CogSysIII Lecture 4/5: Empirical Evaluation of Software-Systems

Human Computer Interaction

Ute Schmid

Applied Computer Science, Bamberg University

last change May 8, 2007

Why and When Empirical Methods?

- User-centred approach to design: design based on real data not on imagination abour users
  - what kind of activities
  - needed support
- User studies at the beginning: collect data, analyse, abstract, build system
- Support of system evaluation: user test in controlled lab conditions or real surrounding

Empirical Research Methods

- Research Methods: “soft” skills
- How to generate a “operational hypothesis”
  e.g., $\bar{x}_1 < \bar{x}_2$ “Number of errors in test group (with help menu) is significantly smaller than number of errors in control group (without help menu)
  starting point: a verbally described hypothesis (derived from some theory of human information processing)
- Take care about “scale niveau” of data, select appropriate test
- Present results of inference statistics in combination with descriptive charts
- Knowledge about statistics necessary
- but: focus on the general procedure of empirical research
- see: Jürgen Bortz, Lehrbuch der empirischen Sozialforschung

Kinds of Empirical Studies

- Case-study vs. random sample
  - Look at a small number of people in detail (typically qualitative data from log-files, other protocolls)
  - Use a representative sample of users (typically in experimental settingsI
- Single shot vs. longitudinal
  - general characteristics vs. learnability questions
  - special problems of longitudinal studies “Regression to the mean”
- In natural setting or in laboratory
  - question of external validity
- Quasi-experimental or experimental

What is an Experiment?

Wilhelm Wundt (1886):
Das Experiment besteht in einer Beobachtung, die sich mit der willkürlichen Einwirkung des Beobachers auf die Entstehung und den Verlauf der zu beobachtenden Erscheinung verbindet.
(1908) additionally: *Wiederholbarkeit* and *Variierbarkeit*

- Willkürlichkeit: assign a treatment (independent variable) to a probe/subject
- Wiederholbarkeit: experiment can be performed at another time (another place, with other probes/subjects) leading to the same results (dependent variable)
- Variierbarkeit: treatment can be assigned deliberatly and systematically

Example

Word superiority effect (Wheeler, 1970)

- Treatment: Subjects are randomly assigned to one of the following conditions: (a) short presentation of a letter (e.g. “d”) , (b) presentation of a word (e.g. “word”) Afterwards: recognition (a) either “d” or “k”, (b) either “word” or “work”
- Dependent variable: recognition errors
- Result, significantly less errors in condition (b)

Quasi-Experiment

- Some characteristics are invariably correlated with the subject (no random assignment of treatment)
- Examples: gender, intelligence quotient, color of a person

Example: Patients in single bed rooms vs. multi-bed hospital rooms Dependent variable: days until discharge

- Problem: unrecognized factors correlated with the dependet variable (single rooms are more sunny, doctors take more time with patients in single rooms etc.)

Experiment and Causal Explanation

- If a treatment can be assigned randomly to subjects, we can assume that all other factors which might influence the outcome of the experiment are distributed randomly over the subjects
- In experiments, we can test hypotheses of effects (differences between different treatments, typically including a control group without treatment)
- Hypotheses about correlations do not allow causal explanations, they only allow statements about the kind and intensity of the co-variation of variables!

Internal/External Validity

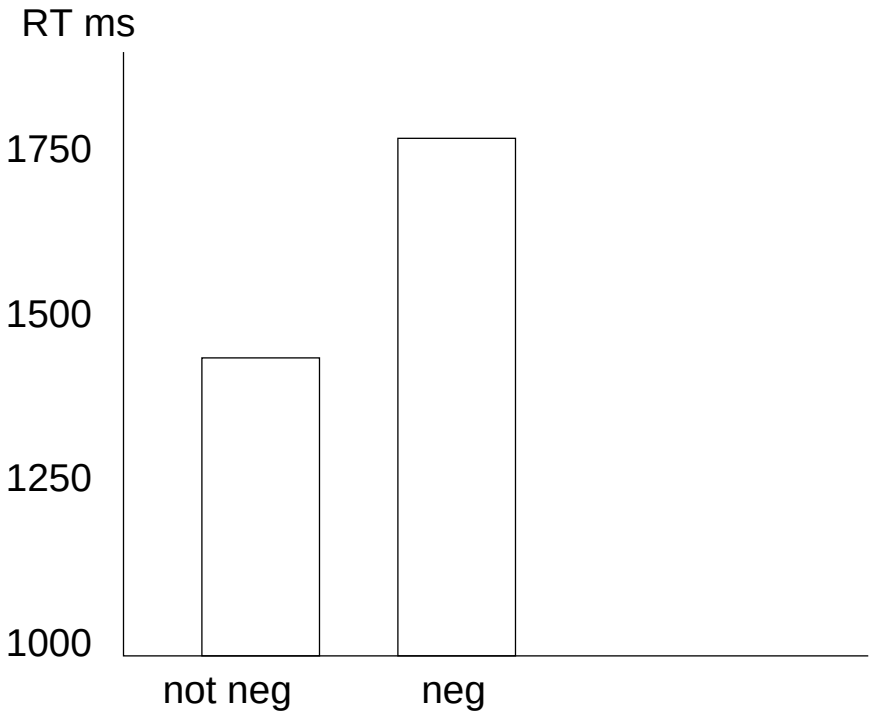
- Internal validity: no (not many) alternative explanations for results (effect can be ascribed to treatment)
- External validity: Generalizability of results from the experimental setting to a larger domain (necessary: representativeness of sample)

Types of Statistics

- Descriptive Statistics: Means and standard deviations of data (kind of measure depends on scale niveau, e.g. median vs. arithmetic middle)
- Inference Statistics: Generalize from sample to population with a certain amount of error (e.g. analysis of variance, chi-square, ...)
- Data Aggregation: cluster analysis, multidimensional scaling, principal component analysis

Presentation of Results

Example:



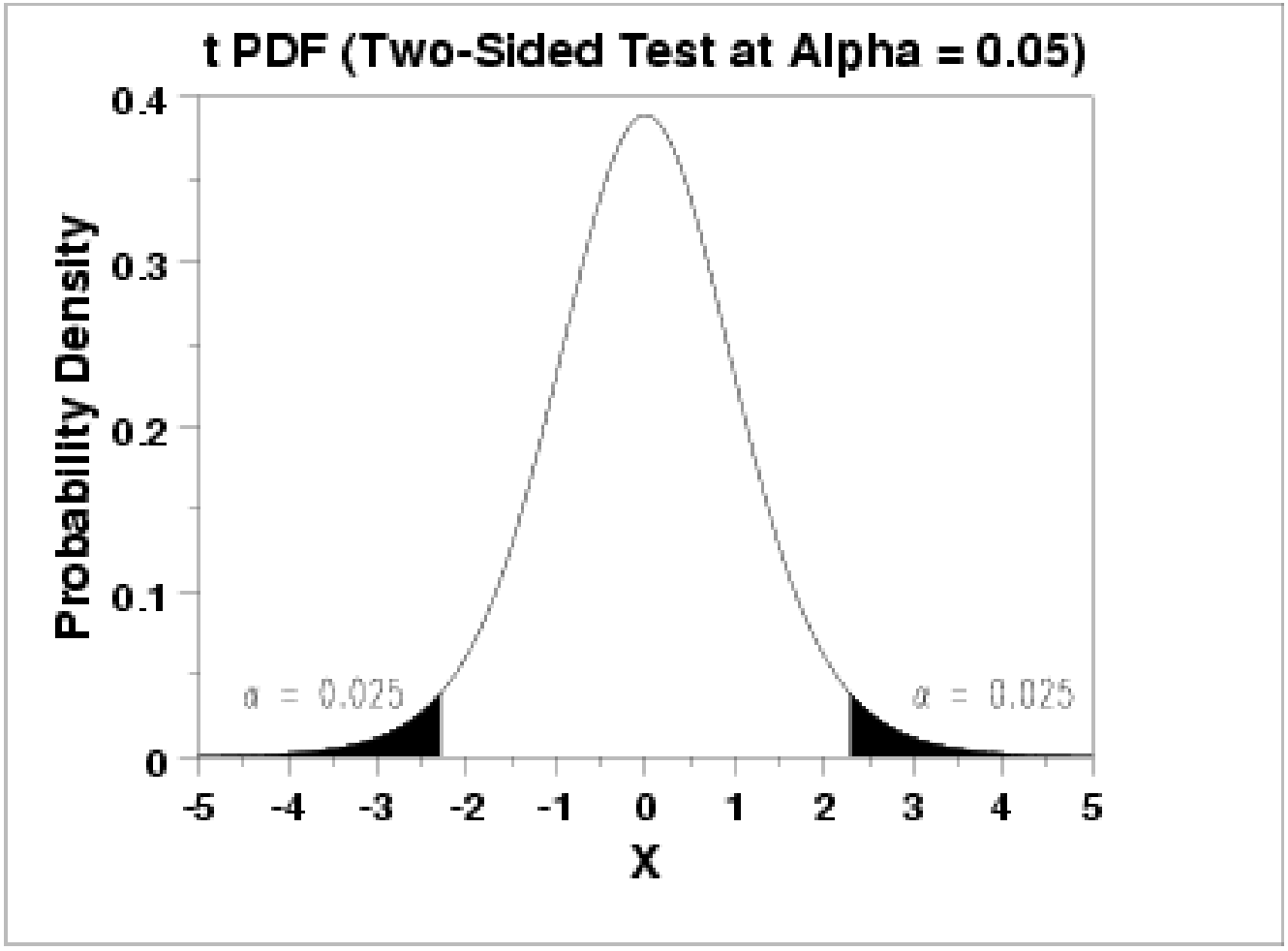
ANOVA $F(1, 31) = 17.09, p < .01$

Testing of Hypotheses about Differences

- general case: analyses of variance (ANOVA)
- special case: two independent treatments, t-Test
- Null-Hypothesis (no effect of treatment) vs. alternative hypothesis (effect of treatment)
- Inference statistics: tests whether the result can occur if it is assumed that the null-hypothesis holds in the population (from which we tested a sample), tests whether differences in the mean of the dependent variable are significant
- Significance: with an error (typically 5 or 1 percent) we can conclude that the null hypothesis does not hold

Accept/Reject Hypothesis

$H_1 : \bar{x}_1 > \bar{x}_2$ $H_0 : \bar{x}_1 - \bar{x}_2 = 0$



Two Kinds of Errors

		In population	
		H_0	H_1
In sample	H_0	correct dec.	β error
	H_1	α error	correct dec.

t-Test for Independent Samples

- H_0: the means of two populations which are normally distributed (Gauss) and with homogenuous variances are not different ($\mu_1 = \mu_2$)
- Specific (directed) alternative hypothesis: $\mu_1 > \mu_2$
- If samples of size n are drawn from a normally-distributed population, the sample means are t-distributed (with $n - 1$ degrees of freedom because $n - 1$ means determine the last mean)

$$t = \frac{\bar{x}_1 - \bar{x}_2 - (\mu_1 - \mu_2)}{\hat{\sigma}_{\bar{x}_1 - \bar{x}_2}}$$

t-Test cont.

- Because null hypothesis $\mu_1 - \mu_2 = 0$

$$t = \frac{\bar{x_1} - \bar{x_2}}{\hat{\sigma}_{\bar{x_1} - \bar{x_2}}}$$

- Because population variance is unknown, it is estimated from the sample

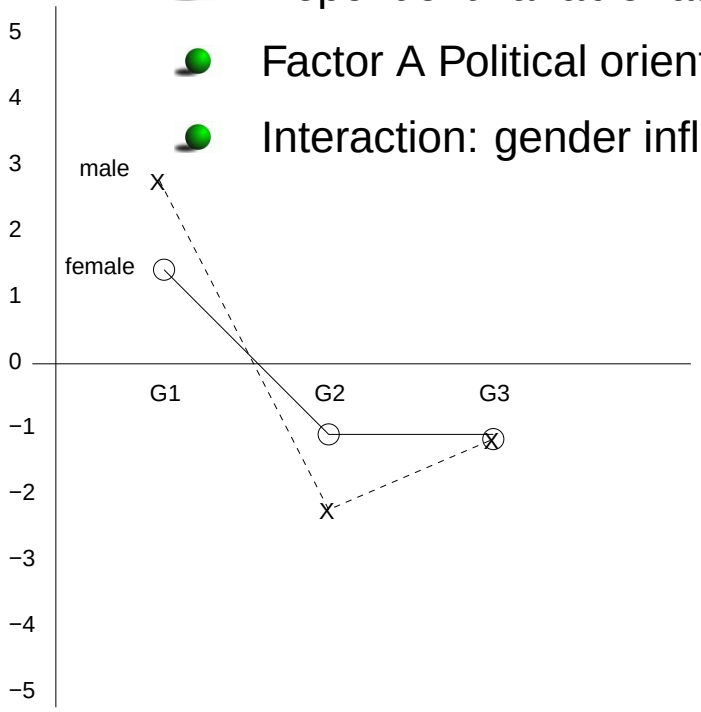
$$\hat{\sigma}_{\bar{x}_1 - \bar{x}_2} = \sqrt{\frac{\sum_{i=1}^{n_1} (x_{i1} - \bar{x}_1)^2 + \sum_{i=1}^{n_2} (x_{i2} - \bar{x}_1)^2}{(n_1 - 1) + (n_2 - 1)} \times (\frac{1}{n_1} + \frac{1}{n_2})}$$

Analysis of Variance

- More general than t-test
- For different designs
  - one factor with arbitrary number of groups
    e.g. factor with 4 groups
    source of noise: sea and wind, street, construction work, airport
  - multi-factorial
  - with repeated measurement
- F-distribution: H_0: no difference in the means, i.e. variance of means is zero
  $H_0 : \mu_0 = \mu_1 = \mu_p$ $H_0 : \sigma_\mu = 0$
  $H_1 : \sigma_\mu = c$
- Degrees of freedom (df) for one factor: $p - 1$

Multifactorial Analysis of Variance

- E.g., two-factorial: factor A with p steps, factor B with q steps
- Interaction: $A \times B$
- Example (quasi-experimental):
  - Dependent variable: attitude to a new law for parental leave (rating -5 to $+5$)
  - Factor A Political orientation (3 groups), Factor B Gender (2 groups)
  - Interaction: gender influence differs for the different political groups



Chi-Square Tests

- For categorial data
- e.g. two independent variables
  H_0: An attribute with k values is independent of an attribute with l values
- Example: Frequency of depression and schizophrenia is distributed differently in different social groups

Which Kind of Statistical Test

- Scale niveau of data
- Distribution of measurements
- Sample Size

Scale Niveau

- Nominal
- Ordinal: rank order
- Intervall: determined up to shift of zero and scaling $\alpha x + \beta$
- Rational: determined zero
- Absolute

Experimental Ethics

- Trade off between scientific progress and human dignity (e.g., threshold of pain, learned helplessness, ...)
- Obligation to inform subjects
- Free decision to participate/quit participation
- Anonymity of results

Experiments in social psychology (e.g. Milgram) vs. in cognitive psychology

Report of Empirical Studies

- Highly standardized
- General structure
  - Introduction
  - Theoretical Background
  - Experiment/Study
  - General Discussion

Introduction

- Section name: Introduction
- Give a short motivation for your study
- What is the general empirical question?
- Relevance of the question
- Need for the study/experiment
- Advanced organizer for rest of paper

Theoretical Background

- Section-name: specific for your study
- E.g.: Color Perception and Interpretation of Diagrams
- Describe state of research related to your study
- Argue why a (further) study is needed to give more precise information
- Formulate the hypothesis in a general way and explain how you derived this hypothesis

Experiment

- Section name: Experiment or Empirical Study
- Subsections
  - Design (operational hypothesis and resulting design: independent and dependent variables)
  - Method
    - Participants
    - Material
    - Procedure
  - Results
  - Discussion

Participants

The study was conducted in May 2004 with students from Bamberg university. 42 students participated in the study (23 male, 19 female, average age 21.3, sd = 2.4). 2 subjects were excluded from the analysis because of missing values.

Material

Four charts were drawn with XX (see fig. 1 to 4). All charts depicted the same imaginary technical device, called “Megafux”. Two charts represented temperatur distribution and two charts distribution of frequencies (factor: content type). For each content type, factor “representation type” was varied such that one chart color coded the distributed value directly on the device and the other chart gave values in a line graph.
For each chart, the subjects had to answer one question “The temperature/frequence is higher on the upper part than on the lower part of the Megaflux: yes/no”.
The information was presented in a webbrowser via http. The dependent variables were obtained via keystrokes using php.

Procedure

Subjects were tested individually. One experminental session took about five minutes time. First subjects read a webpage giving general instructions. Afterwards, the question was presented. The subject clicked the “OK” button if he/she was sure that he/she understood the question. Then a chart appeared. The subjects now clicked either yes or no and the answer together with the time from presentation of the chart until mouse click were stored.

Results

- Give descriptive statistics, such as: overall number of correct answers in percent, overall mean and standard deviation of times.
- Present results for hypothesis in a bar or line graph and give the results of the statistical test
- If you have controlled some possible extraneous variables, such as position of yes/no button, (gender, age), report their influence

Discussion

- Only now relate the findings to your hypothesis
- Is your hypothesis confirmed by the results?
- Are the alternative explanations for your result?
- If your hypothesis could be not confirmed, have you hypotheses about additional factors which have influenced the results? a more specific hypothesis?

General Discussion

- Give a short summary of your hypothesis and the main result
- What are the conclusions of the results (e.g. implication for information presentation)
- Which follow-up studies/experiments should be conducted?

User Study Methods

- Remember: user-centered approach to design
- Different data collection methods
  - Behavioral data: solution times, errors (high quality)
  - Subjective data: extracted from from interviews, observations, questionnaires

Interviews

- Can/should be conducted very early in design (no prototype available)
- unstructured (if you have really no idea) vs. structured
- Points to be covered
  - Explain the purpose of the interview
  - Enumerate activities to be supported
    e.g., form general questions to more specific questions
  - Explore work methods (hardest part, may be difficult to understand for an outsider)
  - Tracing interconnections
  - uncover performance issues

Analysis of Interview Data

- Recording of interviews get transcribed
- Extract categories for the different aspects of users' activities
- There are some software tools for interview analysis http://www.qualitative-research.net
- If necessary: try to make analysis more objective by using different raters which assign text segments to categories and evaluate interrater reliability

Observations

- Best used to observe task performance of users
- Best not interrupt users' activities (non-reactive approach)
- Data collection via
  - Video
  - Logfiles
  - Thinking aloud

Data from Observations

- As in interviews: qualitative data
- Be aware of Hawthorne effect: workers productivity might increase simply as response to being studied
- Most important is a careful design of tasks to be studied (representative, varying from routine to special, ...)
- Usability criteria should be translated into dependent variables!

Questionnairs

- Best used to assess additional aspects of system use which cannot be derived more directly
- E.g.: Subjective impressions as satisfaction, feeling in control
- Questionaire design issues
  - Items should be unambiguous
  - Answer modalities: yes/no, multiple choice, rating scales, open answers
  - Test theoretical questions: reliability, validity
    For tests which are developped for a broader application!