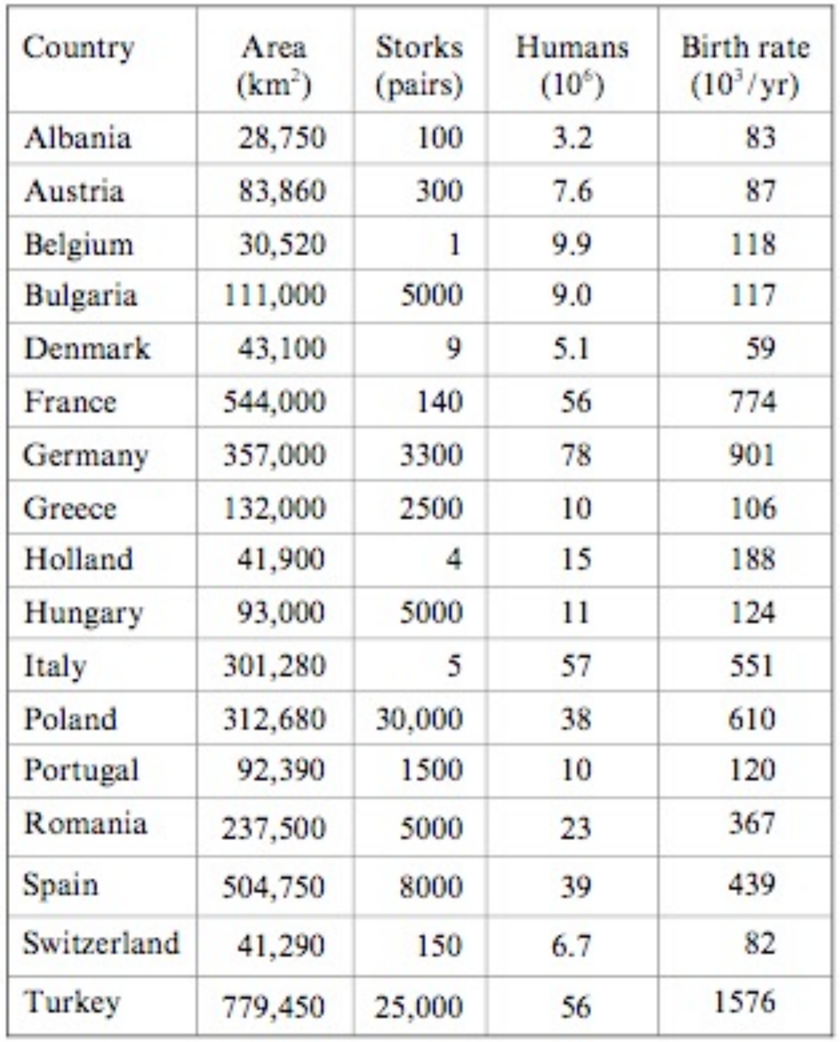
# Storks Deliver Babies?!

- R. Matthews, “Storks Deliver Babies”. Journal of Teaching Statistics, vol. 22, issue 2, pages 36-38, 2001; http://www3.interscience.wiley.com/journal/119039912/abstract/

- There is a correlation coefficient of r=0.62 (reasonably high)
- A statistical test can be employed that shows that this correlation is in fact significant (p = 0.008)

- What are the flaws?



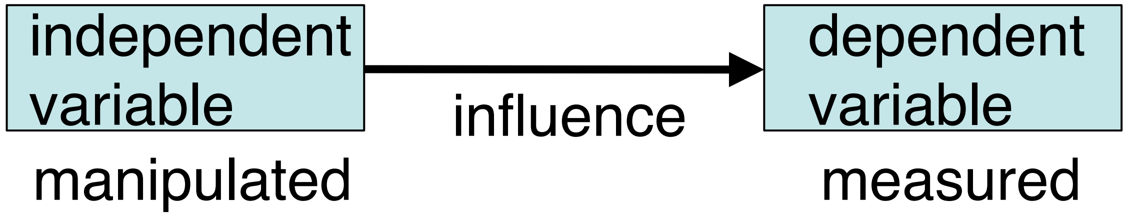


| Country | Area ($km^2$) | Storks (pairs) | Humans ($10^6$) | Birth rate ($10^3$/yr) |
|---|---|---|---|---|
| Albania | 28,750 | 100 | 3.2 | 83 |
| Austria | 83,860 | 300 | 7.6 | 87 |
| Belgium | 30,520 | 1 | 9.9 | 118 |
| Bulgaria | 111,000 | 5000 | 9.0 | 117 |
| Denmark | 43,100 | 9 | 5.1 | 59 |
| France | 544,000 | 140 | 56 | 774 |
| Germany | 357,000 | 3300 | 78 | 901 |
| Greece | 132,000 | 2500 | 10 | 106 |
| Holland | 41,900 | 4 | 15 | 188 |
| Hungary | 93,000 | 5000 | 11 | 124 |
| Italy | 301,280 | 5 | 57 | 551 |
| Poland | 312,680 | 30,000 | 38 | 610 |
| Portugal | 92,390 | 1500 | 10 | 120 |
| Romania | 237,500 | 5000 | 23 | 367 |
| Spain | 504,750 | 8000 | 39 | 439 |
| Switzerland | 41,290 | 150 | 6.7 | 82 |
| Turkey | 779,450 | 25,000 | 56 | 1576 |

**Table 1.** Geographic, human and stork data for 17 European countries

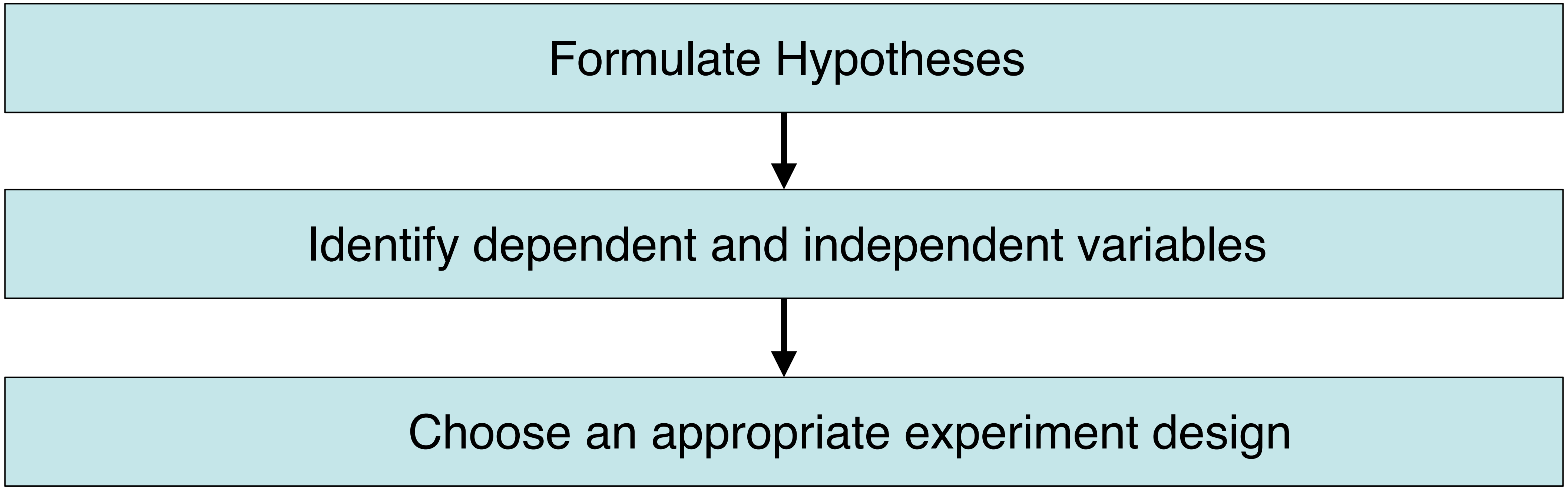
# Independent vs. Dependent Variables

- **Independent** variables
  - Manipulated by the experimenter
  - Conditions under which the tasks are performed
  - The number of different values used is called ***level***, e.g.
    - Font can be *Arial* or *Times* (2 levels)
    - Background can be blue, green, or white (3 levels)
- **Dependent** variables
  - Affected by the independent variables
  - Measured in the user study
  - Objective values: e.g. time to complete a task, number of errors, etc.
  - Subjective values: ease of use, preferred option
  - They should only depend on the independent variables (conditions)



# Hypothesis

- Prediction of the result
- States how a change in the independent variables will effect the measured dependent variables
- By doing an experiment, the hypothesis is either proved or disproved
- **Null hypothesis** predicts that independent variables do not have any effect on the dependent variables
- Formulate hypotheses BEFORE running the study!



Formulate Hypotheses

Identify dependent and independent variables

Choose an appropriate experiment design

# How to Isolate the Cause

1. Control conditions
2. Controlling other factors
   ⇨ Minimize the risk of other factors influencing the experiment
3. Randomizing allocation of participants to experimental and control group, Example: Instruction Manual
   - RQ: Does reading a manual help to use a device (e.g. a mobile phone) more efficiently?
   - Conditions:
     1. Experimental condition: Participants read the manual
     2. Control condition: Participants do not read the manual
   - About half the participants own this device. Imagine all of those would be allocated to the experimental condition, the other ones to the control condition. What happens?

# User Study Design + Statistics

- 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

# Aims of Research

The results of your experiment should be

1. Valid
   - ⇨Results should be accurate
   - ⇨Results should show what you intend to show
2. Reliable
   - ⇨Results should be potentially replicable by anyone
3. Generalizable
   - ⇨Results should have a wider application than the particular circumstances of the experiment
4. Important

In order to be (potentially) important the results need to fulfill the first three criteria!

# Reliability

- Consistency of measurement: Degree to which an instrument measures the same way each time it is used under the same condition with the same subjects
- A measure is reliable if a person‘s score on the same test given twice is similar.
- Two ways of estimating reliability:

  1. Test/Retest
     - Conservative method
     - Two separate times of measurement
     - Compute correlation between the two measurements
     - Assuming the conditions are the same
  2. Internal Consistency
     - Group questionnaire items that measure the same concept
       e.g. two sets of questions that both measure motivation
     - Compute correlation between the two sets
     - **Cronbach‘s Alpha**: split all questions every possible way and compute correlations for all of them ⇨ correlation coefficient

# Maximizing Reliability

- Precise, unambiguous and objective definition of what is being measured.
- Not always easy!
  - Easy examples:
    - Memory ⇨ # items recalled
  - Hard example: measuring effect of frustration on children‘s agression
- Solutions
  - Definition by consensus
    - Find candidates for aggressive activities (e.g. through observations)
    - Independent judges rate aggression of activities
  - Operational definition
    - Experimenter defines aggressive behavior as X, Y, Z for the purpose of this study
    - Whether one agrees to the definition or not, at least the results are true for X, Y, Z

# Validity

- Concerns the relationship between concept and indicator
  - –Measurements show what they are intended to show
- Internal validity
  - –Measurements are accurate
  - –Measurements are due to manipulations, not caused by other factors
  - –Precondition:
    - »Good experimental design
- External validity
  - –Findings are representative of humanity
  - –Not only valid in experiment setting
  - –Precondition:
    - »Good judgement and sometimes intuition

# Example: Brain Weight

- Paul Broca investigated human abilities / intelligence by measuring brain weight (19th century)
- Findings:
  - Brain of Caucasian men > Brain of Caucasian women > Brain of negroes
  - Brain of French men > Brain of German men
- Is brain weight a true score for intelligence?
  - No, because it is known that within all species there is no relationship between brain weight and intelligence
- What other things does brain weight reflect?
  - Relation to body size
  - Age (mainly elderly females and young males, who died in car accidents)

Reliable? ☑

Valid? ☒

# Example: Folding Rule

- A folding rule is only 1.9 m instead of 2 m
- Every time it is used to determine the length of an object, it systematically overestimates the length.

Reliable? ☑

Valid? ☒

# Threats to Internal Validity (1)

- Group threats
  - If experimental and control group are different the study is worthless
- Instrument change, e.g.
  - Different measuring devices
  - Interviewer gets more practiced
- Reactivity and experimenter effects
  - Measuring a person‘s behavior might already change the behavior
  - Social desirability
  - Ideally: Double-blind technique (participant and experimenter unaware of hypotheses and conditions)
- Differential Mortality
  - When testing the same individuals repeatedly
  - E.g. pre-test is not comparable to post-test when many participants drop out

# Threats to Internal Validity (2)

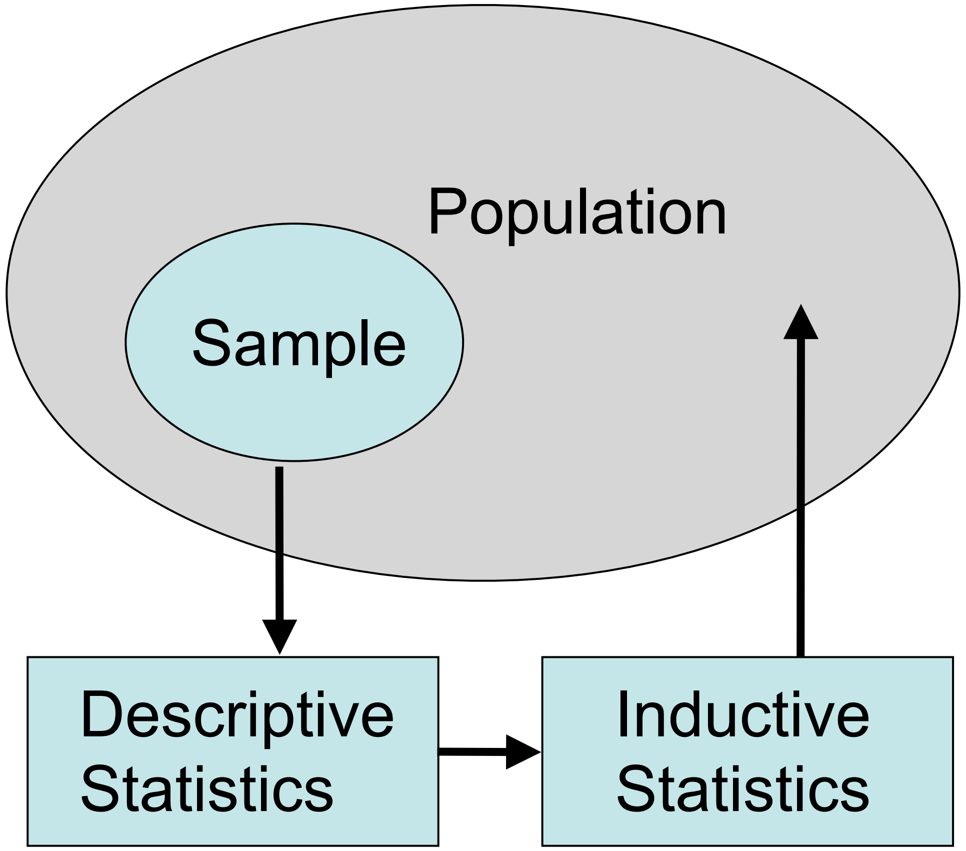
- Regression to the mean
  - If extreme scores were produced on a pre-test, it is more likely that the score is closer to the mean on a subsequent test
  - Problem always occurs when measuring the effect of a problem solution / policy
- Time threats
  - Maturation, e.g. children‘s reading ability
  - Influence of events unrelated to the manipulation that occurred during the treatment, i.e. between pre- and post-test

# Threats to External Validity

- Over-use of special participant groups
  - McNemar 1946: „psychology is largely a study of undergraduate behavior“
  - 70-90% of participants are undergraduates (Rosenthal and Rosnow, 1975)
  - Today: how valid are experiments that are done with Media Informatics students only?
- Restricted numbers of participants
  - Typical threat to reliability
  - Also affects the ability to generalize

# Generalizability

- What do we want to gain from a user study?
  - Result, which is valid for all people
- Test users must be representative
- Descriptive statistics:
  - Tables
  - Diagrams
  - Means
  - ...
- Inductive statistics:
  - Ensure validity for the whole



# Quality of Study Design

- Well designed experiments isolate causal factors well
- Poorly designed experiments leave many alternative explanations of the results ⇨ practically useless
- Data consists of four components:
  1. A „true score“ for the things we hope to measure — maximize
  2. A „score for other things“ that are measured inadvertently — minimize
  3. Systematic (non-random) bias — minimize
     – Should (if at all) affect all participants in the study
  4. Random (non-systematic) error — minimize
     – Should be cancelled out over large numbers of observations

# User Study Design + Statistics

- 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

# Experimental vs. Observational Methods

Two approaches to answering research questions (RQ)

1. **Observational** (= correlational) methods:
   Observe what naturally happens in the environment without interfering
2. **Experimental** methods:
   Manipulate some aspects and observe the effects

| | Experimental | Observational |
|---|---|---|
| Pros | • Isolate and control variables ⇨ allow causal statements | • Natural setting: observe how people behave normally |
| Cons | • Danger of artificial situations ⇨ people might behave differently | • Variables are not isolated<br>• Time consuming |

Compromise:

- Verify causal hypotheses ⇨ confirm findings with more natural observations or
- Identity hypotheses through observations ⇨ verify hypotheses in experiments

# Quasi-Experimental Method

1. Observational
    - No manipulation
    - Record behavior systematically and objectively
    - Strength: observe people how they behave normally (e.g. driving behavior)
    - Downside:
        - No identification of cause and effect
        - Time consuming
2. Quasi-experimenal
    - Sometimes real experiments are not possible (e.g. for ethical reasons)
    - Control over timing of measurement
    - No (complete) control over independent variables
      ⇨ Impossible to isolate cause and effect
3. Experiment
    - Manipulation by experimenter
    - Only way to prove cause and effect

# Quasi-Experiment - Example: Motorcyclists

- RQ: Does daytime headlight use make motorcyclists more detectable?
- Dependent variable: number of accidents
- Experimental design:
  - Randomly allocate large group of motorcyclists to two groups
    - » Experimental group uses headlight during daytime
    - » Control group does not use headlight during daytime
  - Ethical reasons against this allocation!
- Solution: Quasi-experimental design:
  - Find motorcyclists with different preferences
  - Pre-existing difference (⇨ group threat): Other factors related to the preference for/against headlights can influence the dependent variable, e.g.
    - » Older machines
    - » Different safety-consciousness levels
    - » ...

# Experiments on Age- and Gender-Differences

- E.g. is there an age-difference in problem-solving ability?
- Most researchers investigate effects of age and gender as „true“ experiments
- Strictly speaking, they are quasi-experiments:
  - –Participants are not randomly allocated to the groups
  - –Impossible to rule out other reasons than age or gender difference, such as
    - »Born at different times
    - »Different life experience
    - »...

⇨ Be aware of the complications in interpreting the results!

# Types of Experimental Designs

1. Within subjects („repeated measures“)
   - Each subjects is exposed to all conditions
   - Randomize the order of conditions to avoid ordering affects
2. Between groups (“independent measures”)
   - Separate groups of participants for each conditions
   - Careful selection of groups is essential
3. Hybrid (“mixed”) designs
   - Combination of between-group and within-subject variables

| | Pros | Cons |
|---|---|---|
| Within subjects | • Fewer participants required (n) | • Carry-over (learning) effects<br>• Sometimes impossible (e.g. gender) |
| Between groups | • No carry-over effects<br>• Less fatigue | • More participants required (n * [number of conditions])<br>• Usually harder to show significance |

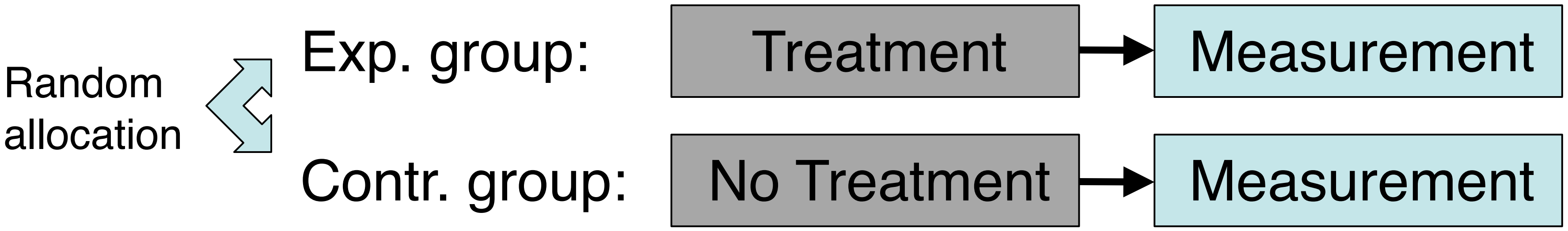
# The Importance of Randomization

- In all types of experiments randomization is crucial:
  - In within-subject designs ⇨ order of conditions
  - In between-group designs ⇨ allocation to groups
- If you fail to randomize your results can not be interpreted
- Example (between groups): Milk experiment in the 1930s
  - Huge and expensive experiment with 20 000 school children
  - Examine nutritious effects of milk
  - Teachers „randomly“ assigned children to
    - » Experimental group (received milk every day)
    - » Control group (received no milk)
  - Teachers subconsciously tended to assign poor children to the experimental group
  - Result:
    - » Control group were by far superior in weight and height
    - » The whole study was worthless due to the lack of randomization

# Types of Between Group Designs (1)

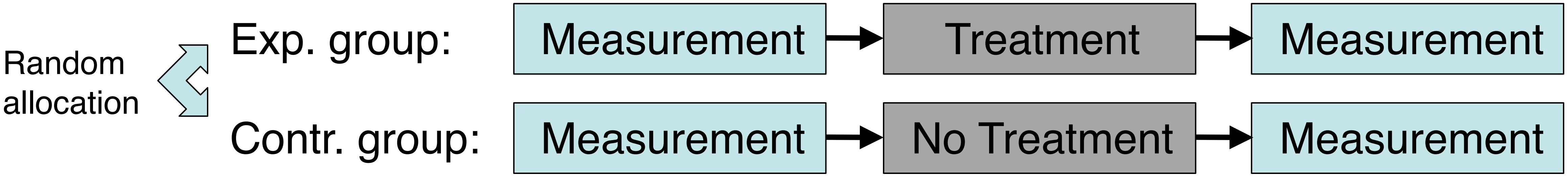
Objective: randomized group allocation ⇨ avoid group threats

1. Post-test only control group design



- Weakness: no way of knowing if randomization fails to produce equivalence

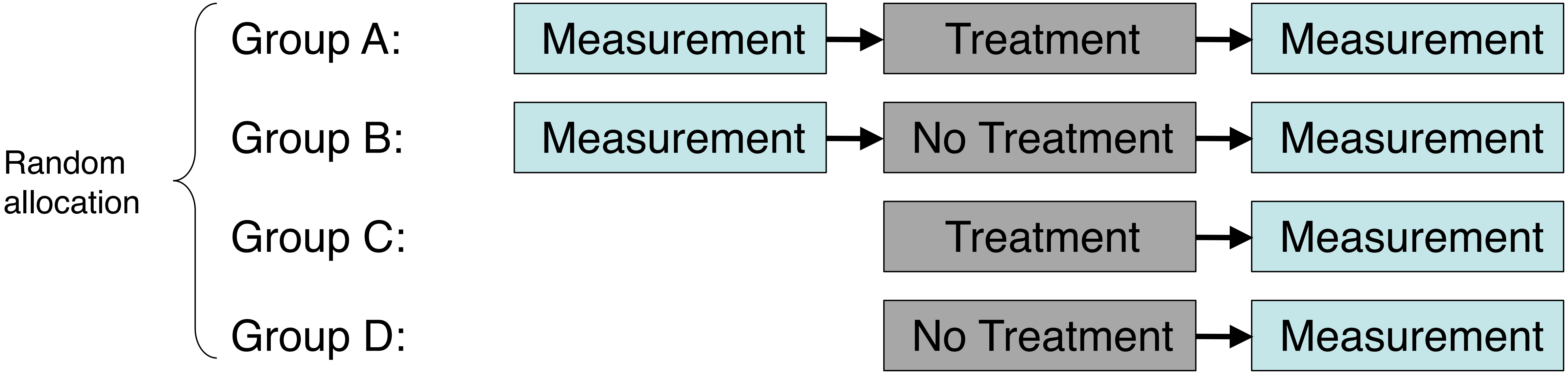
2. Pre-test / post-test control group design



- Pre-test guarantees equivalence
- Weakness: pre-test might affect the performance

# Types of Between Group Designs (2)

3. Solomon four-group design



- Two experimental groups (A and C)
- Two control groups (B and D)
- Groups A and B show the effects of presence/absence of the treatment
- Groups C and D show the effects of the pre-testing
- Very expensive in time and # participants ⇨ rarely used

# Types of Within Subject Designs

- Objective: random / counterbalanced order of conditions
- Trivial for 2 conditions:
  Half of the participants start with condition A, the other half with condition B
- For more than 2 conditions:
  - Randomize order
  - Systematically counterbalance the order (Latin Square Design):
    - » There are n! different orders for n conditions
    - » Instead of running n! different orders (= groups), only use n and still avoid order effects
    - » Idea: Every condition is placed at each „position“ once
    - » Each order is used by one of the n groups of participants
    - » Weakness:
      - Unbalanced for odd numbers of conditions
        e.g. n = 3: A before B twice, B before A once

n = 3 ⇨ ABC, BCA, CAB
n = 4 ⇨ ABCD, BADC,
CDAB, DCBA

# Multi-Factorial Designs

- All designs covered so far: manipulation of only 1 variable

Note: Do not confuse

- – Multiple levels of one variable (e.g. different doses of a drug) with
- – Multiple variables (e.g. (1) different drugs taken at (2) different times of the day)

- Advantage of using multiple variables:
  - – Analyze how multiple variables interact
  - – Not much extra work in within subject designs (only more task(s))
- Disadvantage:
  - – Experiments with more than 2 - 3 variables are difficult to interpret!
  - – Much extra work in between group designs (#groups multiplies)
- Number of experimental conditions = product of the variables' levels, e.g.
  - – Font can be Arial or Times (2 levels)
  - – Background can be blue, green, or white (3 levels)
  - – => 6 experimental conditions

# How to study: Example: Questionnaires

- Use:
  - Getting information on users' subjective satisfaction
  - Information on possible anxieties, issues, etc.
  - Find out about system usage, (dis-)likability of specific features
  - Distribution to many people, can be filled in offline
  - (Parts) can be re-used
- Problems:
  - People need not tell the truth
  - Answer can be incorrect (e.g. people often rather say what they should have done instead of what really happened)
  - People can easily deny answering it
  - Can easily be too long, too complex to understand, not give expected results
  - Questions can be leading
  - Open questions can provide much data but are more difficult to analyse
- Consider alternatives
  - interviews ('live questionnaires'); system logging; observations; focus groups; ...

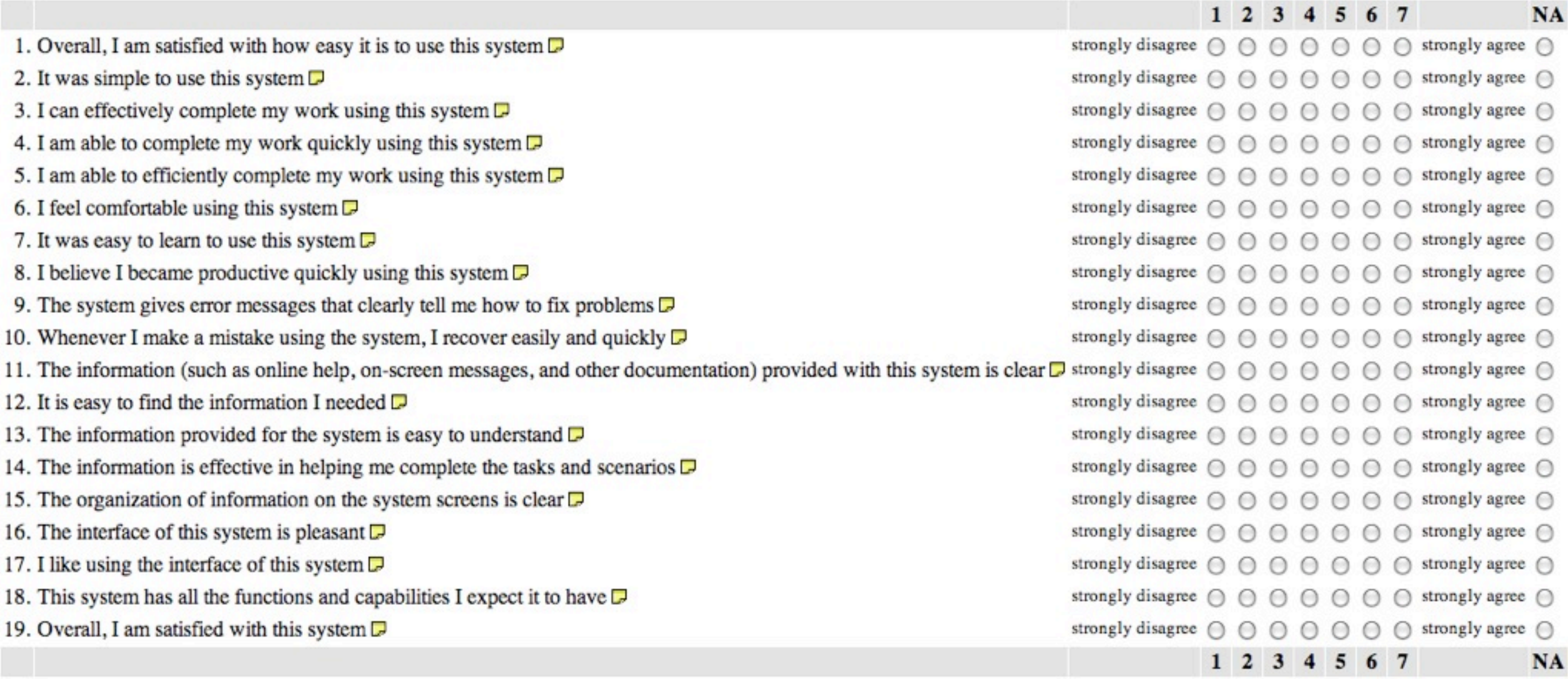
# How to study: Example: Questionnaires

- Things to keep in mind:
  - Only include questions whose answer is relevant (know what to do with the data in advance)!
  - Make it short, clear, and understandable (test it before distribution)!
  - Include open questions (only) if necessary (e.g. ask for exceptional events etc.)
  - Allow quick, short answers, e.g. one number or use scales or checklists

- Examples:
  - (see next page) Lewis, J. R. 1995. IBM computer usability satisfaction questionnaires: psychometric evaluation and instructions for use. Int. J. Hum.-Comput. Interact. 7, 1 (Jan. 1995), 57-78. DOI= http://dx.doi.org/10.1080/10447319509526110
  - QUIS: The Questionnaire for User Interaction Satisfaction (not freely available)
    It contains a **demographic questionnaire**, a measure of **overall system satisfaction** along six scales, and hierarchically organized measures of nine **specific interface factors** (screen factors, terminology and system feedback, learning factors, system capabilities, technical manuals, on-line tutorials, multimedia, teleconferencing, and software installation). Each area measures the users' overall satisfaction with that facet of the interface, as well as the factors that make up that facet, on a 9-point scale.

# IBM Computer Usability Satisfaction Questionnaire

- General set of questions
  - Can be applied to a wide area of interfaces / devices / systems
  - But gives mostly general results



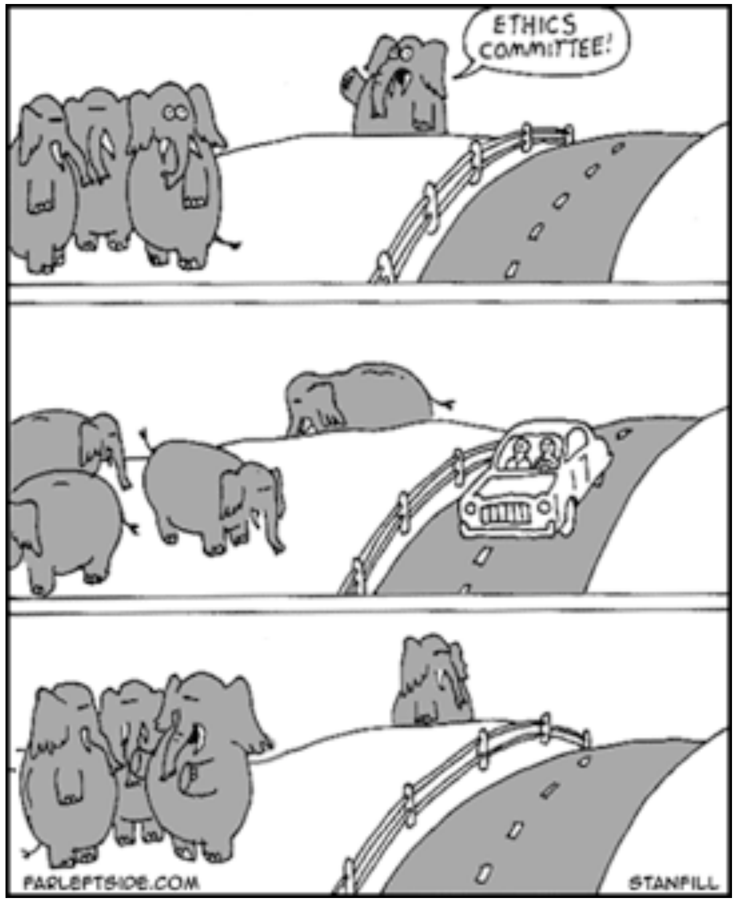
| | | 1 | 2 | 3 | 4 | 5 | 6 | 7 | | NA |
|---|---|---|---|---|---|---|---|---|---|---|
| 1. Overall, I am satisfied with how easy it is to use this system | strongly disagree | ○ | ○ | ○ | ○ | ○ | ○ | ○ | strongly agree | ○ |
| 2. It was simple to use this system | strongly disagree | ○ | ○ | ○ | ○ | ○ | ○ | ○ | strongly agree | ○ |
| 3. I can effectively complete my work using this system | strongly disagree | ○ | ○ | ○ | ○ | ○ | ○ | ○ | strongly agree | ○ |
| 4. I am able to complete my work quickly using this system | strongly disagree | ○ | ○ | ○ | ○ | ○ | ○ | ○ | strongly agree | ○ |
| 5. I am able to efficiently complete my work using this system | strongly disagree | ○ | ○ | ○ | ○ | ○ | ○ | ○ | strongly agree | ○ |
| 6. I feel comfortable using this system | strongly disagree | ○ | ○ | ○ | ○ | ○ | ○ | ○ | strongly agree | ○ |
| 7. It was easy to learn to use this system | strongly disagree | ○ | ○ | ○ | ○ | ○ | ○ | ○ | strongly agree | ○ |
| 8. I believe I became productive quickly using this system | strongly disagree | ○ | ○ | ○ | ○ | ○ | ○ | ○ | strongly agree | ○ |
| 9. The system gives error messages that clearly tell me how to fix problems | strongly disagree | ○ | ○ | ○ | ○ | ○ | ○ | ○ | strongly agree | ○ |
| 10. Whenever I make a mistake using the system, I recover easily and quickly | strongly disagree | ○ | ○ | ○ | ○ | ○ | ○ | ○ | strongly agree | ○ |
| 11. The information (such as online help, on-screen messages, and other documentation) provided with this system is clear | strongly disagree | ○ | ○ | ○ | ○ | ○ | ○ | ○ | strongly agree | ○ |
| 12. It is easy to find the information I needed | strongly disagree | ○ | ○ | ○ | ○ | ○ | ○ | ○ | strongly agree | ○ |
| 13. The information provided for the system is easy to understand | strongly disagree | ○ | ○ | ○ | ○ | ○ | ○ | ○ | strongly agree | ○ |
| 14. The information is effective in helping me complete the tasks and scenarios | strongly disagree | ○ | ○ | ○ | ○ | ○ | ○ | ○ | strongly agree | ○ |
| 15. The organization of information on the system screens is clear | strongly disagree | ○ | ○ | ○ | ○ | ○ | ○ | ○ | strongly agree | ○ |
| 16. The interface of this system is pleasant | strongly disagree | ○ | ○ | ○ | ○ | ○ | ○ | ○ | strongly agree | ○ |
| 17. I like using the interface of this system | strongly disagree | ○ | ○ | ○ | ○ | ○ | ○ | ○ | strongly agree | ○ |
| 18. This system has all the functions and capabilities I expect it to have | strongly disagree | ○ | ○ | ○ | ○ | ○ | ○ | ○ | strongly agree | ○ |
| 19. Overall, I am satisfied with this system | strongly disagree | ○ | ○ | ○ | ○ | ○ | ○ | ○ | strongly agree | ○ |
| | | 1 | 2 | 3 | 4 | 5 | 6 | 7 | | NA |

# User Study Design + Statistics

- 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

# Ethical Considerations

- Be aware of the influence and power of the experimenter
- Responsibility to the participants!
- Some research institutions have an ethics committee, which examine details of your proposed study before you can run the experiment.
- If not, you should still follow some guidelines:
  - – Protect the participants‘ confidentiality
  - – Protection from physical and psychological risks (of psychological or medical experiments)
  - – Informed Consensus: Inform participants about:
    - » The experiment (in particular risks)
    - » Their rights (in particular withdrawal from the study)
    - » Confidentiality
  - – Inform participants, that the system is evaluated - not the user.
    - » If something does not work, it is never the user‘s fault!
  - – Debriefing: Tell participants what the study was about in the end



# Informing the Participants About the Study

**Inform the participant about:**

- General purpose of the study
- Procedure
  - Amount of time
  - Breaks
  - ...
- Their right to withdraw from the study at any time
- Confidentiality
- Risks
- The system is evaluated - not the user:
  Interest is in aggregated data of all participants, not in the individual ones!

**Never reveal:**

- Hypotheses
- Conditions

# Example Consent Form

- **Participants Consent Form**

- **Study** ______________________________ **Institution** __________________________
- Name: _________________________________ Date of Birth: ____________________
- Email: ___________________________________________________________________
- Phone:__________________________________

- I have been informed on the procedure and purpose of the study and my questions have been answered to my satisfaction.
- I have volunteered to take part in this study and agree that during the study information is recorded (audio and video as well as my interaction with the system). This information may only be used for research and teaching purpose.
- I understand that my participation in this study is confidential. All personal information and individual results will not be released to third parties without my written consent.
- I understand that I can withdraw from participation in the study at any time.

- Date: ____________________ Signature:_____________________________________

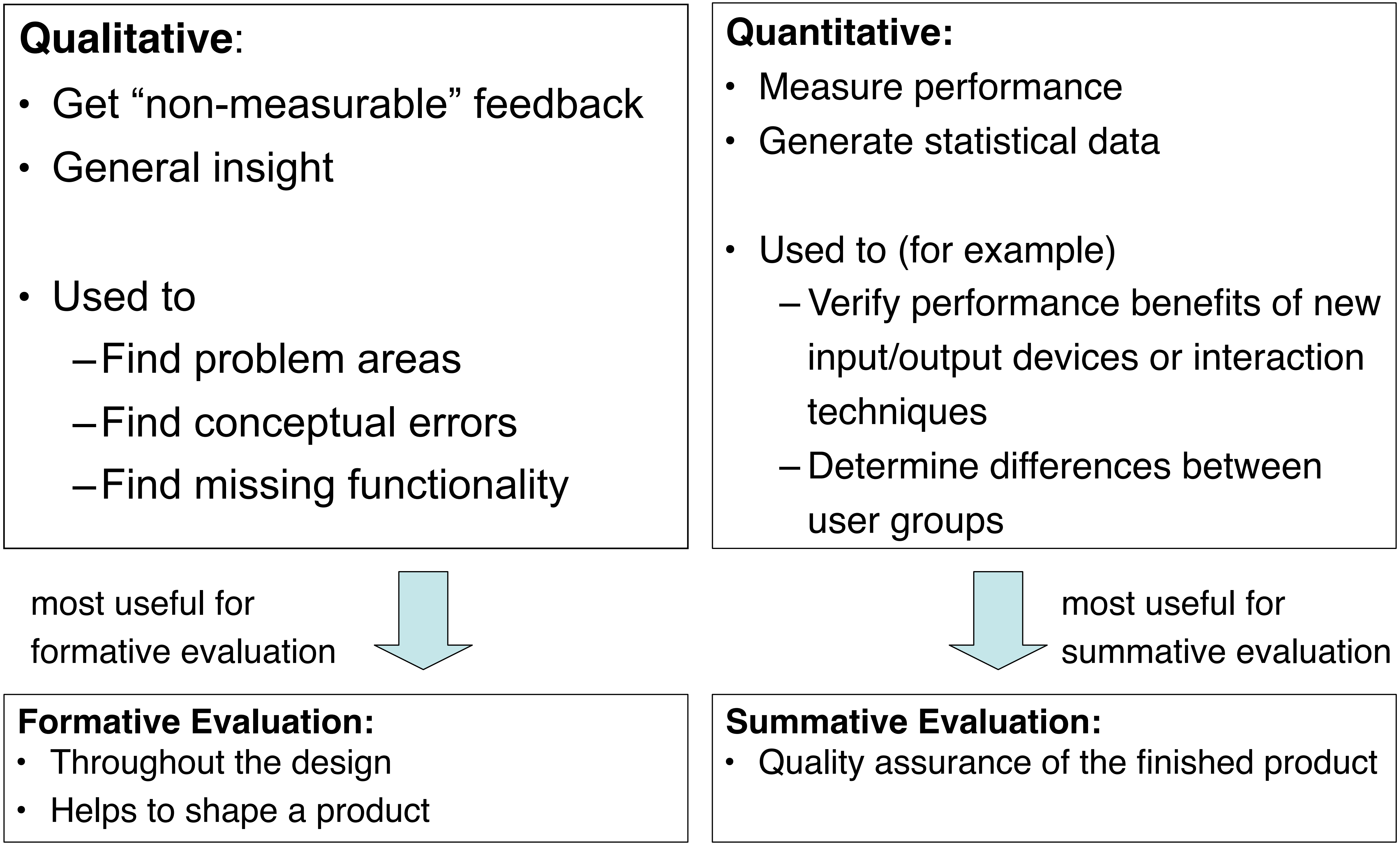
# User Study Design + Statistics

- 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

# What is Evaluated in HCI Research?

- Depends on the stage of a project:
  - Ideas and concepts
  - Designs
  - Prototypes
  - Implementations
  - Products in use
- Differentiate between assessing learnability or interaction
  ⇨ train the user before the tasks?
- Approaches
  - Formative evaluation
    - Throughout the design
    - Helps to shape a product
  - Summative evaluation
    - Quality assurance of the finished product

# Qualitative vs. Quantitative User Studies



**Qualitative**:

- Get “non-measurable” feedback
- General insight

- Used to
  - Find problem areas
  - Find conceptual errors
  - Find missing functionality

**Quantitative:**

- Measure performance
- Generate statistical data

- Used to (for example)
  - Verify performance benefits of new input/output devices or interaction techniques
  - Determine differences between user groups

most useful for formative evaluation

most useful for summative evaluation

**Formative Evaluation:**

- Throughout the design
- Helps to shape a product

**Summative Evaluation:**

- Quality assurance of the finished product

# Procedure

1. Set goals (hypotheses)
2. Design the experiment
3. Do a pilot study
4. Recruit users
5. For each user, typically:
   - Inform the user about the experiment
   - Consent form
   - Do a survey on
     - » Demographics
     - » Questions related to the experiment (e.g. left- / right-handedness)
   - Give instructions on the task
   - Let the user do the tasks and measure the variables
   - Be available for questions and (informal) feedback
6. Analyze the results ⇨ accept / reject hypotheses

# Recruiting and Participants

- The number of subjects needed depends on
  - Project
  - Goals
  - Setup

  Minimal size is about 10 subjects
- Participants should be representative for the user group
  - Age
  - Background (e.g. technical vs. not technical)
  - Skills
  - Experience
  - …

  In most cases your team members are NOT representative!

# Specification of the Experiment Setup

The experiment should be set up to be reproducible

⇨ write a specification describing everything which is necessary for reproducing the experiment:

- Hard- and software in use
- Detailed description of self-built prototypes
- The environmental conditions
  - –Light conditions
  - –Atmosphere
  - –...
- Skills of the test users, e.g.
  - –„All participants have to be professional designers”
  - –“The candidates should have no experience on using eye-trackers”
  - –…

# What You Should Keep in Mind

- Don’t learn how to conduct the experiment during the user study. Think about what to do in case of problems in advance, e.g. how to proceed if the mobile phone of a user gets an incoming call during a test run?
  - Stop the recording and repeat the test run?
  - Stop the test and don’t use the data?
- Times can be recorded automatically by the testing software or stopped manually with a watch.
- Allow and plan enough time
- Allow and plan for people to exercise their right to leave any time

# Example User Study Design - Variables

- Imagine you want to compare different mobile phone input methods:
  - T9 vs. Multi-Tap (2 conditions)

- Dependent variables?
  - Time
  - # Errors
- Independent variables?
  - Input method: 2 levels: Multi-tap and T9
  - Text to input: 1 level: text with about 10 words

# Example User Study Design - Hypotheses

- Hypotheses

  H-1: Input by multi-tap is quicker than T9

  H-2: fewer errors are made using multi-tap input compared to T9

- Null-Hypotheses

  H0-1: No difference in the input speed between multi-tap and T9

  H0-2: No difference in the number of errors between multi-tap input and T9

- Experimental Method
  - > Within subjects
  - > Randomized order of conditions
  - > Users 1, 3, 5, 7, 9 and 11 perform T9 then Multi-tap
  - > Users 2, 4, 6, 8, 10 and 12 perform Multi-tap then T9

# Example User Study Design - Other Aspects

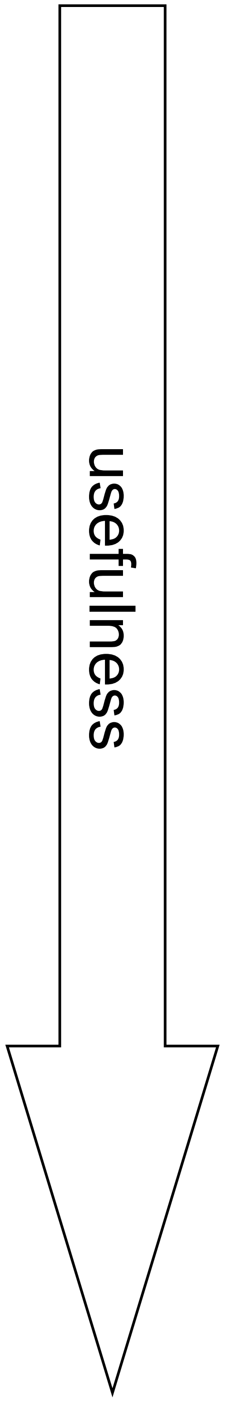
- Different texts in first and second run?
  - Variable “text” would have two levels
    ⇨ 4 experimental conditions:
    - » Users 1, 5 and 9 perform T9/Text1 then Multi-tab/Text2
    - » Users 3, 7 and 11 perform T9/Text2 then Multi-tab /Text1
    - » Users 2, 6 and 10 perform Multi-tab/Text1 then T9/Text2
    - » Users 4, 8 and 12 perform Multi-tab/Text2 then T9/Text1
- Particular phone model?
- How to measure
  - Completion time (e.g. stop watch or application?)
  - Number of errors/corrections observed
- Participants
  - How many?
  - Skills
  - Computer user, Phone/T9 users?

# User Study Design + Statistics

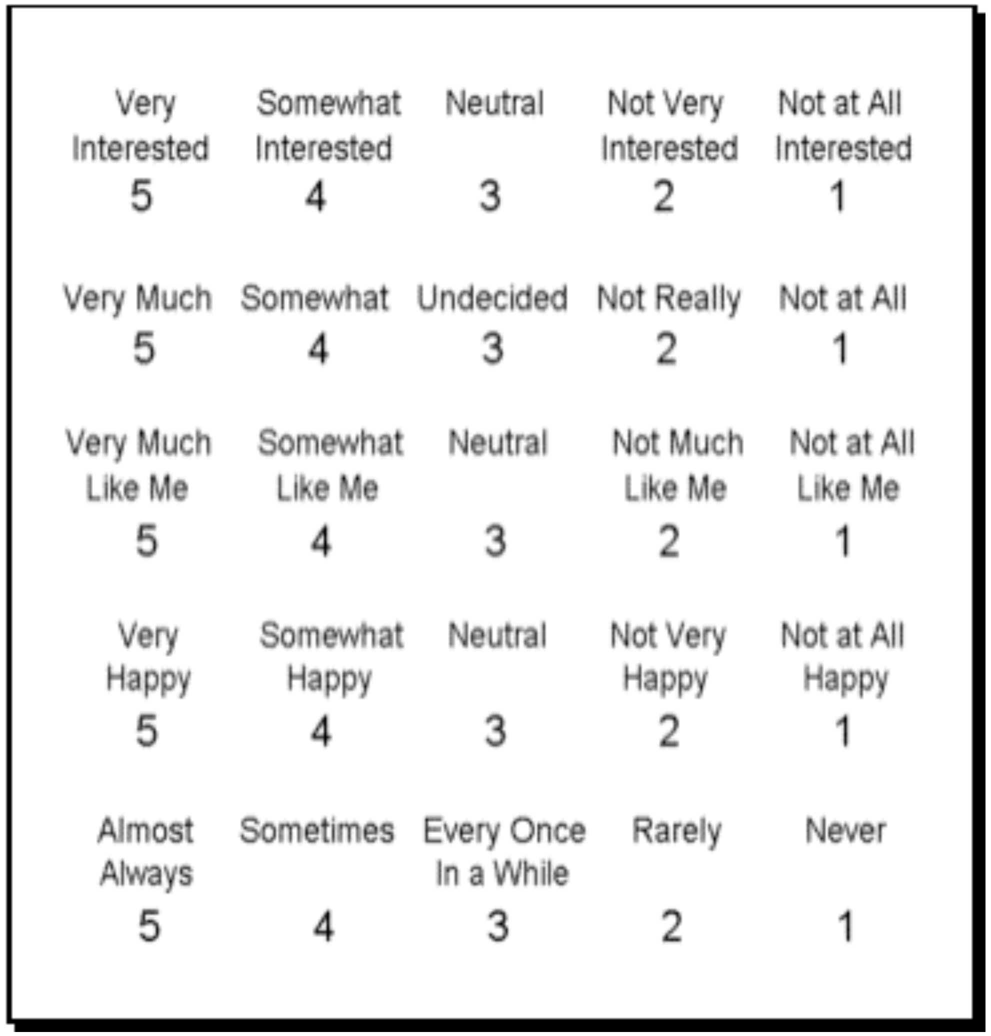
- The Purpose of User Studies
- Research Aims: Reliability, Validity and Generalizability
- Research Methods and Experimental Designs
- Ethical Considerations
- HCI-related and practical information for your own studies
- Interpretation of Data and Presentation of Results

# 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



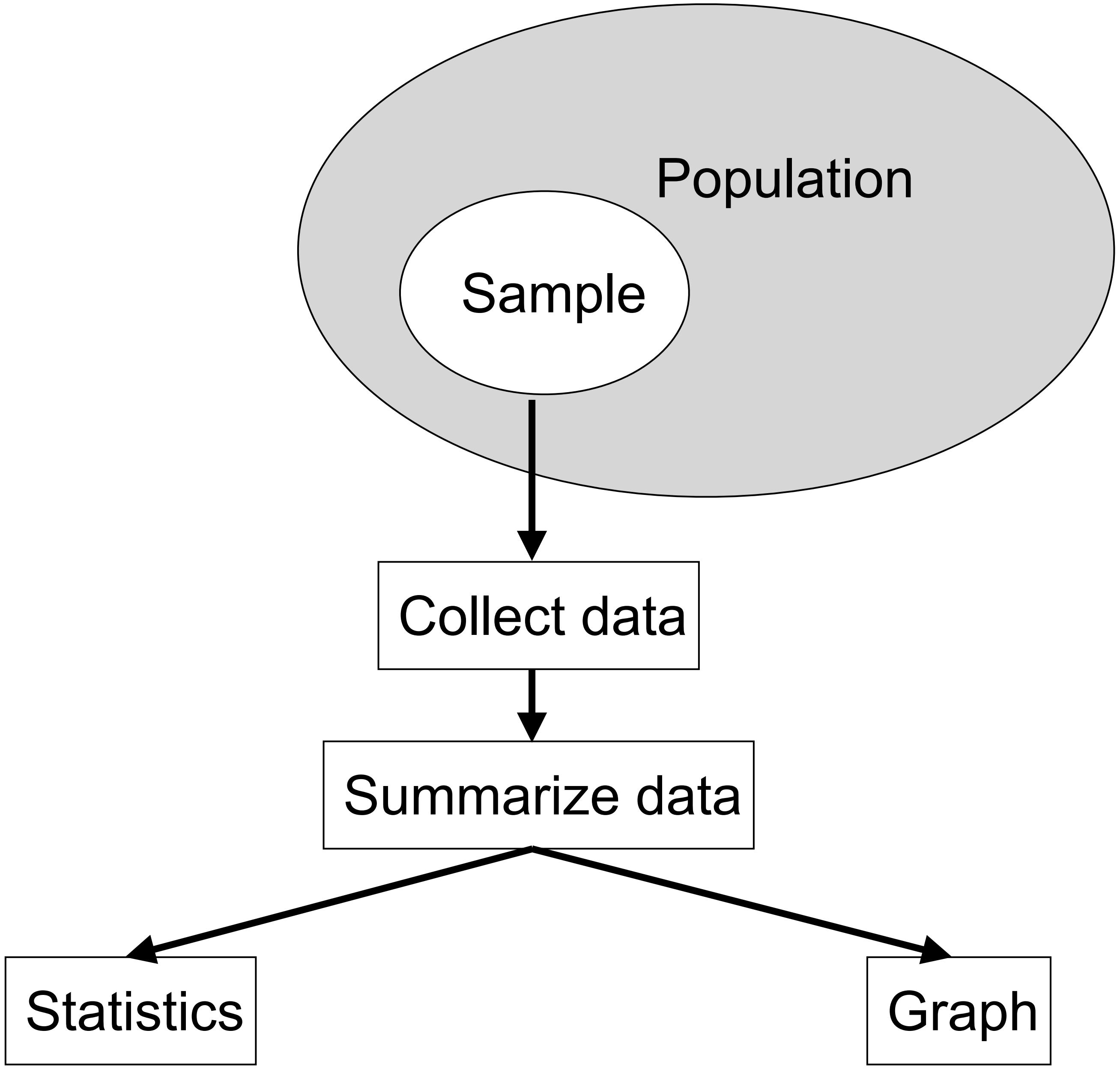
| | | | | |
|---|---|---|---|---|
| Very Interested | Somewhat Interested | Neutral | Not Very Interested | Not at All Interested |
| 5 | 4 | 3 | 2 | 1 |
| Very Much | Somewhat | Undecided | Not Really | Not at All |
| 5 | 4 | 3 | 2 | 1 |
| Very Much Like Me | Somewhat Like Me | Neutral | Not Much Like Me | Not at All Like Me |
| 5 | 4 | 3 | 2 | 1 |
| Very Happy | Somewhat Happy | Neutral | Not Very Happy | Not at All Happy |
| 5 | 4 | 3 | 2 | 1 |
| Almost Always | Sometimes | Every Once In a While | Rarely | Never |
| 5 | 4 | 3 | 2 | 1 |

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

- 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, ...

# Summarizing Data

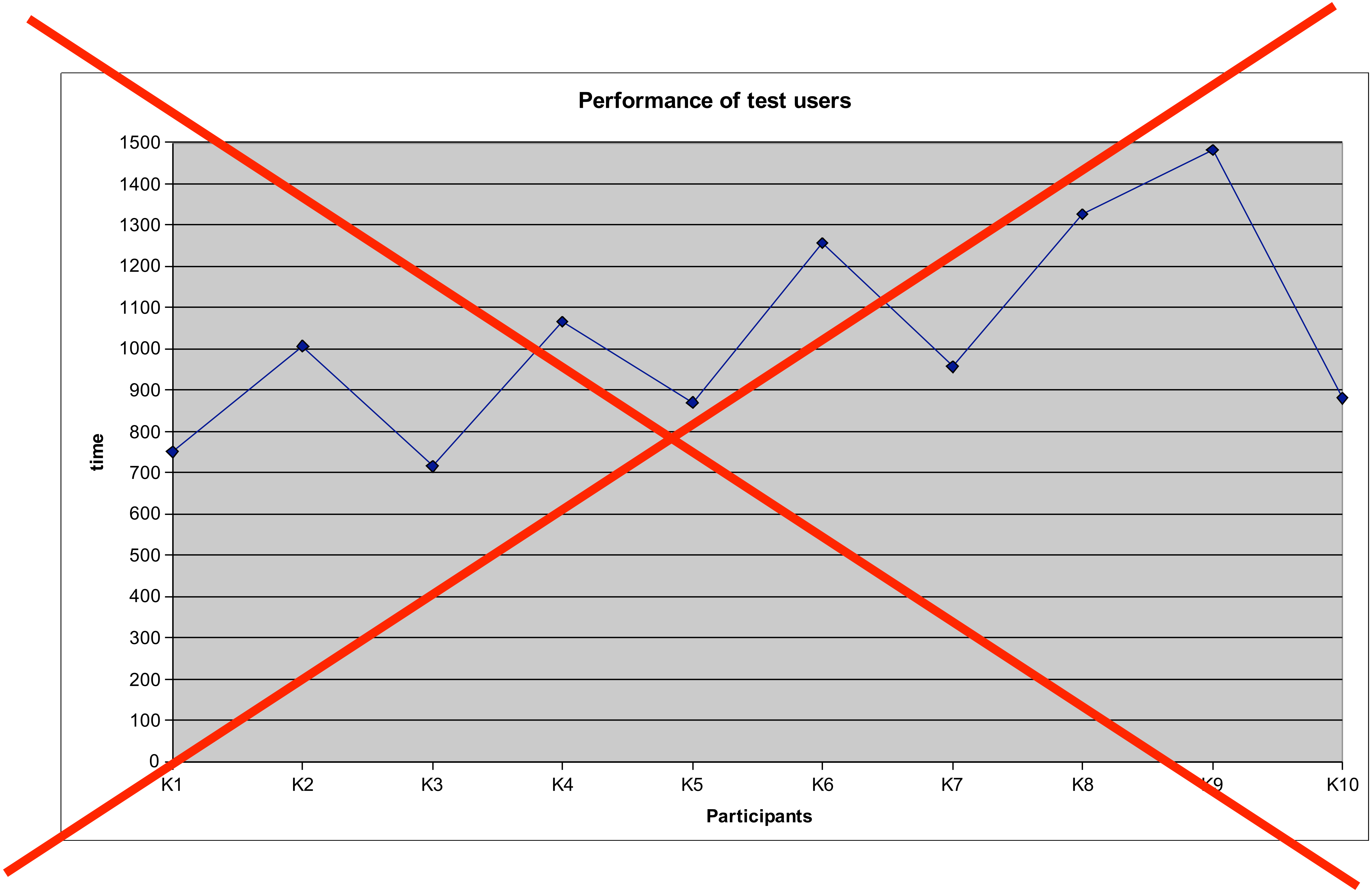
- Collected data needs to be summarized
  - Recognize patterns
  - Aggregate data
- Two ways:
  - Statistics
  - Graph



(e.g. mean, median, mode)

(e.. frequency distribution)

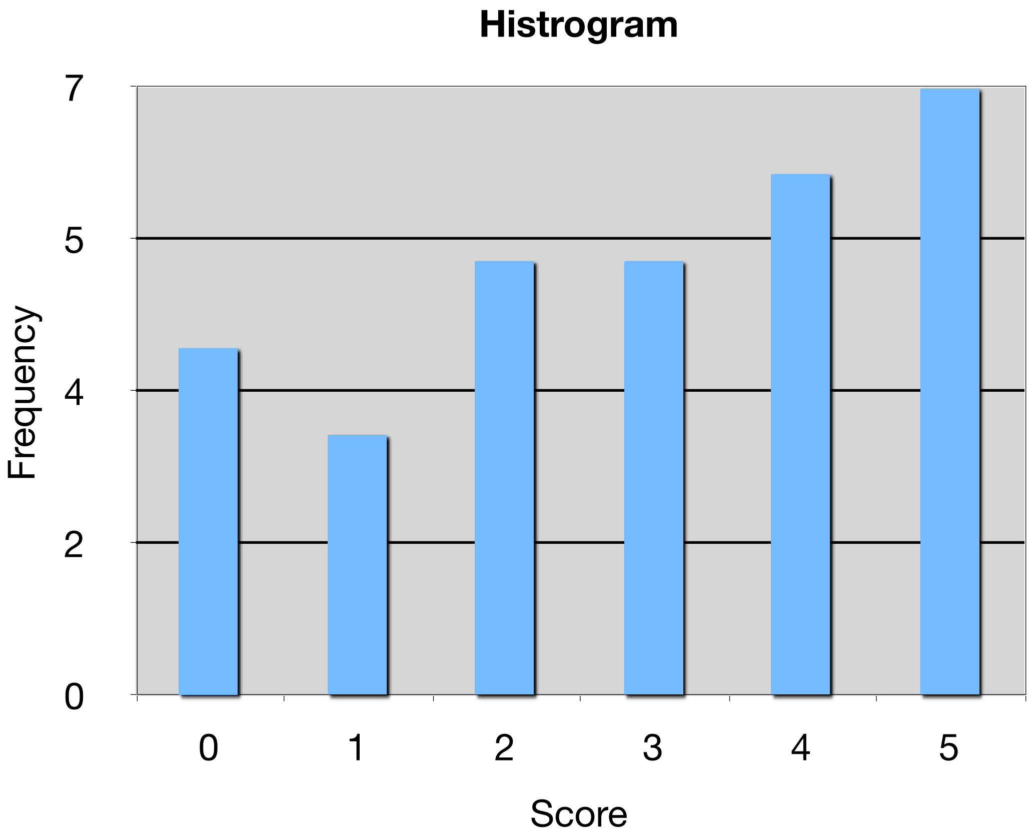
# Don’t Do This


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:

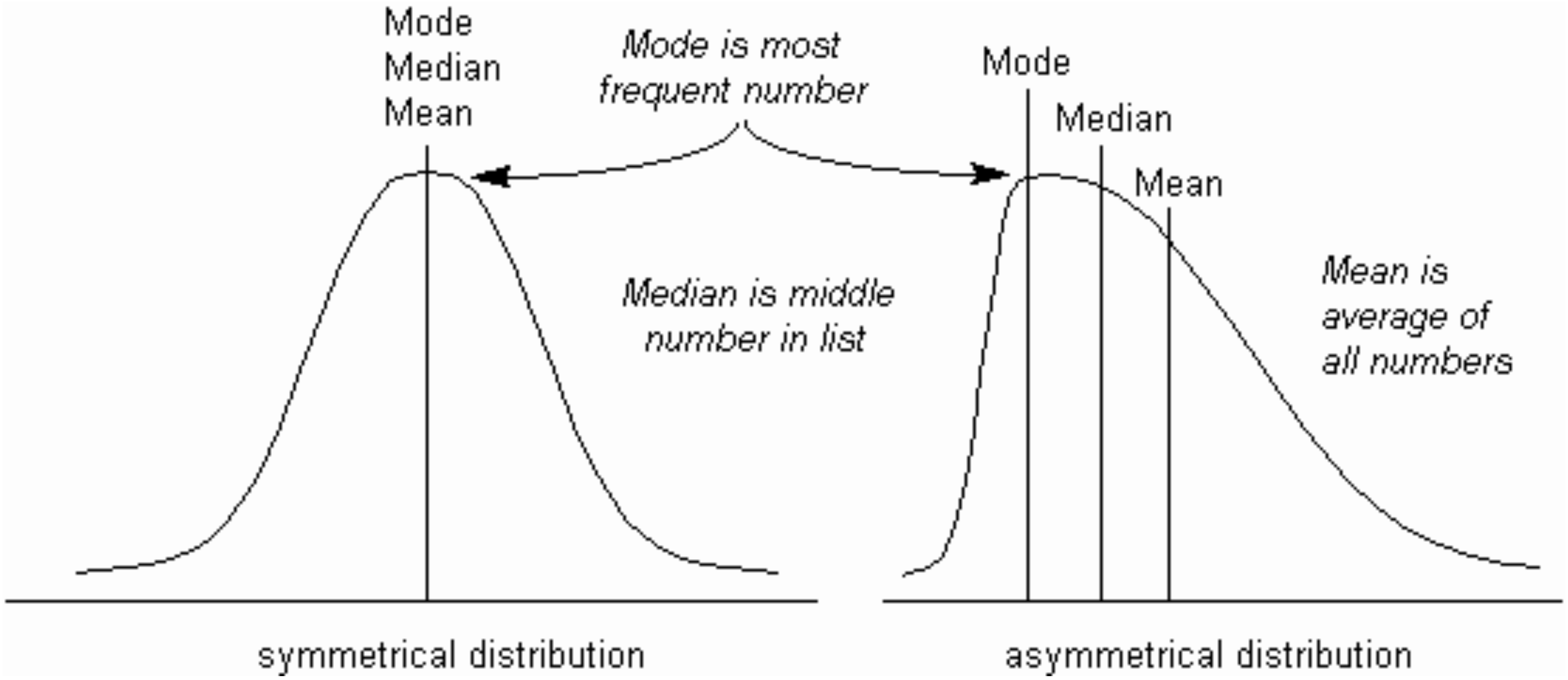
| Days | Frequency | Frequency (%) |
|---|---|---|
| 0 | 4 | 13% |
| 1 | 3 | 10% |
| 2 | 5 | 17% |
| 3 | 5 | 17% |
| 4 | 6 | 20% |
| 5 | 7 | 23% |



# 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
Median
Mean
Mode is most frequent number
Mode
Median
Mean
Median is middle number in list
Mean is average of all numbers
symmetrical distribution
asymmetrical distribution

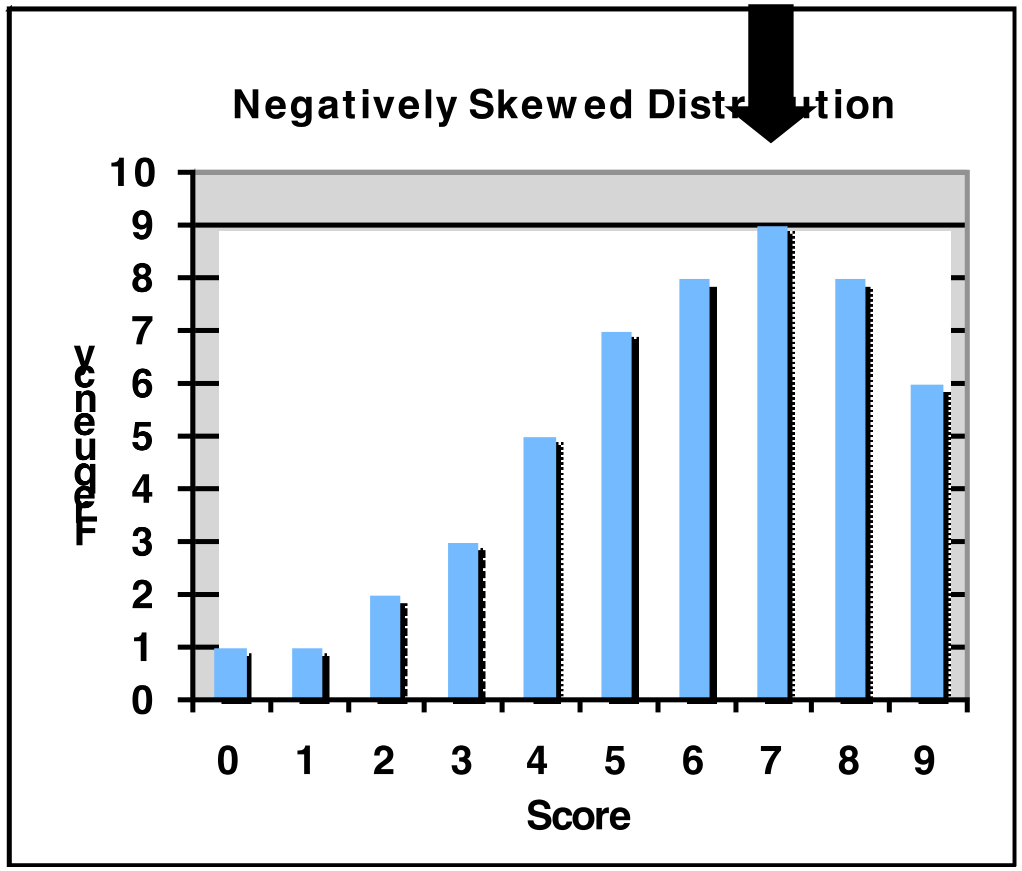
http://www.syque.com/quality_tools/toolbook/Variation/measuring_centering.htm

# 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

mode



# 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

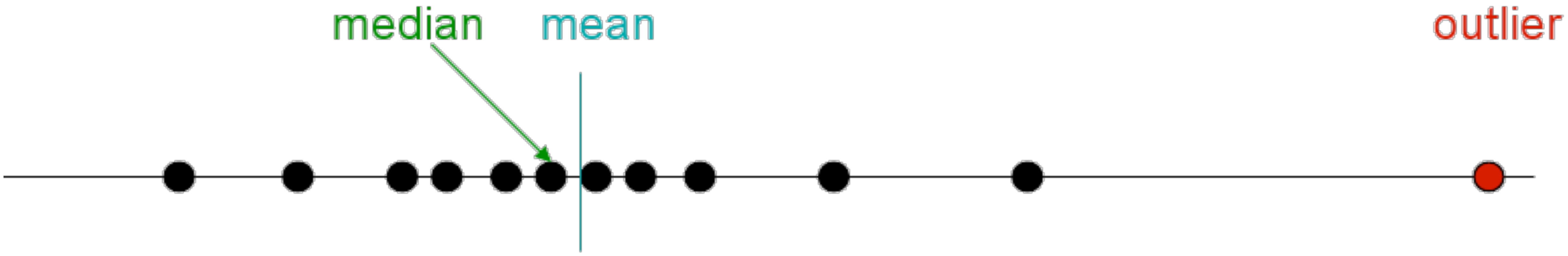
if n is odd: $x_{(n+1)/2}$

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

# Mean (M)

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

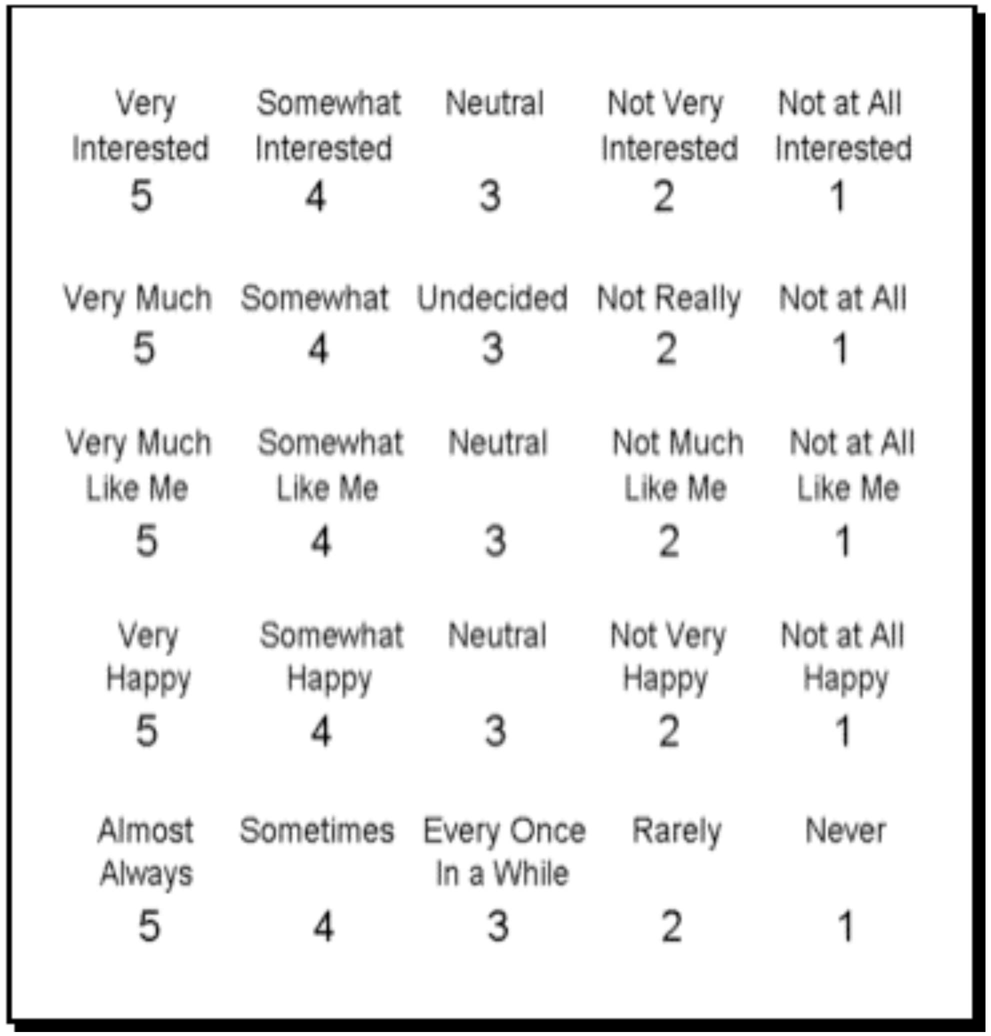
- Sum of all scores divided by #scores:
- Most often used if ‘average’ is mentioned
- Pros:
  - – Considers every score
    - ⇨ most accurate summary of the data
  - – Resistant to sampling variation: removing one sample changes the mean far less than mode or median
- Cons:
  - – Heavily affected by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



# 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



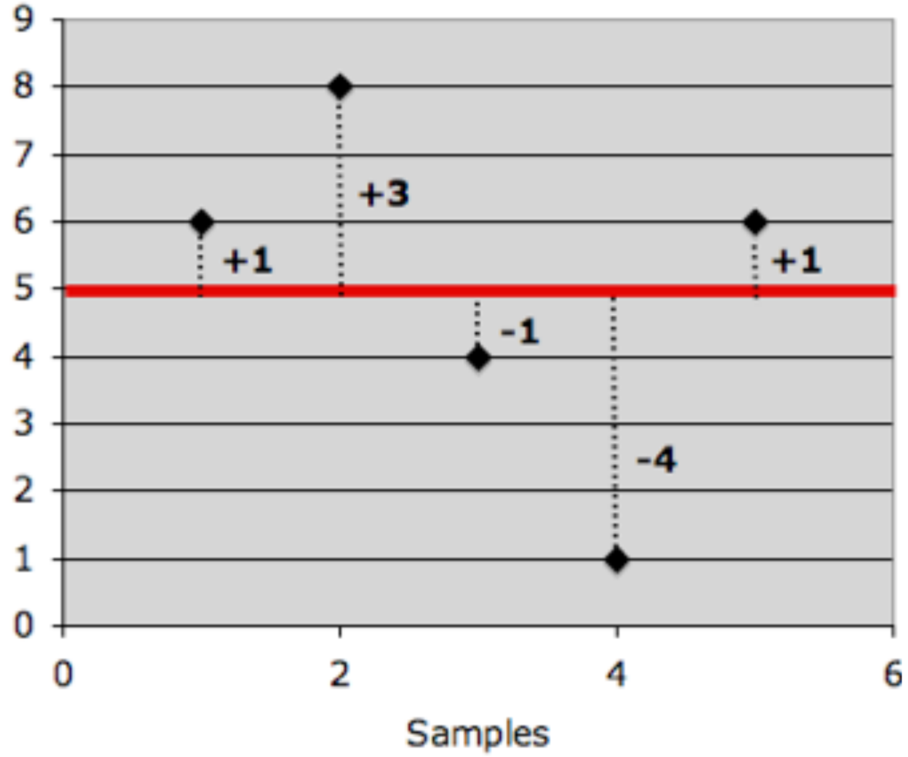
| | | | | |
|---|---|---|---|---|
| Very Interested | Somewhat Interested | Neutral | Not Very Interested | Not at All Interested |
| 5 | 4 | 3 | 2 | 1 |
| Very Much | Somewhat | Undecided | Not Really | Not at All |
| 5 | 4 | 3 | 2 | 1 |
| Very Much Like Me | Somewhat Like Me | Neutral | Not Much Like Me | Not at All Like Me |
| 5 | 4 | 3 | 2 | 1 |
| Very Happy | Somewhat Happy | Neutral | Not Very Happy | Not at All Happy |
| 5 | 4 | 3 | 2 | 1 |
| Almost Always | Sometimes | Every Once In a While | Rarely | Never |
| 5 | 4 | 3 | 2 | 1 |

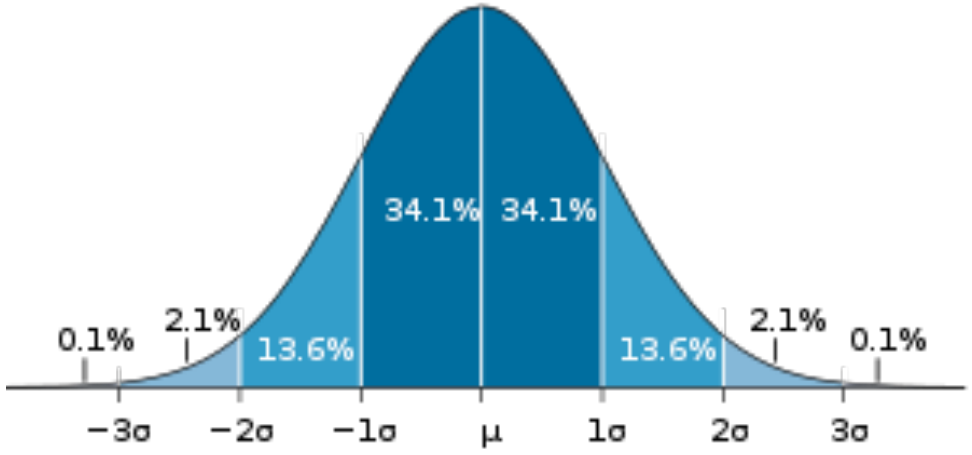
# Standard Deviation and Variance

- 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 $x_1$, $x_2$, … $x_n$ 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 = √Var(X)

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

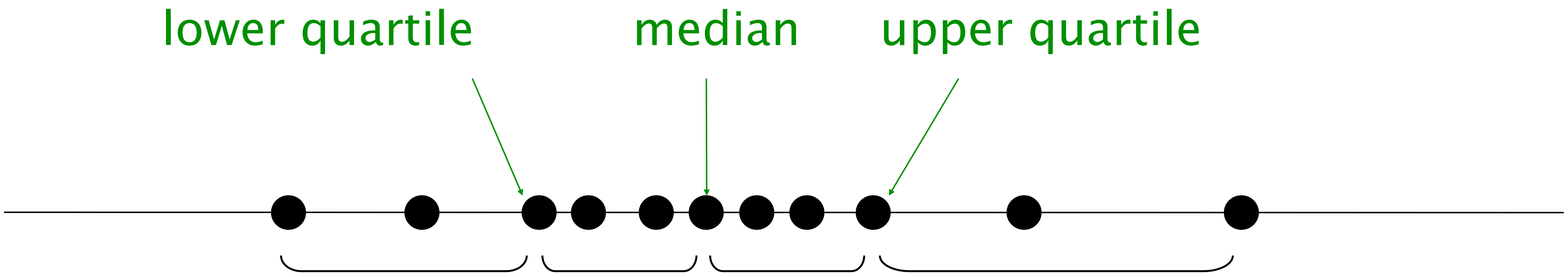




http://en.wikipedia.org/wiki/Normal_distribution

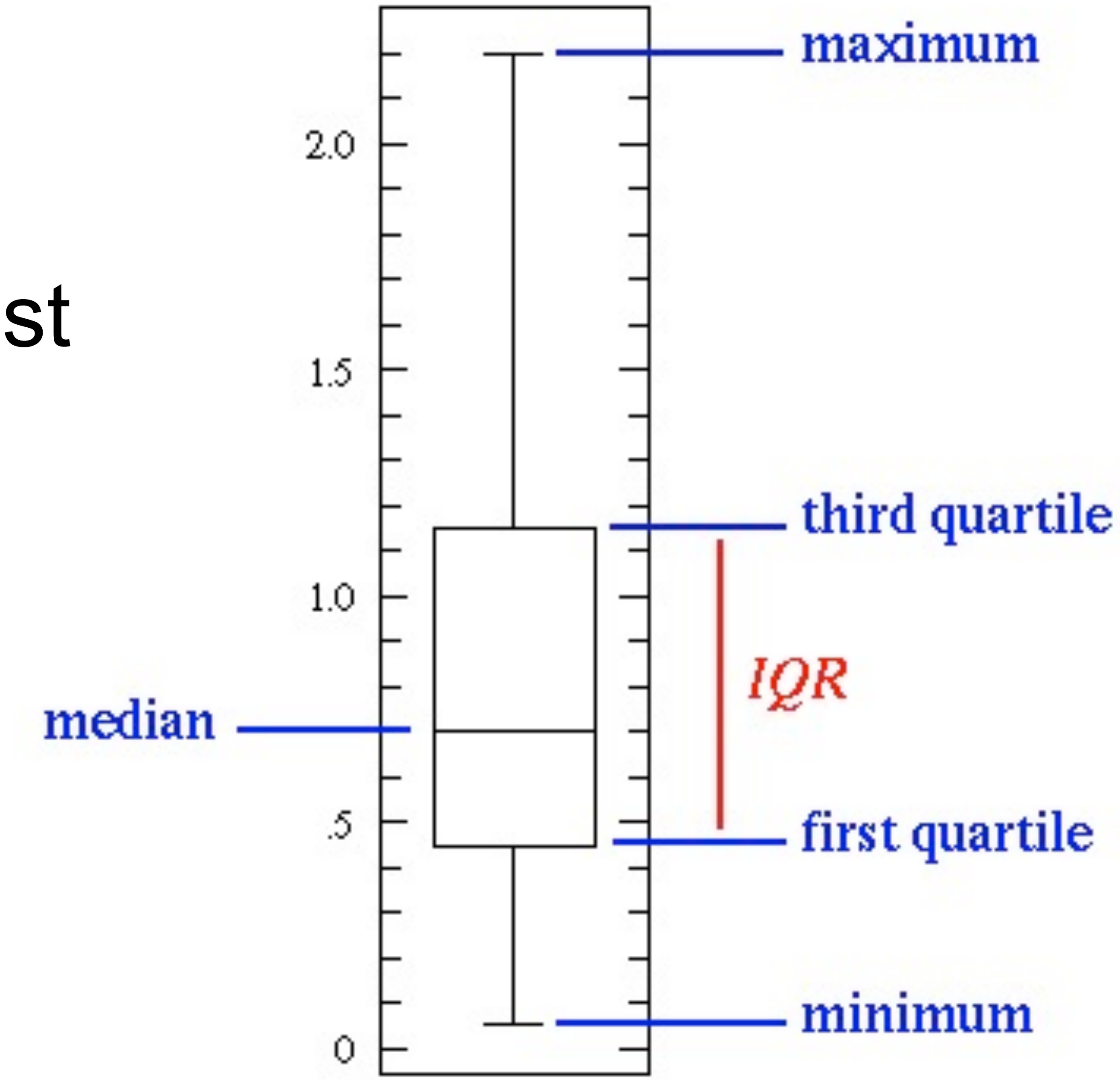
# Quantile, Quartile and Percentile

- 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

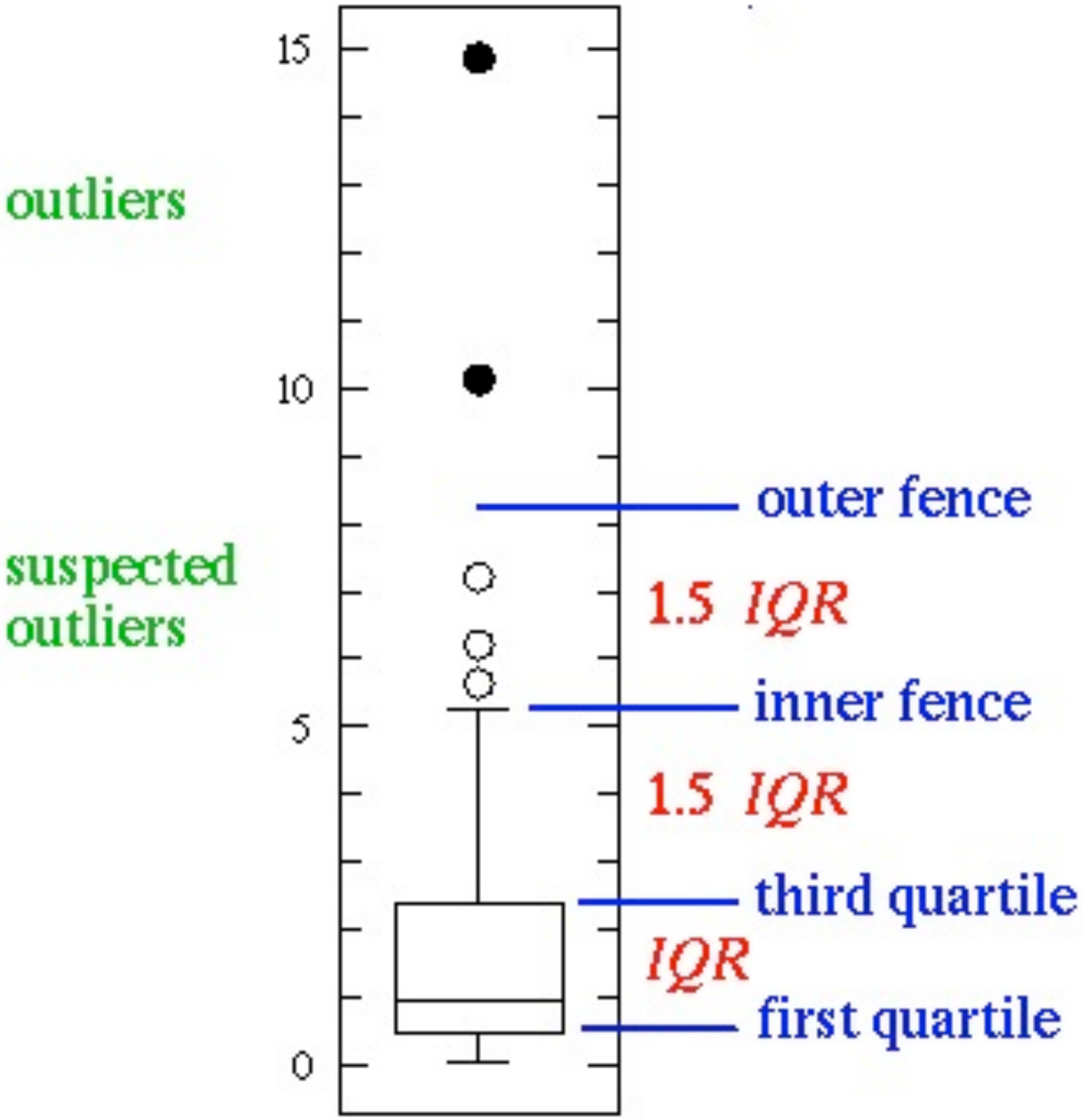
- Also known as
  - box-and-whisker diagram
  - candlestick chart
- Quick overview of the most important values



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



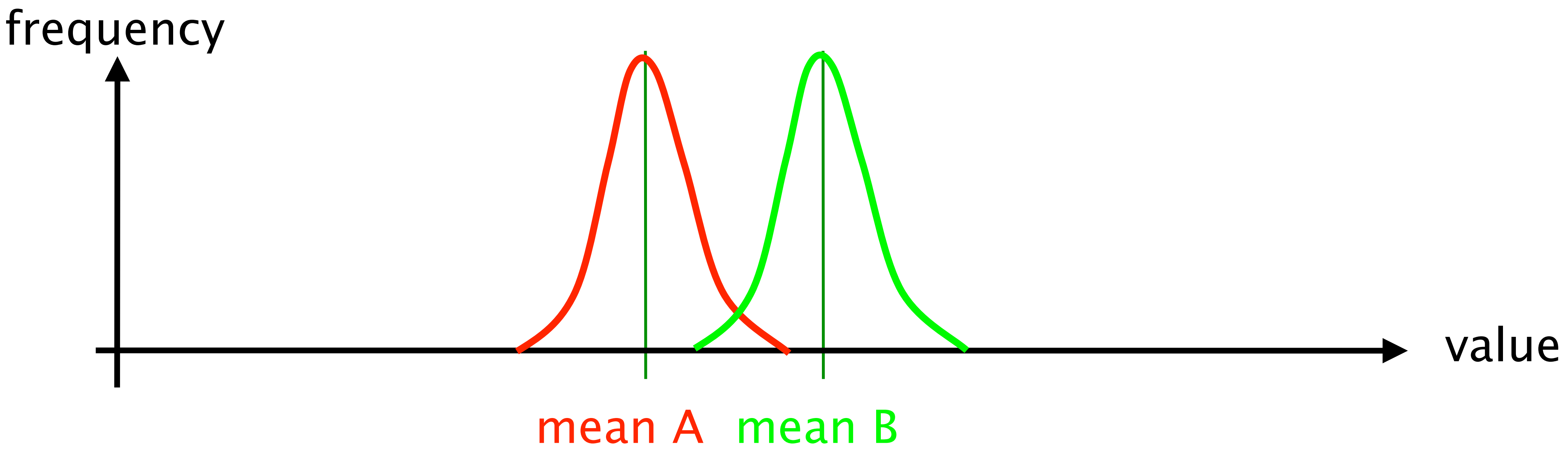
- Outliers are not generally bad – they give valuable information
- With large data sets outliers can often not be avoided

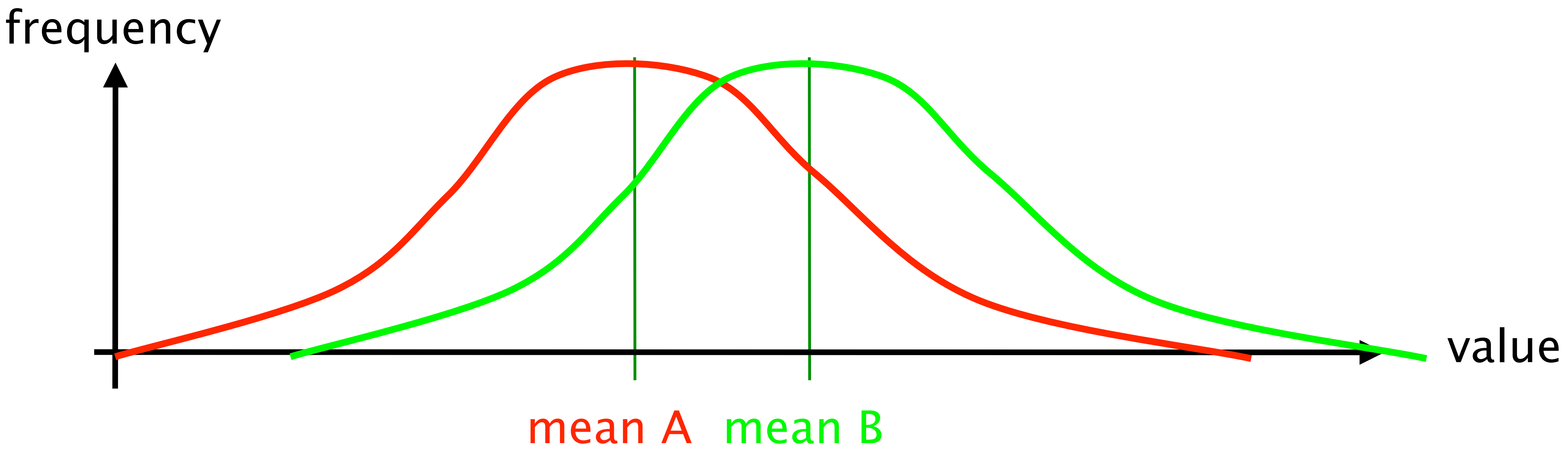
# Creating Boxplots with Excel

- 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?

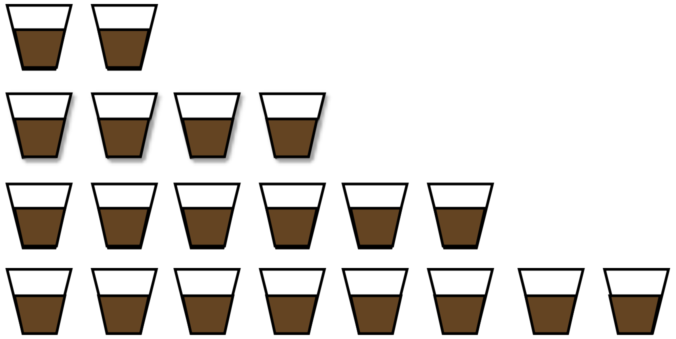




# Example: Pepsi Challenge

- 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 have different causes:
  - 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%

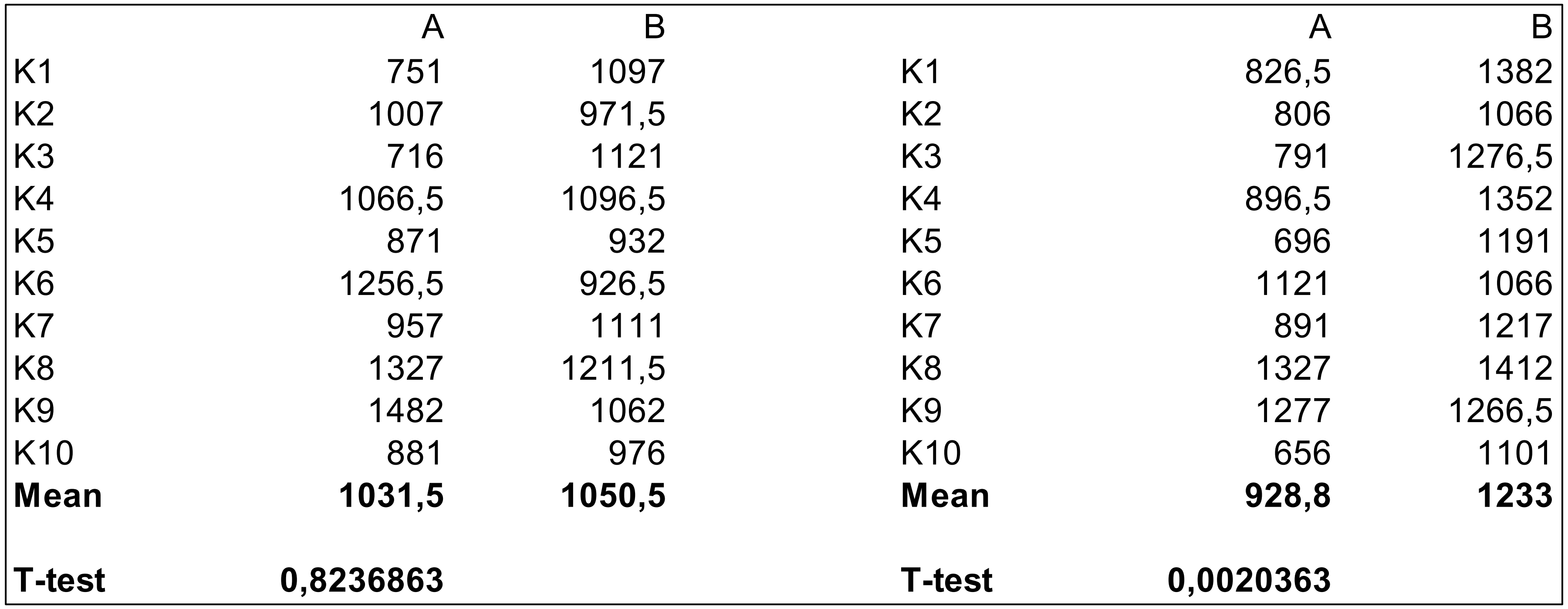
# Significance

- 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)



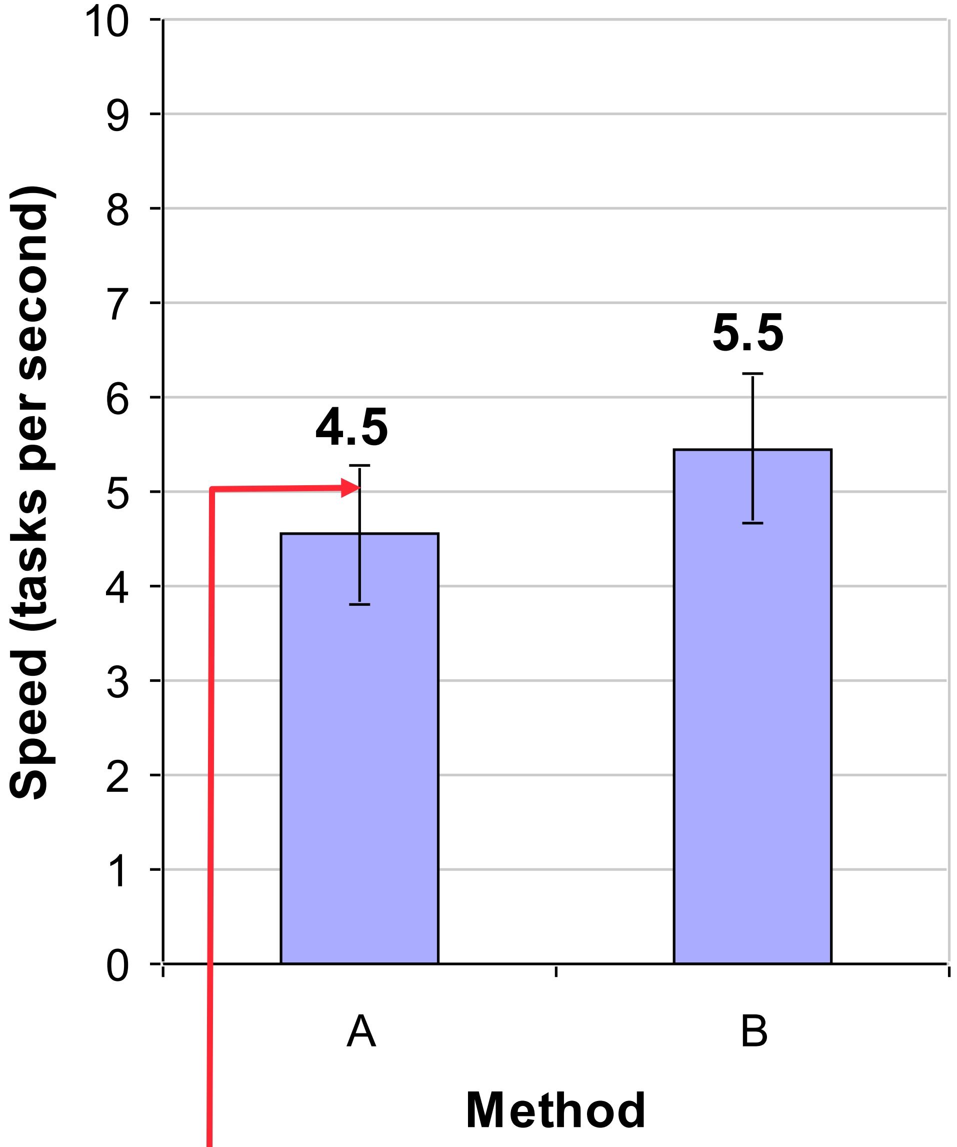
|  | A | B |  | A | B |
|---|---:|---:|---|---:|---:|
| K1 | 751 | 1097 | K1 | 826,5 | 1382 |
| K2 | 1007 | 971,5 | K2 | 806 | 1066 |
| K3 | 716 | 1121 | K3 | 791 | 1276,5 |
| K4 | 1066,5 | 1096,5 | K4 | 896,5 | 1352 |
| K5 | 871 | 932 | K5 | 696 | 1191 |
| K6 | 1256,5 | 926,5 | K6 | 1121 | 1066 |
| K7 | 957 | 1111 | K7 | 891 | 1217 |
| K8 | 1327 | 1211,5 | K8 | 1327 | 1412 |
| K9 | 1482 | 1062 | K9 | 1277 | 1266,5 |
| K10 | 881 | 976 | K10 | 656 | 1101 |
| **Mean** | **1031,5** | **1050,5** | **Mean** | **928,8** | **1233** |
|  |  |  |  |  |  |
| **T-test** | **0,8236863** |  | **T-test** | **0,0020363** |  |

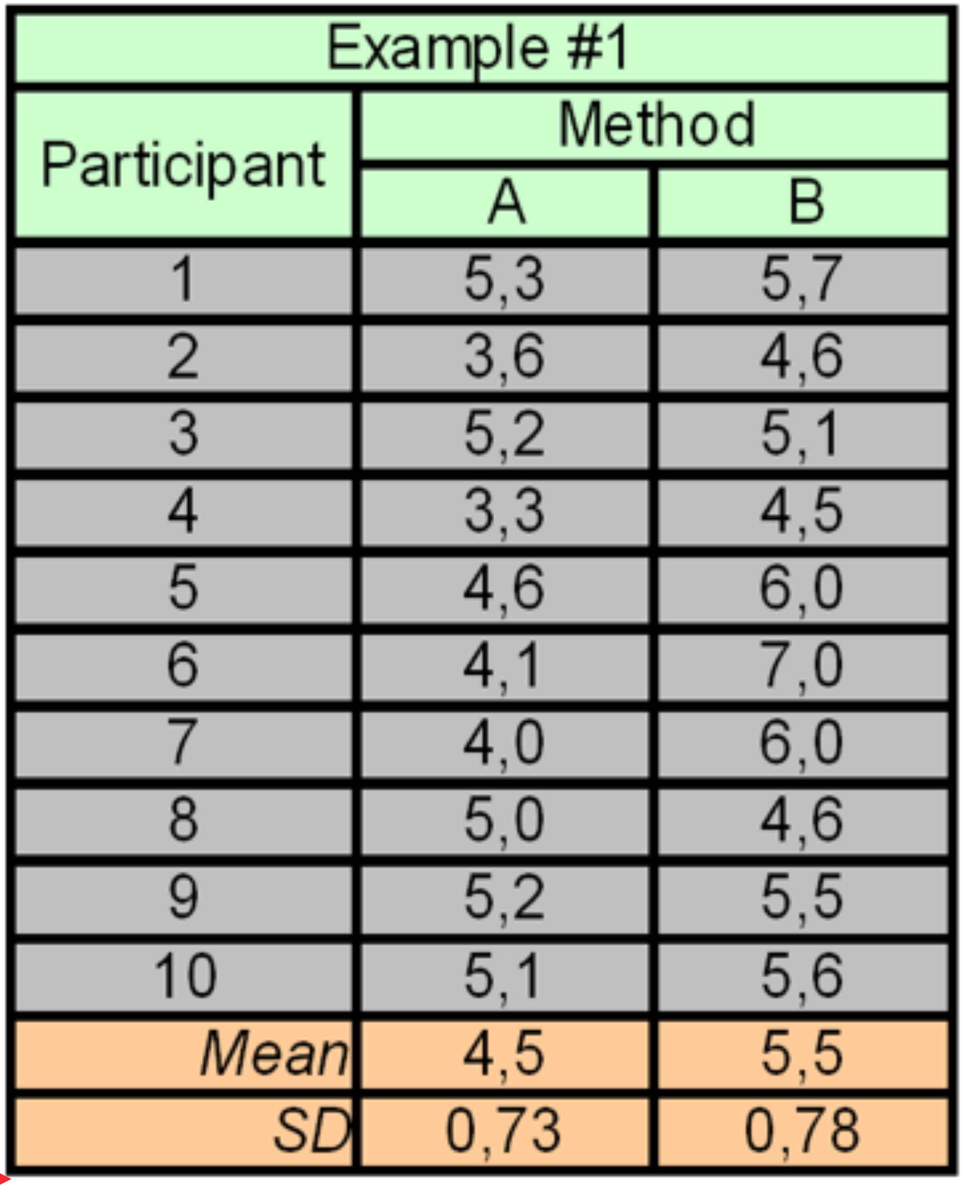
# Analysis of Variance (ANOVA)

- 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

# Significant Example



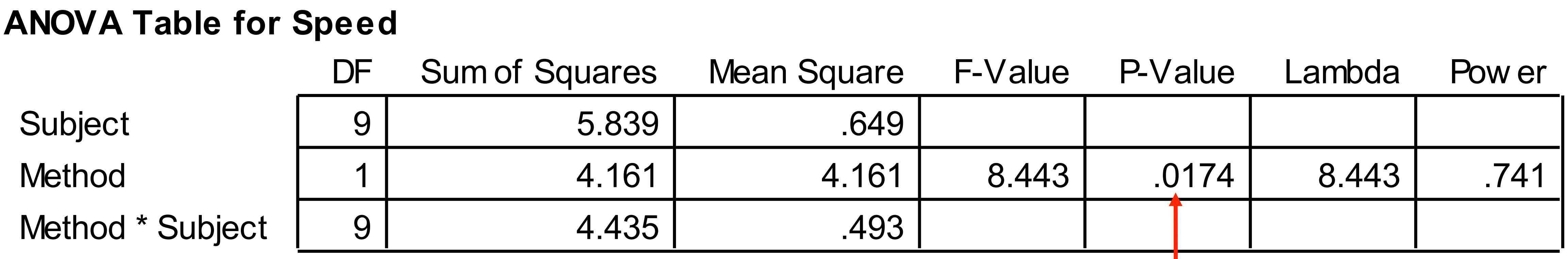


| Example #1 | | |
|---|---|---|
| Participant | Method | |
| | A | B |
| 1 | 5,3 | 5,7 |
| 2 | 3,6 | 4,6 |
| 3 | 5,2 | 5,1 |
| 4 | 3,3 | 4,5 |
| 5 | 4,6 | 6,0 |
| 6 | 4,1 | 7,0 |
| 7 | 4,0 | 6,0 |
| 8 | 5,0 | 4,6 |
| 9 | 5,2 | 5,5 |
| 10 | 5,1 | 5,6 |
| *Mean* | 4,5 | 5,5 |
| *SD* | 0,73 | 0,78 |

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…

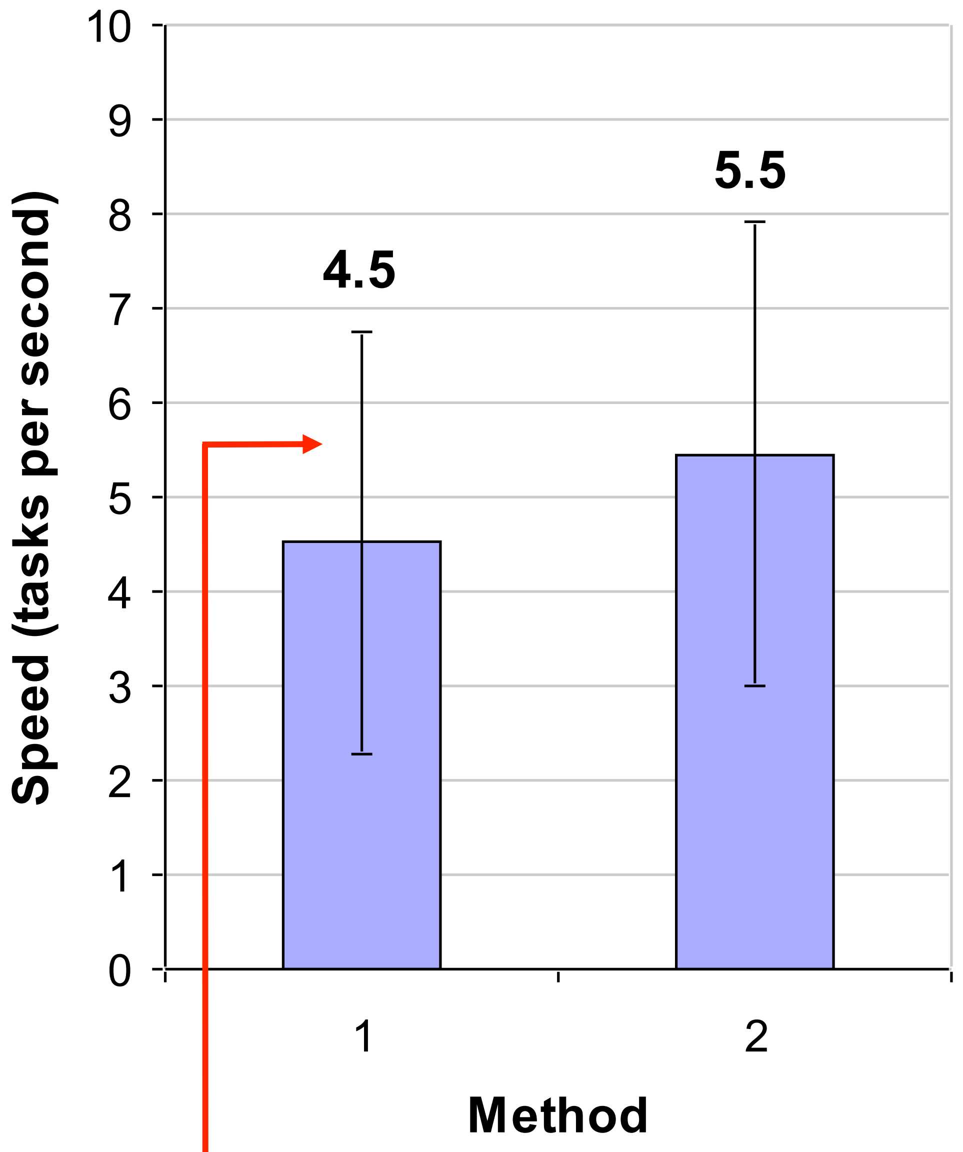
$F_{1,9} = 8.443, p < .05$

Thresholds for “p”

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

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

# Not Significant Example

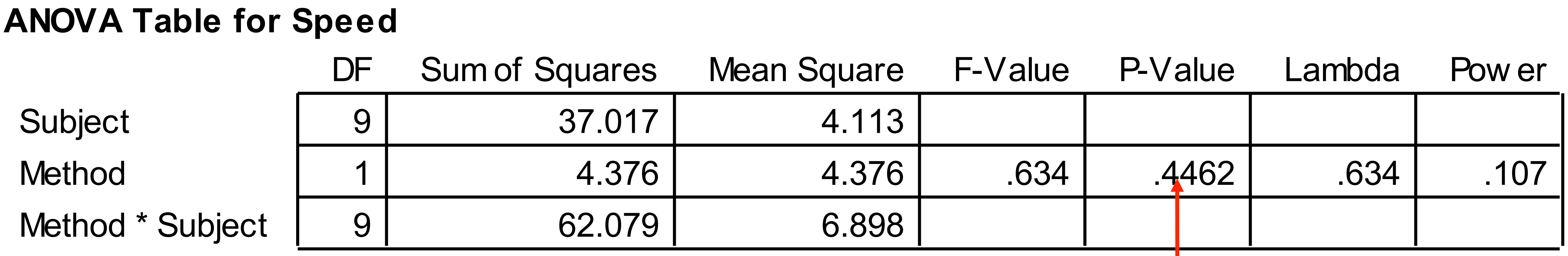


| Example #2 | | |
|---|---|---|
| Participant | Method | |
| | A | B |
| 1 | 2.4 | 6.9 |
| 2 | 2.7 | 7.2 |
| 3 | 3.4 | 2.6 |
| 4 | 6.1 | 1.8 |
| 5 | 6.4 | 7.8 |
| 6 | 5.4 | 9.2 |
| 7 | 7.9 | 4.4 |
| 8 | 1.2 | 6.6 |
| 9 | 3.0 | 4.8 |
| 10 | 6.6 | 3.1 |
| *Mean* | 4.5 | 5.5 |
| *SD* | 2.23 | 2.45 |

Error bars show
±1 standard deviation

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

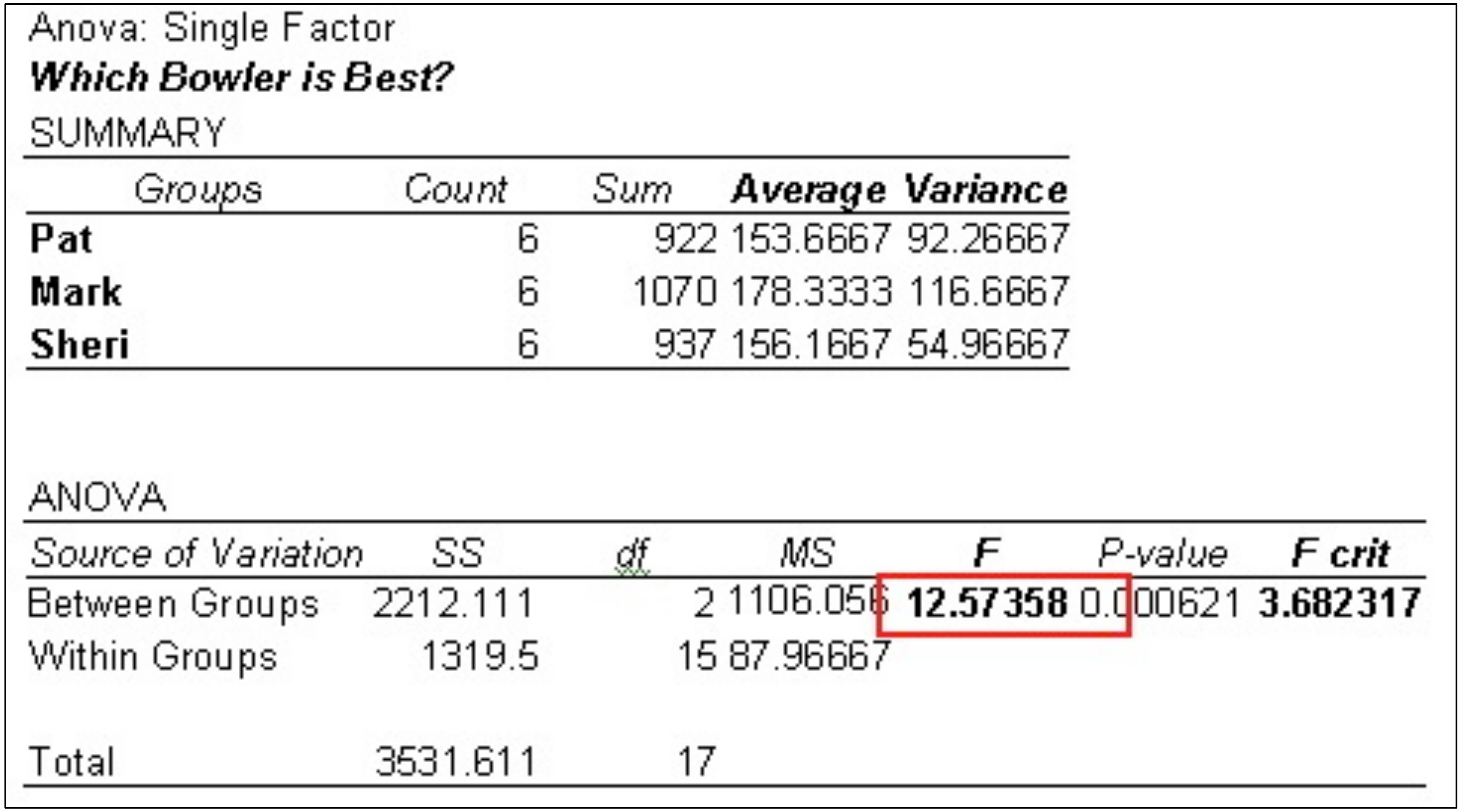
Note: For non- significant effects, use "ns" if

- 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

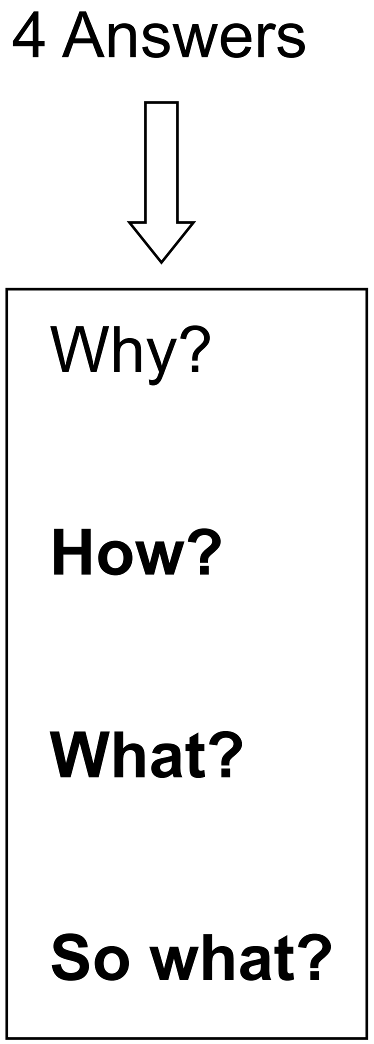
# Overview Parametric and Non-Parametric Tests

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



4 Answers

Why?

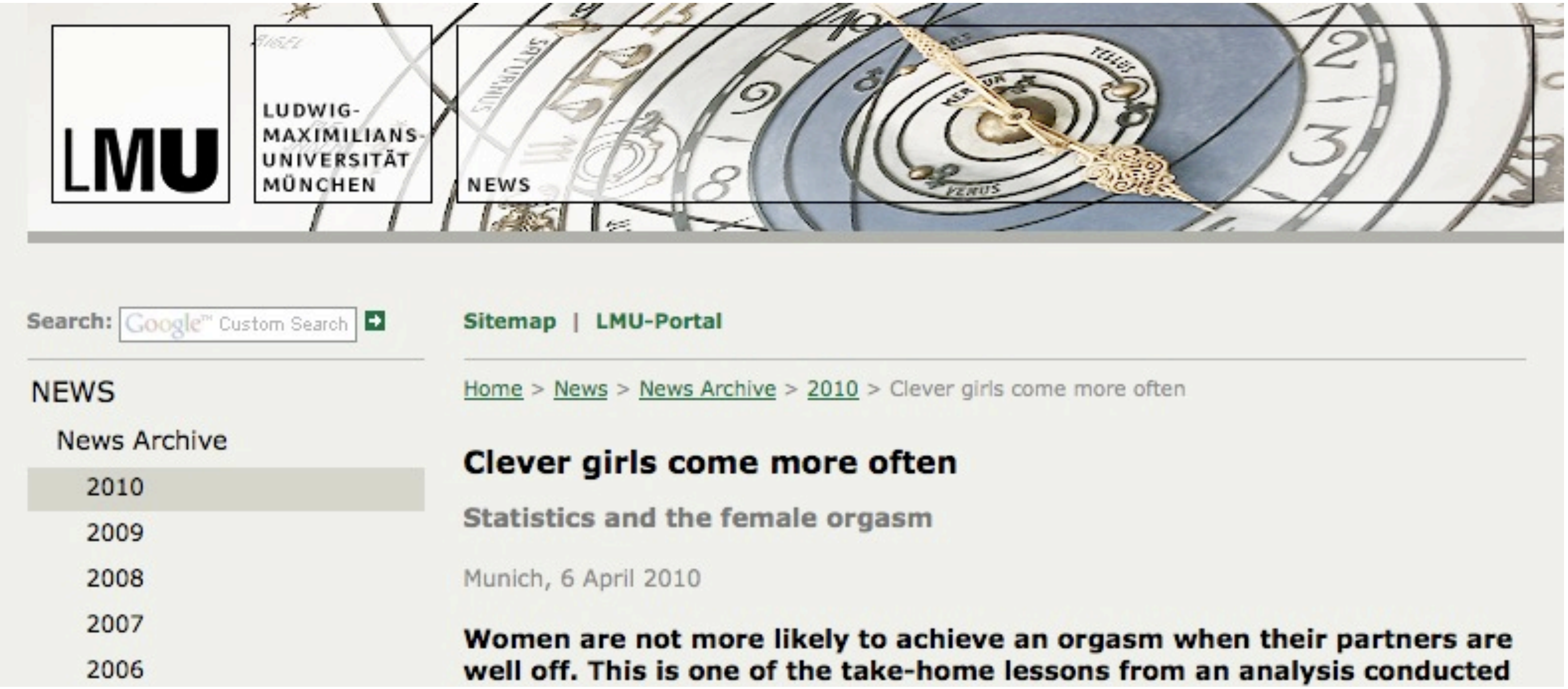
**How?**

**What?**

**So what?**

# Reporting Study Results

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


LMU
LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN
NEWS
Search: Google Custom Search
Sitemap | LMU-Portal
NEWS
News Archive
2010
2009
2008
2007
2006
Home > News > News Archive > 2010 > Clever girls come more often

**Clever girls come more often**

**Statistics and the female orgasm**

Munich, 6 April 2010

**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**

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)

# 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

- 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 ...