Course "Empirical Evaluation in Informatics"

Generic Empirical Method

Lutz Prechelt
Freie Universität Berlin, Institut für Informatik

1. Formulate goal and question
2. Select method, design study
3. Find or create observation context
4. Observe and collect data
5. Evaluate observations
6. Interpret results and draw conclusions

Example: N-version programming

"Empirische Bewertung in der Informatik"

Allgemeines Vorgehen für Empirie

Prof. Dr. Lutz Prechelt
Freie Universität Berlin, Institut für Informatik

1. Ziel und Frage formulieren
2. Methode auswählen und Studie entwerfen
3. Beobachtungssituation finden oder herstellen
4. Beobachten und Daten sammeln
5. Beobachtungen auswerten
6. Ergebnisse bewerten und Schlüsse ziehen

Bsp.: N-Versions-Programmierung

Generic method overview

How to conduct an empirical study:

1. Decide on ultimate goal
2. Formulate question for the study
3. Characterize the observations sought
4. Design the study
5. Find or create the observation context
6. Observe
7. Analyze observations
8. Interpret results

Not a phase model!
(rather holistic or iterative)

(1) Decide on overall goal

- Empirical studies rarely solve practical problems
  - Rather, they provide knowledge and understanding

- Each study is performed in a certain context
  - The **specific goal** of one study can only be fully understood within this context
  - The context defines an **overall goal**
    - (more typically a hierarchy of such goals)
  - The study must contribute to that goal
  - When designing a study, it is important to understand the overall goal well

Example context: High-reliability systems

Freie Universität Berlin

- **Overall goal:**
  Understand how best to produce ultra-reliable software

- Such software is relevant for systems that are
  - extremely expensive (e.g. Ariane rocket) or
  - life-critical (e.g. airplane control, nuclear reactor control)

- Proposed development approaches (most can be combined):
  - Super-intensive validation (testing)
  - Using super-high-level languages (executable specifications)
  - Program development by formal transformations
  - Mathematical program verification
  - **N-version programming**

Example topic: N-version programming

- Build multiple, very different implementations of the same program
  - Typically, 3 such "versions" are used

- Use them all in parallel, with identical inputs
  - If all is well, they will produce identical outputs
  - If not, apply voting to find the correct result
  - The N-version program will fail only when a majority of the versions fails at the same time
    - (assuming the voting has been implemented correctly!)
    - Hopefully, concurrent failures are rare.
    - Then the SW would be very reliable

(2) Formulate question for the study

- The **specific goal** for one study needs to be formulated precisely
  - Typically in the form of a **question** (or a few related questions) to be answered

- This question is the yardstick against which *credibility* and *relevance* of a study will be measured
  - If the question is vague, credibility will always be low
  - and relevance will be difficult to judge.
  - If the question is good, relevance is easy to see
    - if the study really answers the question convincingly.
    - Obtaining a satisfactory answer to the question must be realistic.

Example research question: Independence of failure

- The reliability of an N-version program will be the better, the less correlated the different versions' failures are
  - It will be ideal if the versions' failures are statistically independent (i.e. not correlated at all)
  - N-version proponents often assume this independence

- A good specific study question could be:
  - *"Are the failures of the versions within an N-version program indeed statistically independent or not?"*

Generic method overview

How to conduct an empirical study:

1. Decide on ultimate goal
2. Formulate question for the study
3. **Characterize the observations sought**
4. **Design the study**
5. Find or create the observation context
6. Observe
7. Analyze observations
8. Interpret results

(3) Characterize the observations sought

- First step in the design of the actual study:
  - What information do I need for answering the question?

- Determine:
  - The kind of information
  - The amount, precision, and reliability of information

- If the question of the study is complex, it can be quite difficult to understand what information will be needed
  - In particular, the sensitivity of the study (and hence the amount and precision of information required) can often be estimated only very roughly

Example specific goal: Testing the independence of failures

Freie Universität Berlin

- It is plausible that the assumption of independence of failures is wrong:
  - Argument: Some programming mistakes are due to intricacies of the problem and will tend to occur more frequently than random mistakes

- Thus, we seek observations of the following kind:
  - We thoroughly apply N-version programming
  - We measure the relative frequency of concurrent failures
  - We expect to find more of these than should happen if independence of failure was true
    - To check this, we need to know the correct output in each case
    - To make sure the effect is clear, we should use a lot more than 3 versions

(4) Design the study

- Once we understand what information we need, we can select an appropriate empirical method:
  - Benchmark
  - Controlled experiment
  - Quasi-experiment
  - Case study
  - Survey
  - Literature study, simulation, meta-study, etc.

- The details of the study design process vary a lot from one method to the other
  - Will be described in subsequent lectures
  - Is complex: A whole course could easily be taught only on the details of any one method alone

Example study design: N-version experiment

Freie Universität Berlin

- The independence of failures in N-version programs was investigated by John Knight and Nancy Leveson

- They chose the form of an existence proof (with high control):
  - **Subjects:** Students with good programming experience
    - from two different universities (to increase diversity)
  - **Task:** The "launch interceptor" problem
    - A high-quality specification is available
    - Input: a set of coefficients describing radar reflections plus a set of configuration parameters
    - Output: 241 boolean values (15*15 plus 15 plus 1)
  - **Procedure:** Each subject creates his/her own version; explicitly totally independent of the others
    - Subjects are asked to produce best possible quality, must test extensively
    - Receive 15 sets of inputs/outputs for debugging

Example: N-version experiment design (2)

- Experiment design (continued):
  - Each submitted program version has to pass an acceptance test
    - 200 random test cases, different for each program
      - to avoid fixed types of failures slipping through
    - comparison to output of a *"gold"* program (believed correct)
  - Accepted programs undergo heavy usage simulation
    - executed for 1 million inputs each
      - equivalent to the whole life span of real use
    - measure all failures of all versions
    - compute expected and actual number of concurrent failures

Generic method overview

How to conduct an empirical study:

1. Decide on ultimate goal
2. Formulate question for the study
3. Characterize the observations sought
4. Design the study
5. **Find or create the observation context**
6. **Observe**
7. Analyze observations
8. Interpret results

(5) Find or create the observation context

- Once the study has been designed, we need participants
- Often, though not always, these are human subjects
  - Sometimes it might be programs etc.
- They need to have or acquire the prerequisite knowledge
  - Empirical studies may involve giving specific training
  - Participants should be from the original target population of whatever it is that we study
- We must be able to characterize them
  - e.g. their knowledge, experience, motivation, constraints
  - else generalizability (and hence relevance) will be unclear
- Finding people who are both competent and willing to cooperate can be extremely difficult
  - This is one of the main difficulties for empirical studies in software engineering

(5) Find or create the observation context (c'd)

- We usually need to explain the study to the participants

- This can be difficult if
  - the study involves important but unusual constraints,
  - the participants are not well motivated, or
  - the study objective must be kept secret
    - because knowing it would spoil the study

- We need to provide the participants with
  - a working environment
  - the required input materials
  - perhaps guidance
  - perhaps supervision

(5) Find or create the observation context (c'd)

Freie Universität Berlin

- We need to provide the **measurement infrastructure**
  - Make sure no data is lost
  - Make sure measurements are precise, correct, and sufficiently detailed
  - Measurement should be unintrusive
  - Measurement should be robust against unwelcome events
- Details are much different between different [kinds of] studies
  - in particular if the participants are not human
    - e.g. in retrospective studies of existing data or for many types of benchmarking
- All study designs and implementations need **pilot testing** and several rounds of **improvement**
  - much like software

Example observation context: N-version experiment

Freie Universität Berlin

- 27 students with good programming experience
  - graduate and senior level
  - 9 from U Virginia, 18 from UC Irvine
- received an introduction to N-version ideas
- received the specification of the "launch interceptor" task
  - clarifications handled by email
- work alone; using their own methods, tools etc.
  - important to make sure no interaction occurs
- acceptance test is administered by the experimenter
- no measurement is required during the experiment
  - all measurement is part of the analysis phase

(6) Observe

- Once the study has been started, the actual data collection is going on
- The format of this is very different for different kinds of studies:
  - Benchmarks and experiments: measure various dependent variables
  - Case studies: measure quantitative variables and collect diverse qualitative observations
  - Surveys: Actively interview people and collect individual answers OR just sit back and wait until filled-in questionnaires arrive
  - etc.
- If design or implementation of the study are bad, all you will get is garbage

Example: Observations for NVP experiment

- The N-version experiment is an existence proof experiment
  - A rare form of controlled experiment:
  1. There is no comparison group
     - Rather, the 27 individual implementations will be compared
       - Each implementation represents a group, so-to-speak.
  2. In the NVP case, there are almost no observations going on while the experiment runs
     - Only the results of the acceptance tests
     - All relevant observations are made on the submitted programs after the end of the actual experiment

- Quite unusual!

Generic method overview

How to conduct an empirical study:

1. Decide on ultimate goal
2. Formulate question for the study
3. Characterize the observations sought
4. Design the study
5. Find or create the observation context
6. Observe
7. **Analyze observations**
8. **Interpret results**

(7) Analyze observations

- Once the observation stage of the study is over, we analyze the data we collected in order to answer the study question
  - Analysis may start during the observation stage already

- Quantitative data is analyzed by applied statistics
  - Initially: exploratory data analysis
    - using e.g. descriptive statistics and visualization
  - If we know exactly what we are looking for: inferential statistics

- Qualitative data is analyzed by qualitative research methods
  - e.g. Protocol Analysis or Grounded Theory Methodology (GTM)
  - this is beyond the scope of this course

(7) Analyze observations (cont'd)

- If our study design and conduct were good, our data should contain the answer to the study question
  - And appropriate analysis should produce the answer
    - Often the answer is not as clear as one would like
  - The analysis often gets more complicated than expected
    - e.g. because the data are dirty
- If the answer cannot be found, we either
  - have made a mistake
    - usually in the study design
  - or were unlucky

Example analysis: Failures in the N versions

Freie Universität Berlin

- Again, the N-version experiment is unusual in this respect:
  - The analysis is straightforward
  - And the answer obtained is extraordinarily clear

(Failures are determined by comparing each program's output to the gold program.
All programs now use the same 1 million test cases.)

Version	Failures	Pr(Success)
1	2	0.999998
2	0	1.000000
3	2297	0.997703
4	0	1.000000
5	0	1.000000
6	1149	0.998851
7	71	0.999929
8	323	0.999677
9	53	0.999947
10	0	1.000000
11	554	0.999446
12	427	0.999573
13	4	0.999996
14	1368	0.998632

etc.

Example: frequency of concurrent failures

- How often more than one program failed on any of the n = 1 000 000 test cases:

Number	Probability	Occurrences
2	0.00055100	551
3	0.00034300	343
4	0.00024200	242
5	0.00007300	73
6	0.00003200	32
7	0.00001200	12
8	0.00000200	2

- K = 1255 <u>concurrent</u> failures overall
  - where two or more versions fail on the same test case

Example: Assumption of independence

- Given failure probability p_i of each version, the probability that no version fails (in any one test, if failure is independent) is:

$$P_0 = (1 - p_1)(1 - p_2)\ldots(1 - p_N)$$

- Probability that exactly one version fails:

$$P_1 = \frac{P_0 p_1}{1 - p_1} + \frac{P_0 p_2}{1 - p_2} + \cdots + \frac{P_0 p_N}{1 - p_N}$$

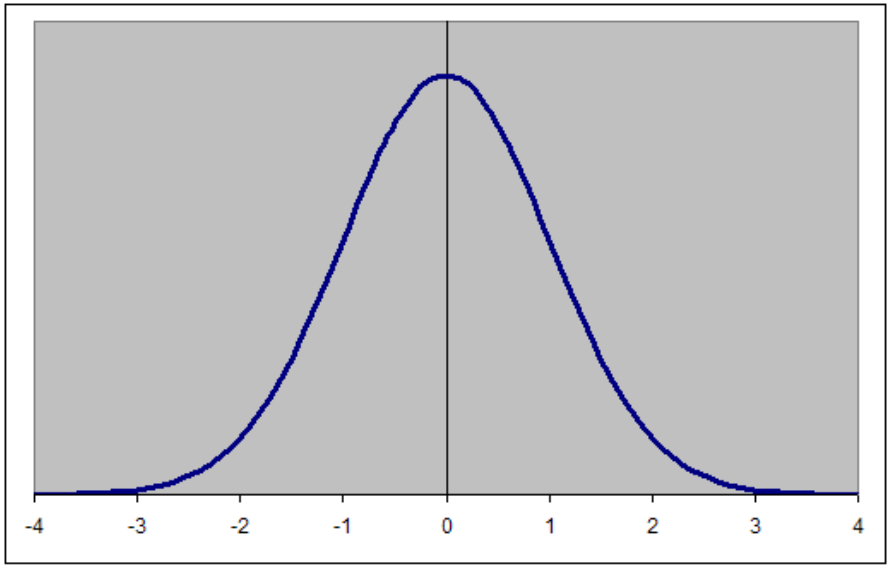
- Probability that more than one version fails:

$$P_{more} = 1 - P_0 - P_1$$

- If independence is true, z will be $N_{0,1}$-distributed:

$$z = \frac{K - nP_{more}}{(nP_{more}(1 - P_{more}))^{1/2}}$$

- In our case: z = 100.51



(P(z < 2.33) = 0.99)

(8) Interpret results

- After analyzing the data, we need to draw conclusions:
  - What do we now know?
  - What not?
  - What can we expect from generalizing the results?
  - What further empirical studies should be done in order to complete the understanding?

- Again, the form of these conclusions and how to derive them is very different depending on method and study

Example conclusions: The independence assumption

- Immediate **result**: The probability of getting such a number of concurrent failures if the failures occurred independently in our experiment is far lower than 1%
- **Conclusion**: In this setting, the assumption of independence was violated
- Further **conclusion**: Reliability conclusions based on the assumption of independence might be too optimistic
- **Conjecture**: Independence of failure is not typically the case is N-version programs
  - N-version programming helps
  - but not as much as one might have hoped

- **Suggestion**: The assumption should be further investigated before critical decisions are based on calculations that used it

Literature

- John Knight, Nancy Leveson: *"An experimental evaluation of the assumption of independence in multi-version programming"*, IEEE Transactions on Software Engineering, January 1986

- Knight, Leveson: *"A Reply to the Criticisms of the Knight and Leveson Experiment"*, ACM Software Engineering Notes, January 1990
  - The validity of the experiment has been attacked seriously,
    - but the attacks are themselves invalid.
  - This is a rebuttal of these attacks and is an extremely interesting read
    - amusing, too.

Summary

For performing an empirical study, one needs to:

- Understand and formulate exactly
  what one intends to find out
- Design the study:
  General method, concrete approach, and setup
- Find or create the setting in which to observe
- Observe and record the observations as data
- Analyze the data
- Interpret the results and draw conclusions

Thank you!