Research Methods

Experiment Design

Dr. Sven Magg, Prof. Dr. Stefan Wermter

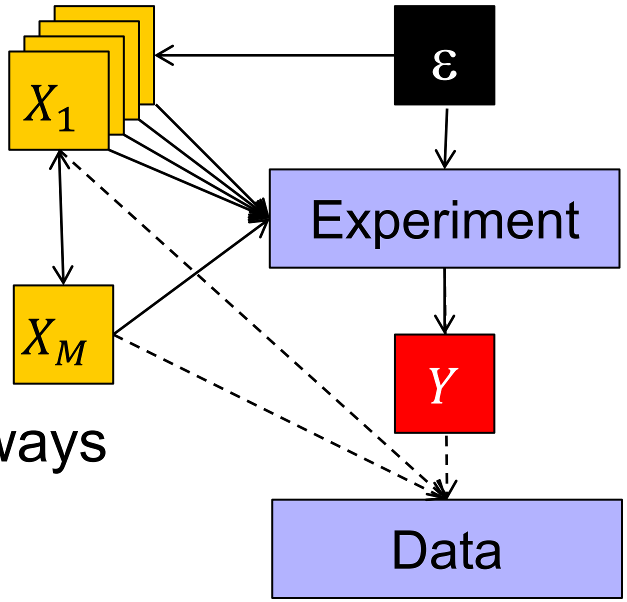
Knowledge Technology

http://www.informatik.uni-hamburg.de/WTM/

Plan for today!

1. How to handle “Effects”
2. Manipulation Experiments
3. What defines a “good” study?
4. Concept of control and Blinding
5. Confounds and spurious effects
6. Guidelines for experiment design

Finding X



- Step 1:
  - Define factors X_M and $X_{1 \dots n}$
  - Define outcome Y
  - One by one, find valid and reliable ways to measure Xs and Y

- Outcome of step 1:
  - A tree for X and Y, listing all factors and their measurements (=variables)
  - Decision of which variable(s) to manipulate and which to control
  - Optimal: A diagram of interactions between variables (model)

Sugar and Children

- Sugar is thought to have an effect on the activity levels of children: Sugar eaten → Child more active
- No effects could be found in several studies

Hoover, Daniel W., and Richard Milich. "Effects of sugar ingestion expectancies on mother-child interactions." Journal of Abnormal Child Psychology 22.4 (1994): 501-515.

- expectation of a sugar effect by parents may itself bring about significant behavioural changes in their children

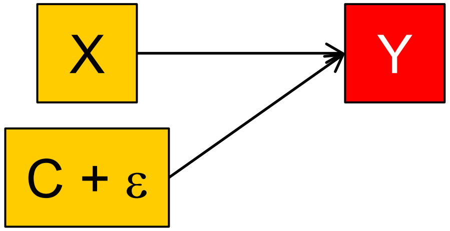
- Hypothesis: “expectancy effects influence mothers' assessments of their children's behavioral sugar sensitivity and thereby influence their subsequent interactions.”

Finding and Measuring X and Y

- Independent, manipulated factor:
  - Knowledge of mother about sugar intake *Category, Binary*
- Dependent factors
  - Assessment of the child's activity by mother *Post-Interaction Questionnaire*
  - Behaviour of mother *Observation + Rating*
- Controlled Factors
  - Age of children *Ratio, Years*
  - Parent's education level *Ratio, Time of education*
  - Previous bias towards child's behaviour *Eyberg Child Behavior Inventory ECBI*
  - Mother's personality *Wesley Cognitive Rigidity Scale*
  - The interaction in each iteration

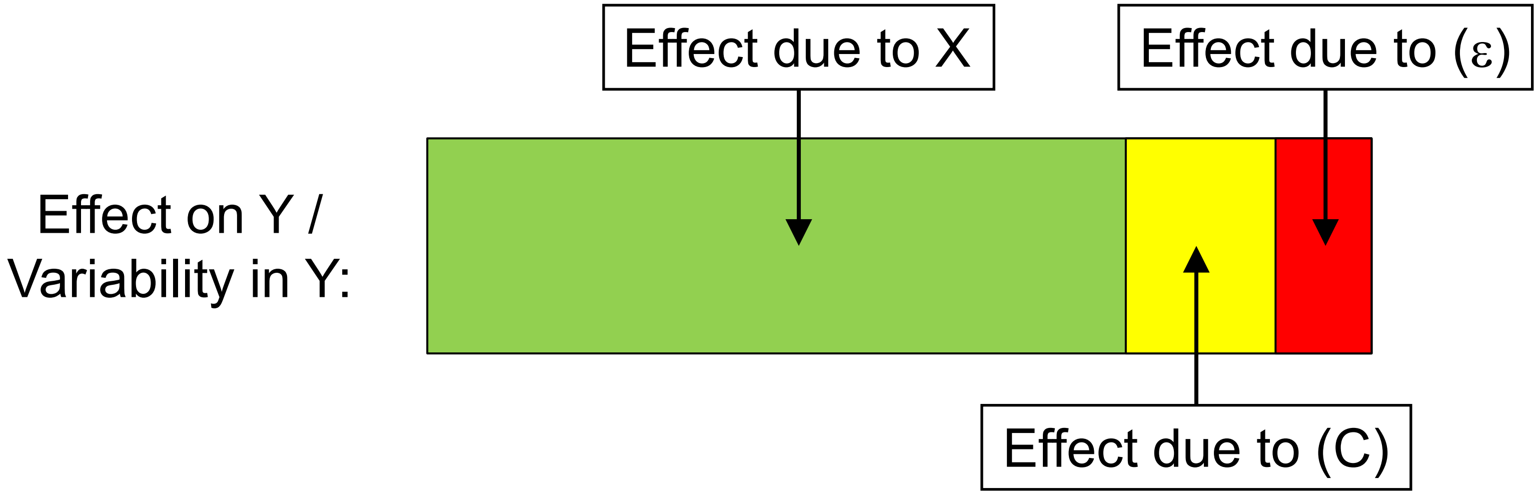
Effects

- Experiments are conducted to
  - correctly attribute the cause of a change (or lack of change) in a dependent variable
  - correctly attribute the causes of effects



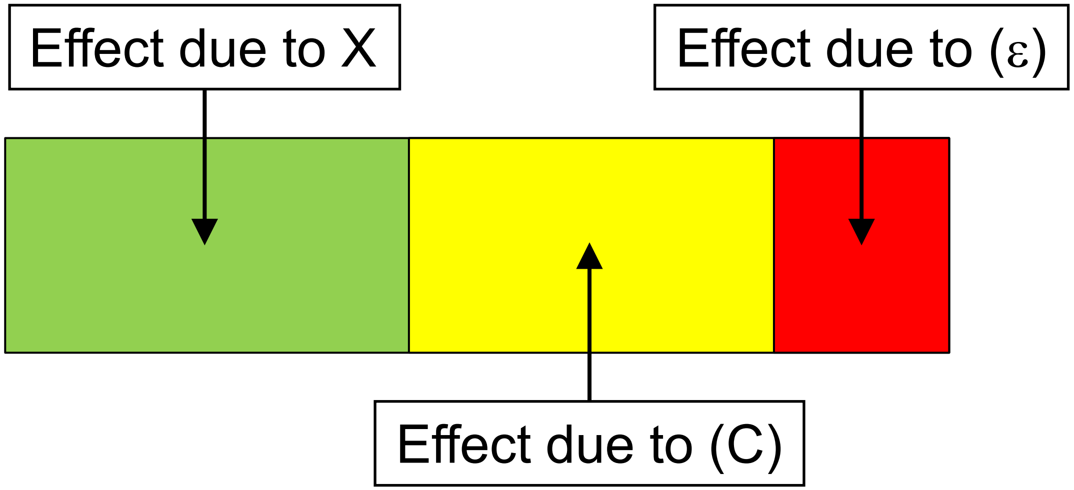
- If X is a cause of Y, then X should produce and effect on Y
- “Effect” usually the variance in Y explained by X

Effects



- The total effect on Y is the sum of different causes
- Effects due to
  - the independent variable(s) (X)
  - measured, controlled variables (C)
  - unmeasured, extraneous variables (ε)
- Aim of a well designed study: $X \gg (C + \varepsilon)$

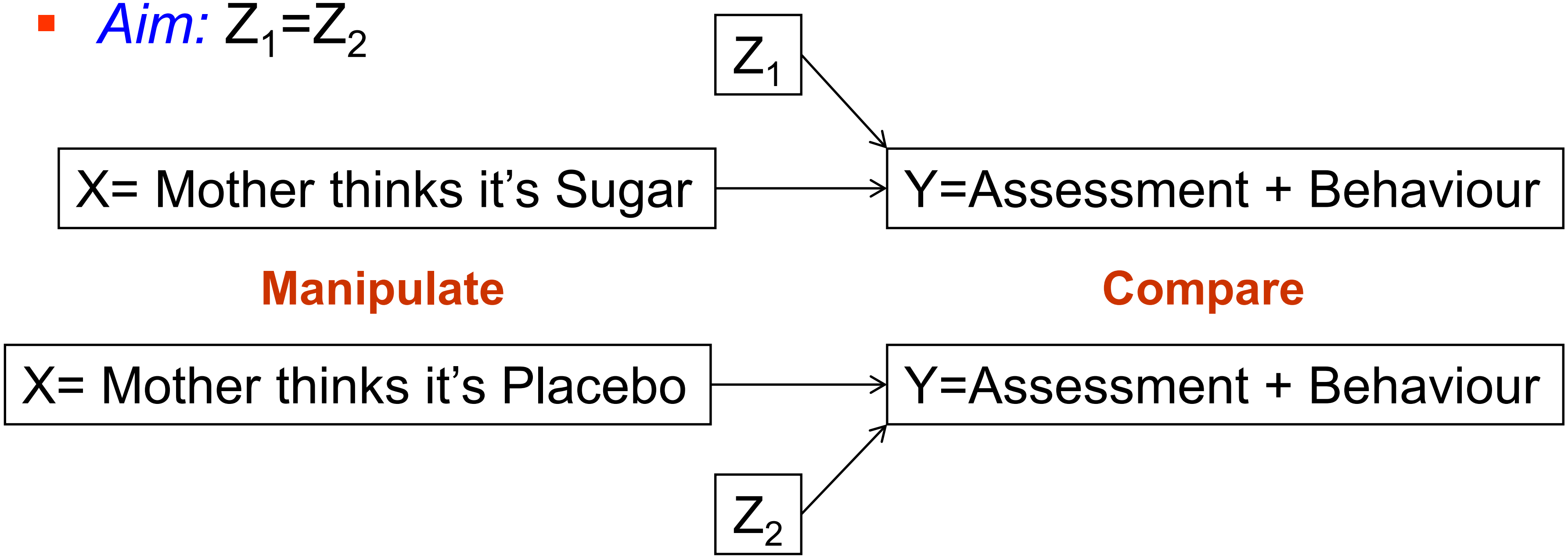
When things go wrong…



- If the effects due to controlled and extraneous variables are too high, we can't be sure whether X is a cause of Y
- If you want to claim that X causes Y, you must rule out all other causes Z=(C+ ε)
- "Ruling out" = Z has negligible effects on Y

Manipulation Experiments

- We manipulate X and measure effects on Y for each value of X
- "each value of X" = *Condition*
- *Variability in Y through X and Zs in each condition*
- *Aim:* $Z_1=Z_2$

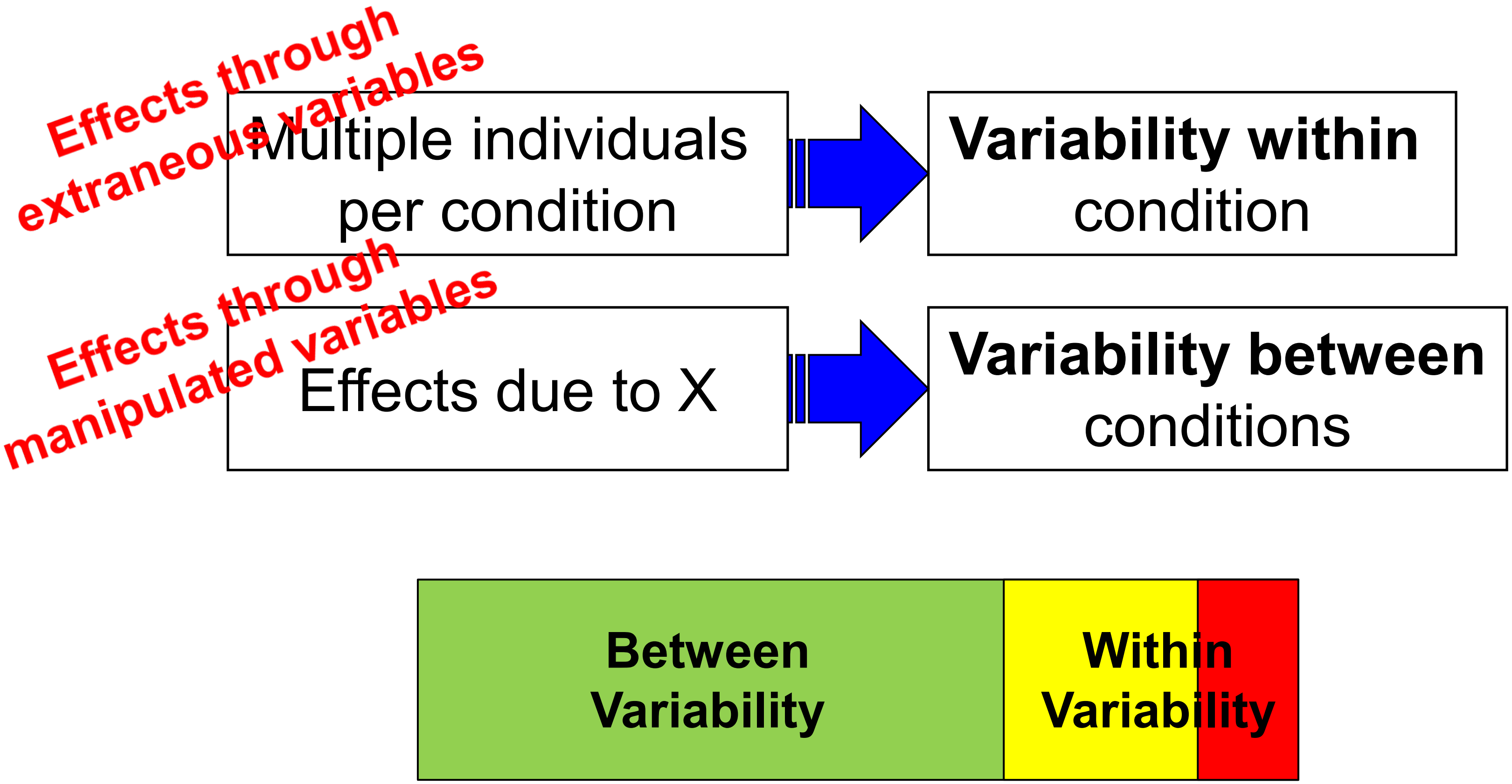


Variability everywhere

- Practically it's almost impossible that $Z_1=Z_2$ if we are conducting experiments outside simulation
- Humans and robots as individuals will inevitably introduce variation between individuals

- In our expectation effect study:
  - Each mother
    - has her own bias, behaviour, predisposition, etc.
    - will assess the interaction differently each time
  - Each child will act differently each time
  - Experimenters may vary, act differently each time, etc.

Variability everywhere



- If between-variability is large compared to within-variability, we speak of significant effects

Group Task!

3 10

A confounding variable has an association with both a dependent and independent variable…..

1. **What does this mean for the variability?**
2. **What would be an example from our sugar study?**

Confounding Variables

- A confounding variable has an association with both the dependent and independent variable

⇒ There is a systematic association between the confounding variable and an experimental condition!

⇒ Either or both affect Y! The effects are confounded

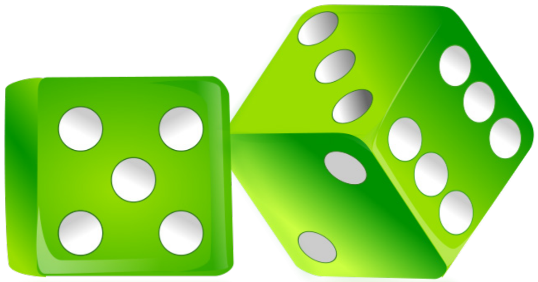
- What could a confound be in our example study?
  - Measured or extraneous variables not evenly distributed over conditions
- For measured variables we can check…..
- Is there a way to check for unmeasured variables?

Design Strategies

- We want to design the experiments such that
  - we avoid possible confounds
  - reduce within-variability

- Within-variability is introduced by
  - differences between our subjects
  - differences between single trials, e.g.
    - environmental differences,
    - human subjects (learning, variability in behaviour),
    - human experimenters
    - ….

Randomisation



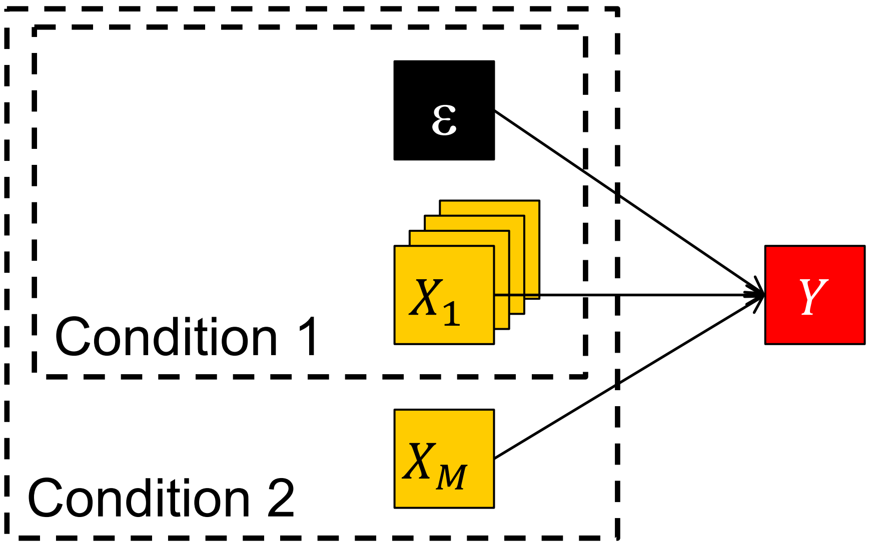
- We want to reduce the chance of a confound

- Sugar study subjects: 31 boys aged 5-7 and their mothers
- Randomly assign the subjects to the two conditions
- In the study:
  - 16 subject in the sugar condition, 15 in the placebo
  - Checked good randomisation for most factors in Z:
    - age, parent's education, ECBI scales (Intensity & Problem)

- Does randomisation solve our confound problem?
  - Not for sure, but it's the best strategy we have…..

The Concept of Control

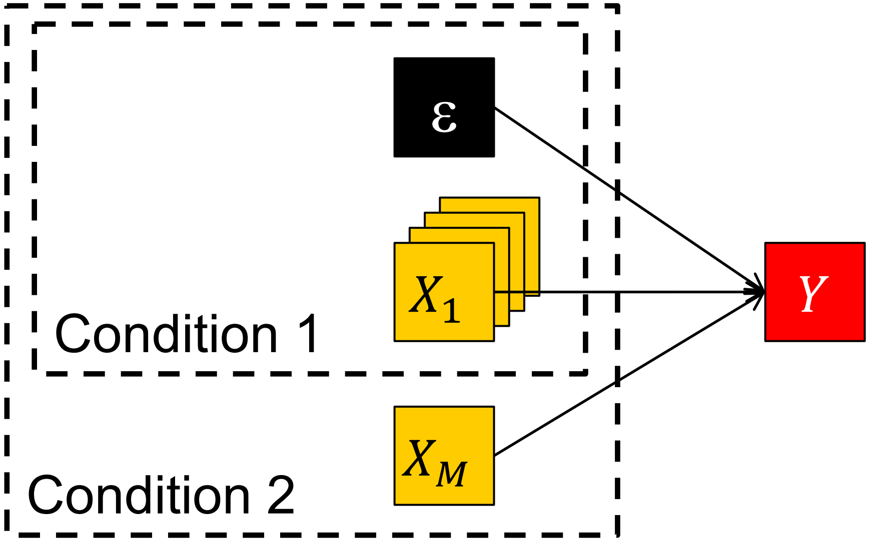
- If I can’t reduce within-variability, maybe I can measure the effect of all extraneous variables and subtract it?



- Create 2 conditions:
  1. U → Y
  2. U + x → Y
- If U captures all controlled and extraneous variables, the difference of effect on y between condition 1 and 2 is the effect of x
- Condition 1 shows the within-, condition 2 all variability

The Concept of Control

- How can we capture all controlled and extraneous variables?
  - Rigorous trial procedure to keep everything but x constant
  - Using a Placebo
  - Blinding



- Good news: Using a computer (e.g. simulation) gives you the first for free and automatically eliminates the other two, or does it?

Placebo Studies

- U should include everything but x
- If x includes a procedure/action/etc., this already introduces effects through extraneous factors
- Why?
  - Expectation effects
  - Other subtle differences like timing, changes in action sequence, ….

$\Rightarrow$ The procedure/action/etc. has to be part of U

- Often difficult to find appropriate replacement

Placebo Studies

- Aim of a placebo trial:
  Subjects cannot detect which condition they are assigned to

⇒ **Single-blind trial**
Subjects are blinded towards their group affiliation

- Examples:
  - Give *x* as medication in form of a pill (pill without active substance in *U* administered the same way)
  - Study involving medical operations (fake operation in *U*)
  - Robot performs similar but insignificant action

Blinding

- Assume we want to test effects of a homeopathic substance
  - Placebo study against pill without active ingredient
  - Fix administration procedure (Wording, sequence of actions, behaviour of experimenter, etc.)
- How can we be sure that the conductor of the experiment does not give clues without knowing?
  - Subtle change in voice, facial expression, etc.
  - In the apple experiment: Did knowing the apple change your behaviour?

⇒ Blinding the experimenter to the subject's group affiliation

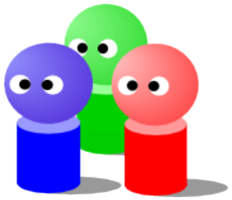
Double-Blind Studies

- Blinding the experimenter to the subject’s group affiliation
  - The person in contact with the subject **has no knowledge** about the condition the subject was assigned to

- We **double-blind**, because we don’t want that
  - the experimenter can disclose information to the subject
  - some biases of the experimenters affect our study
  - people who rate the subjects do not bias the result

- Often difficult to keep experimenters from knowledge
- Many biases of the experimenter can not be avoided!

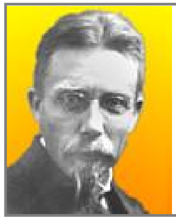
Group Task!

 2  5

Are we already double-blinding when we use computers as interface between experimenter and human?

Double-Blind Studies

- Experimenter biases not necessarily captured by blinding:
  - The experimenter has some stakes in the results
  - A group of experimenters have a systematic bias
- Nobody is free of bias and your design has to address this



"We may fondly imagine that we are impartial seekers after truth, but with a few exceptions, to which I know that I do not belong, we are influenced—and sometimes strongly—by our personal bias; and we give our best thoughts to those ideas which we have to defend. *- August Kroth 1874-1949*

- **Triple-Blind** studies:
  - Even the people doing the analysis are blinded towards which group's results they analyse
  - In many cases difficult to set up properly
  - Rarely seen

Protocol and Procedure

- **Experiment Protocol**
  - Description of the study design (control groups, blinding, etc.)
  - If human subjects are involved, this has to be reviewed by an ethics committee

- **Experimental Procedure**
  - The step-by-step description of what happens to each subject
  - Needs enough detail to make certain each subject is treated the same!
  - Especially important with multiple experimenters!

Back to the effects of sugar….

- Strict experimental procedure:
- Pre-Study:
  1. Selection of mothers with high expectancy through phone interviews.
  2. In interview: Information that children will be given placebo or sugar
  3. Tell mothers: Children have to fast beginning at bedtime the night before and data will be collected before the first meal in the morning
- Pre-Interaction:
  1. Separation of mother and child by blinded assistant (different rooms)

Mothers:	**Boys**:
2. Informed consent 3. Deliver expectancy manipulation 4. Wait 30min “for drink to work” 5. Receive interaction instructions 6. Sugar Expectancy Questionnaire 7. Manipulation Reminder	2. Informed consent 3. drink “Kool-Aid” flavoured with aspartame 4. Brief questioning about content of study 5. Play & watch Winnie the Pooh

Back to the effects of sugar….

- Scripted interactions between experimenter and subjects:
- Expectancy Manipulation
  - Sugar condition:
    “mothers were told their son would receive a large dose of sugar in a drink tasting like very sweet Kool-Aid. They were also told that they would wait one-half hour to give the drink "time to work" and then they would be observed playing with their sons and while asking them to do certain identified tasks.”
  - Placebo condition:
    “[…] same information but were told their children had been selected for the placebo group and would therefore have Kool-Aid flavored with NutraSweet. They were asked not to mention this to their children and told that NutraSweet does not have behavioral effects on children.”

Back to the effects of sugar….

- Scripted Interaction Task
  “They were told that the researchers would be videotaping their sons' behavior for 25 min. In the first 10 min, they were instructed to allow their sons to play with toys in the room and to "feel free to play with him." In the last 15 rain [sic], the task segment of the interaction, the mothers were to get their sons to pick up the toys, copy some simple geometric drawings, and then build a LEGO house together.”
- Manipulation Reminder
  - Sugar condition: *“O.k., [Name] had his sugar drink about half an hour ago and now we can get started.“*
  - *Placebo condition: “O.k., [Name] had his NutraSweet Kool-Aid and now we can get started.“*

Back to the effects of sugar….

- Observational Data
  - Videotapes of the interaction were “coded” by raters blind to experimental condition on verbal and nonverbal behaviours:
    - commands (# of command statements)
    - proximity (% of time bodies overlap on tape)
    - criticisms (# of negative evaluative remarks by mother to child)
    - talking (frequency of utterances & proportion of time talking)
    - friendliness/warmth (both partners, overall rating)
  - Each variable was independently coded by 2 sets of 2-3 raters. Correlation between the sets to judge reliability
  - Several other variables were dropped due to low reliability
  - Children wore Actometers at wrists to measure activity (mins)

Effects to watch out for!

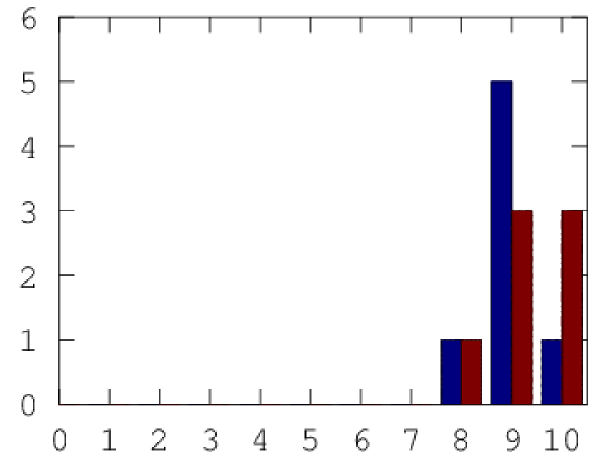
- Sometimes study is well designed but still goes wrong
- Some common things to look out for:
  - Boundary Effects (Ceiling and floor effects)
  - Regression effects
  - Order effects
  - Sampling Biases

- Run **pilot study** (small groups, dry run)
  - to test your design, not the hypothesis ("debug" the protocol)
  - calibrate measurements and parameters (e.g. number of subjects, number of trials, etc.)
  - test your measurements (reliability & validity) and analysis

Ceiling & Floor effects

Group	1	2	3	4	5	6	7	Ø
Test	9	10	9	8	10	10	9	9,29
Control	9	9	8	9	10	9	9	9,0

Scores on 7 tests on a scale 0-10



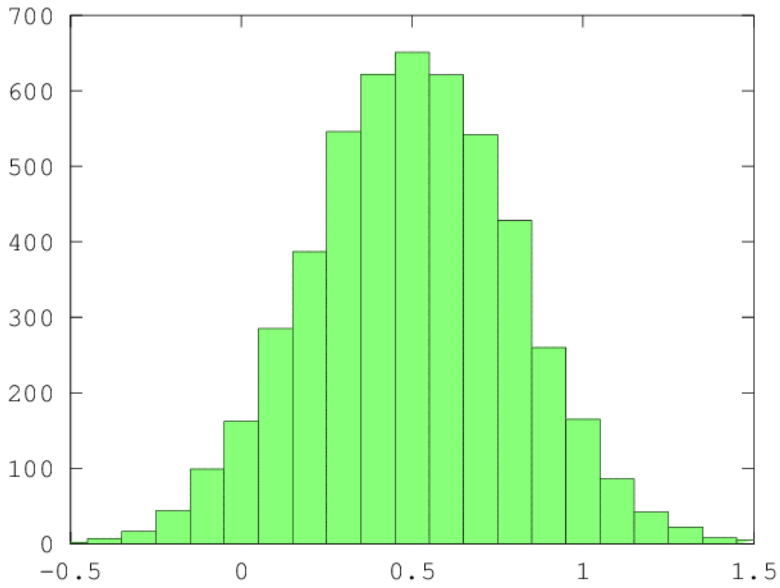
- Scores between test and control are almost equal
- Both a close to maximum value!
- Maybe the tests were too simple?
- Watch for boundary effects when results are near a possible maximum/minimum value that can be reached

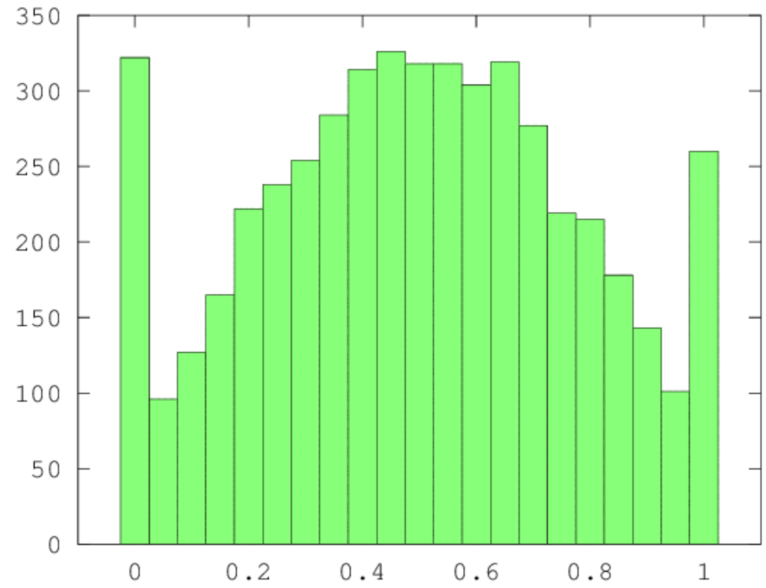
Ceiling & Floor effects

- How to detect boundary effects?
  1. Run a pilot study
  2. Estimate best/worst bounds for recorded values
  3. If both test & control are near this boundary
     $\Rightarrow$ Boundary effect may be lurking

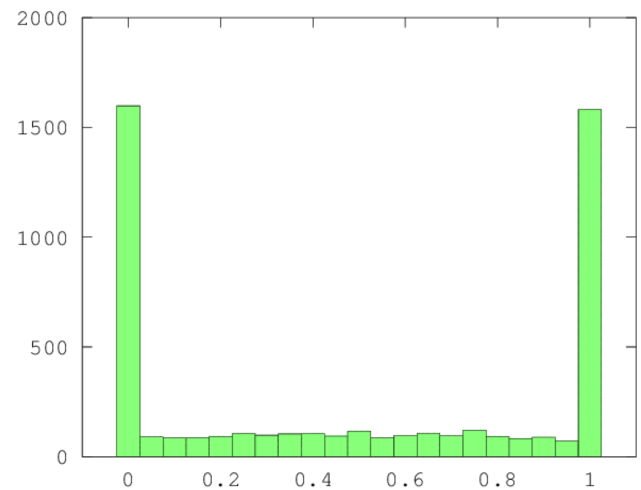
- Just because your are far away from the absolute boundary of your scale does not mean there is no boundary effect!
- Practical boundary not necessarily the max/min of the recording scale!!

Other boundary effects





- Cut-Off points often create statistical "artifacts"
- Be careful with random numbers and limits (think about deviation!)
- Better to "reflect" values at limits



Regression Effects

Test:	1	2	3	4	5	6	7	Ø
Alg 1	0	2	4	5	8	9	10	5.4
Alg 2	3	7	5	6				

- Algorithm 1 is tested on 7 problems
- Due to time issues, only test improved algorithm on problems where Alg1 performed below average
- Claim: Algorithm 2 is an improvement!

- *Regression towards the mean!*
- We expect values to be better if the result depends on a chance component!
- For a second test, always choose a representative sample

Order Effects

- Often there are sequences in the experiment procedure
  - Robot has to complete a series of tasks
  - Humans are presented a sequence of stimuli
- What kind of effects can the order of the sequence have?
  - The order has effects on the performance
  - Two groups receive same sequence, but one is more sensitive to a specific order than the other
- Often subtle!
  - garbage collection in Java at different points in time
  - Learning curve different depending on sequence in training

Order Effects

- How to detect order effects?

- **Counterbalancing**
- Run problem on all permutations of a sequence
  - Expensive!
  - Maybe just run a few to test

Sequence:	Alg1:	Alg2:
a,b	10	15
b,a	12	20

- Often impossible to run all permutations
- Sometimes the sequence is already part of the study
- If you are not sure: Include counterbalancing in pilot study!

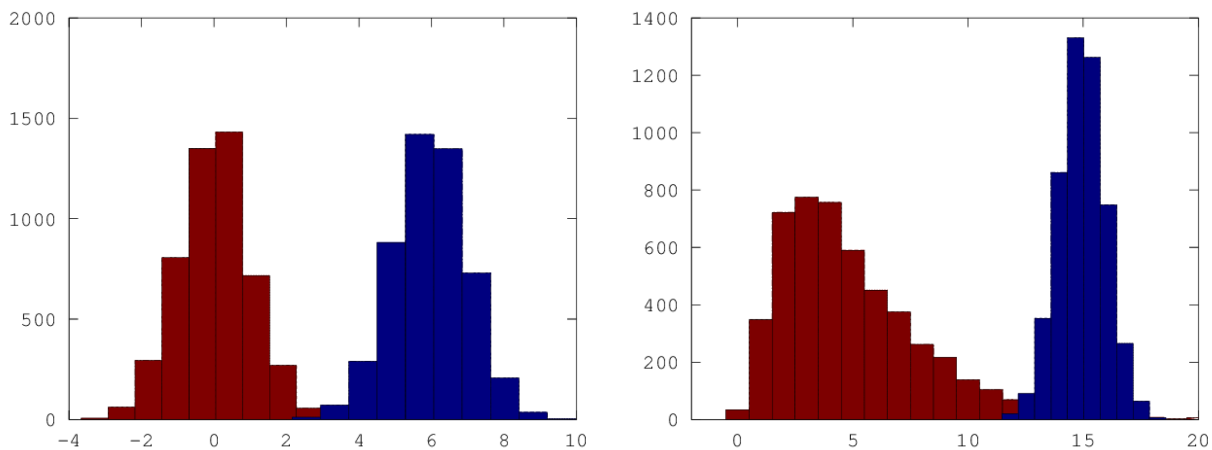
Sampling Bias

- When the collected sample is disproportionally biased towards a result
  - Statistics on computer usage per day taken from a questionnaire distributed at the Informatikum
  - Internet poll on political opinion on the Fox News Website
  - Internet polls in general

- Can be very subtle
  - Assume gender is independent variable, #siblings is noise
  - But, parents often have children until they have a boy
  - i.e. gender and #siblings are NOT independent

Sampling Bias

- How to detect sampling bias?
- If we have a well-designed experiment with a random sample: mean of y is different for different levels of x
- BUT: shape of distribution of y the same



- If the shape changes for different levels of x, this hints at another factor influencing membership in the sample!

What have we learned?

1. Collect information
   - Define and **write down** all factors (independent, dependent, extraneous) as table or diagram
   - Define variables and measurements for each factor you want to measure
     - scales (including estimate of expected values)
     - frequency of recording (once at beginning or end, at intervals)
     - estimate validity and reliability
2. Design Experiment
   - Set up protocol, define procedures (even for small studies)
   - Define procedure of analysis
   - Get someone to review protocol and procedures

What have we learned?

3. Run a pilot study
   - test the protocol, procedures AND analysis (“debugging”)
   - check for validity and reliability of measures
   - look for spurious effects and sampling bias
4. Discuss and interpret results
   - Do they show what you have expected?
   - Do they actually answer your question?
   - Did you address all competing hypotheses?
5. Run and repeat the experiment
   - Repetition to disclose effects of interaction between independent and noise variables

Homework!

Redesign your Apple Experiment

1. **Create (partial) List of factors, including:**
   - **Variables and scales**
   - **Measurements, estimation of validity/reliability**
2. **Define (at least one) procedure, explicitly addressing the points in 1.**
3. **Discuss possible spurious effects and how to avoid them**

Deadlines: **- November 26, 12:00 noon, Experiment protocol as PDF**

- November 27, Review and Improvement