veritas
iustitia
libertas
Freie Universität Berlin

Seminar: Empirische Forschungsmethoden in der Softwaretechnik

Einführung

Stephan Salinger

Vorbemerkungen und Organisatorisches (1)

- Sebastian Jekutsch, Stephan Salinger: R008
- Termine
  - KW14 (29.03.-02.04.2004) o. KW15 (05.04 – 08.04.2004)
  - 9:15 bis ca. 13:00Uhr und 14:15 bis ca. 18:00Uhr
  - Raumnummer wird ausgehängt
- Informationen
  - Mailingliste: se_s_empirie@lists.spline.inf.fu-berlin.de
  - Homepage: http://www.inf.fu-berlin.de/inst/ag-se/teaching/S-EMPIRIE-2004/
  - TWiki: http://projects.mi.fu-berlin.de/w/bin/view/SE/

Vorbemerkungen und Organisatorisches (2)

- Ablauf
  - Anmeldung
  - Vorbesprechung und Themenvergabe
  - Jeder Teilnehmer (TN) hält einen Vortrag (60 -70 Minuten)
  - Jeder Teilnehmer ist Schriftgutachter für zwei weiter TN
    - Begutachtung von Folien und Ausarbeitung
    - TN schickt spätestens 10 Kalendertage vor Vortrag Folien und Ausarbeitung an Schriftgutachter
    - Schriftgutachter schickt spätestens 7 Tage nach Erhalt Begutachtungen zurück
  - Jeder Teilnehmer ist Vortragsgutachter für zwei weitere TN
    - Vortragsgutachter macht handschriftliche Notizen zu bestimmten Punkten während des Vortrags
  - Details siehe im TWiki

Vorbemerkungen und Organisatorisches (3)

- Scheinkriterien
  - Halten eines Vortrages
    - Verständlichkeit und Korrektheit der Darstellung
    - Interessante Darstellung
    - Kritische und konstruktive Auseinandersetzung mit dem Artikel
    - Ggf. über eine über die in den Artikeln hinausgehende Einführung in die in den Experimenten untersuchte Materie
  - Erarbeiten einer Ausarbeitung
    - Einreichen beim Veranstalter bis spätestens 7 Kalendertage nach Vortrag (PDF).
    - Soll über die einfache Übersetzung und Umformulierung hinausgehen
    - Eigene Meinung muss erkennbar sein
    - Muss auf den Schwerpunkt der Veranstaltung fokussiert sein
    - Abstraktion von diesbezüglich weniger interessanten Details
  - 2 x Schriftgutachter, 2 x Vortragsgutachter
  - Anwesenheit und Beteiligung bei allen Terminen

Bibliographie

- Lutz Prechelt
  Kontrollierte Experimente in der Softwaretechnik, Potential und Methodik
  Springer-Verlag Berlin Heidelberg. 2001
- Dewayne Perry, Adam Porter, Lawrence Votta
  Empirical Studies of Software Engineering: A Roadmap
  In Proceedings of the conference on the future of software engineering, pages 345-355, 2000
- Walter F. Tichy
  Should Computer Scientists Experiment More?
  IEEE Computer, IEEE Computer Society Press, pages 32-40, 1998
- L.B. Christensen
  Experimental Methodology,
  sixth ed. Needham, Heights, Mass.: Allyn and Bacon, 1994.

Ziele des Seminars (1)

- Einführung in die empirischen Verfahren in der SWT
- Kennen lernen von bestimmten empirischen Forschungsmethoden
  - Kontrollierte Experimente
  - Umfragen
- Berücksichtigung unterschiedlicher inhaltlicher Rollen (Wirkungen) von empirischen Verfahren in Forschungsprogrammen
  - Hypothesentest
  - Erkundung
  - Quantifizierung von Effekten
  - Modellbildung
  - Absicherung von Erkenntnissen
  - Verbreiterung von Erkenntnissen

Ziele des Seminars (2)

- Kennen lernen von typischen (und untypischen) Themengebieten der empirischen Forschung in der SWT
  - Inspektionen
  - Objektorientierung
  - Personal Software Process (PSP)
  - ...
- Fokus der Betrachtungen:
  - Experimententwurf, -implementierung und -durchführung
  - Stärken und Schwächen im Vorgehen
- *Nicht* im Mittelpunkt steht:
  - Methoden zur Datenerfassung
  - Methoden zur Datenvalidierung
  - Datenanalyse/Statistik

Ziele dieser Präsentation

- Klärung der folgenden Fragestellungen:
  - Was ist Empirie und was bedeutet Empirie in der SWT?
  - Wie grenzen sich sich die unterschiedlichen Forschungsmethoden voneinander ab?
  - Wie sehen die Rollen von Experimenten aus (und wie spielen sie zusammen)?
  - Wie kann man die Güte von Experimenten klassifizieren?
- Im Fokus steht nicht:
  - Beschreibung der unterschiedlichen Auswertungsmethodiken (Datenerfassung, -validierung und Analyse)
  - Statistik

Zusammenfassung Ziele

- **Die heutige Präsentation soll folgende Fragen klären:**
  - Was ist Empirie bzw. Empirie in der SWT?
  - Welches sind die zentralen Begriffe?

- **Das gesamte Seminar soll Einblicke in die Beantwortung folgender Fragen liefern:**
  - Welche systematischen Vorgehensmodelle gibt es?
  - Welche systemischen Probleme existieren?
  - Wie können bessere Studien erstellt und durchgeführt werden?

Inhalt

1. Empirie in der SWT
2. Überblick empirische Forschungsmethoden
3. Zentrale Begriffe und Prinzipien
4. Rollen kontrollierter Experimente
5. Wann ist ein Experiment gut?

Inhalt

1. **Empirie in der SWT**
2. Überblick empirische Forschungsmethoden
3. Zentrale Begriffe und Prinzipien
4. Rollen kontrollierter Experimente
5. Wann ist ein Experiment gut?

Empirie in der SWT: Ausgangsposition

- 1994: Untersuchung von Tichy, Lukowicz, Prechelt und Heinz:
  - 40% der wissenschaftlichen Untersuchungen in der Informatik, die eine empirische Auswertung aufweisen müssten, besitzen eine solche nicht.
  - In der SWT sogar 50%
- D.h.: Bislang werden Beiträge (in der SWT) oftmals kaum oder gar nicht auf ihre Nützlichkeit untersucht.

Empirie in der SWT: Bsp. für Fragestellungen (1)

Entwurfsmuster

- Als Vorteile von Entwurfsmuster werden die folgenden Eigenschaften postuliert:
  - Entwickler lernen schnell besseres Design
  - Entwickler werden produktiver
  - Qualität der Software wird besser
  - Kommunikation zw. Entwicklern wird besser
  - Kommunikation bei der Wartung wird besser
- Prechelt, Unger, Phillipsen und Tichy untersuchten (1997) folgende spezielle Fragestellung:
  - Ist ein Programm, welches ausdrücklich die Benutzung von Entwurfsmustern im Programmtext beschreibt besser wartbar als ein Programm, welches Entwurfsmuster nur stillschweigend benutzt.

Empirie in der SWT: Definition Empirie

- **Empirie**

  griech.: Erfahrung, auf Beobachtung beruhend

- **Empirie (empirical experience)**

  Bezeichnung für Erkenntnisse, die auf *Erfahrungen* beruhen. Empirische Beobachtungen oder Aussagen beziehen sich auf *Wahrnehmungen* und/oder sind von solchen abgeleitet. Die Methode der Theoriebildung wird *Induktion* genannt.

- **Induktion**

  In der Logik und Mathematik versteht man unter Induktion die *Schlussfolgerung von Einzelfällen auf das Allgemeine*, aber auch das *Schlussfolgern von Beobachtungen auf Gesetzmäßigkeiten* (Gegenbegriff: Deduktion).

Empirie in der SWT: Definition Empirie in der SWT (1)

Der Begriff Empirie umfasst in der SWT:

- Gesamtheit aller *Forschungsmethoden* zur Beobachtung des Verhaltens oder der Wirkung softwaretechnischer Artefakte
- Gesamtheit aller *Anwendungen dieser Forschungsmethoden*
- Das wissenschaftliche Prinzip, sich, wo nötig, zum Wissensgewinn auf tatsächliche Ereignisse und Beobachtungen zu stützen

Empirie in der SWT: Definition Empirie in der SWT (2)

Empirische Bewertung bedeutet in der SWT:

- *Praktische* Verwendung und Erprobung eines Werkzeuges, einer Methode oder eines Modell um deren Eigenschaften zu *verstehen* und *beschreiben* zu können.

Im Gegensatz dazu steht die **spekulative Bewertung**:

- Durch mehr oder weniger stringentes logisches Schließen werden die erwarteten Eigenschaften auf Basis von mehr oder weniger plausiblen und größtenteils unausgesprochenen Annahmen ohne Empirie hergeleitet.

Empirie in der SWT: Forschungsansätze in der SWT (1)

- Grundsätzlich muss Forschung in der SWT *erfahrungsgeleitet* sein. Hierbei gibt es zwei unterschiedliche Weisen vorzugehen.
- **Ingenieurmäßiger Ansatz** in der SWT-Forschung
  - Zielt unmittelbar auf die Erschaffung von Artefakten (softwaretechnische Methoden, Softwarewerkzeuge), die für einen Praktiker nützlich sind (*ingenieurmäßiger Systembau*).
  - Im Mittelpunkt steht die Realisierung solcher Systeme.
  - Es werden Annahmen darüber gemacht, welche Eigenschaften nützlich wären und wie diese zu erreichen sind.
  - Die Validierung der Ergebnisse erfolgt durch den Markt.
  - Nützlich ist ein System, wenn es von (anderen) Forschern oder von Praktikern aufgegriffen und weiterentwickelt oder benutzt wird.

Empirie in der SWT: Forschungsansätze in der SWT (2)

- **Ingenieurmäßiger Ansatz** in der SWT-Forschung
  - ☺Führt im Erfolgsfalle zu raschen Fortschritten
  - ☹Oft ist in der SWT unklar, wie ein nützliches Werkzeug oder eine nützliche Methode konkret aussehen muss
  - ☹Bei der Flut von Forschungsbeiträgen kann es passieren, dass ein nützlicher Beitrag vom Markt übersehen wird

Empirie in der SWT: Forschungsansätze in der SWT (3)

- **Wissenschaftlicher Ansatz** in der SWT-Forschung
  - Zielt auf den Aufbau von Wissen, das für Praktiker bei der Anwendung der Softwaretechnik nützlich ist.
  - Dieses Wissen hat die Form von *Modellen*, die beschreiben
    - welche Wirkungen bestimmte Methoden oder Werkzeuge *unter gewissen Umständen* haben und
    - wie diese Wirkungen zusammenspielen.

- Das **math. Modell** ist eine idealisierte Beschreibung der Realität. Die Modelle dienen der Erklärung, Prognose und Simulation von Sachverhalten.
- Ein **Modell** ist eine Abstraktion des betrachteten Realitätsausschnittes.

Empirie in der SWT: Forschungsansätze in der SWT (4)

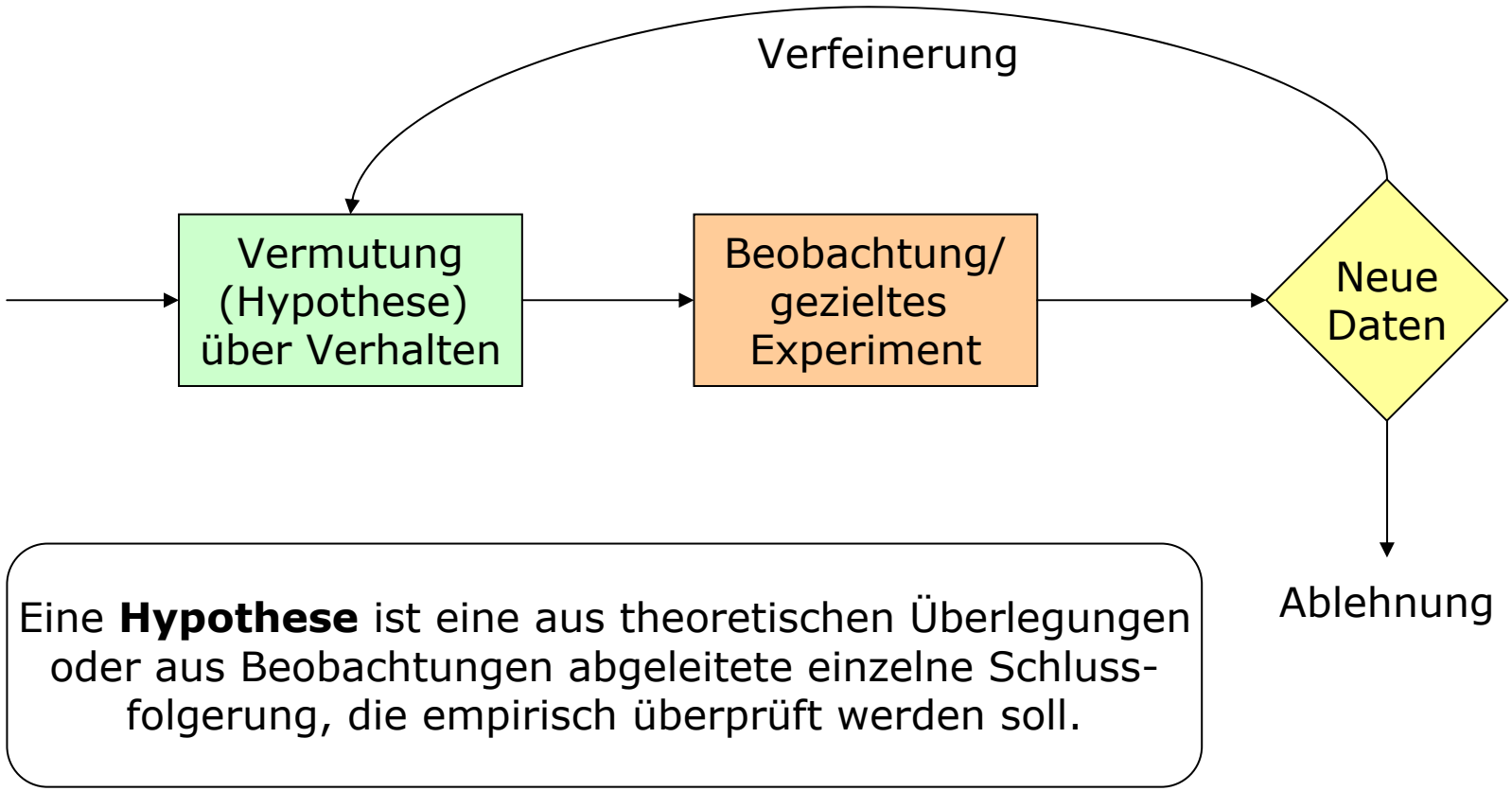
- **Wissenschaftlicher Ansatz** in der SWT-Forschung

  Aufgaben:

  - Validierung des ingenieurmäßigen Ansatzes:
    - Untersuchung von Werkzeugen, Sprachen und Methode
  - Ermittlung und Verbindung von Erkenntnissen, die über das Bewerten einzelner Methoden oder Werkzeuge hinausgehen:
    - Die eigentlichen Modellbildung.

Empirie in der SWT: Forschungsansätze in der SWT (5)

- **Wissenschaftlicher Ansatz** in der SWT-Forschung
  - benutzt Zyklus aus den Naturwissenschaften:



Eine **Hypothese** ist eine aus theoretischen Überlegungen oder aus Beobachtungen abgeleitete einzelne Schluss-folgerung, die empirisch überprüft werden soll.

Empirie in der SWT: Forschungsansätze in der SWT (6)

- **Achtung:**
  - Lässt sich eine Hypothese aufgrund der empirischen Überprüfung als "wahr" entscheiden, spricht das für die zugrunde gelegte Theorie oder Beobachtung – die Theorie ist damit aber nicht "bewiesen".
  - Der Wahrheit kann man sich mittels empirischer Untersuchungen nur *annähern*.

Empirie in der SWT: Rolle der Empirie in der SWT

- In der SWT erfolgt der Wissenszuwachs überwiegend durch Empirie, da die betrachteten Prozesse zu komplex für mathematische Analysen sind und sich ohnehin nur unvollständig abstrakt erfassen lassen.
- Insbesondere ist die Entwicklung von Modellvorstellungen in der SWT nur durch Empirie möglich.
- Fortschritte in der SWT gehen überwiegend vom Systembau aus (gute Ideen gepaart mit Intuition).
- Systembau liefert eine Zahl von Antworten auf Fragen der Form „Wie löse ich das Problem X?“.
- Systembau beantwortet nicht:
  - Welche der vorgeschlagenen Methoden haben welche Vor- und Nachteile?
  - Wie kommen diese Vor- und Nachteile zustande?

Inhalt

1. Empirie in der SWT
2. **Überblick empirische Forschungsmethoden**
3. Zentrale Begriffe und Prinzipien
4. Rollen kontrollierter Experimente
5. Wann ist ein Experiment gut?

Forschungsmethoden: Überblick

- **Forschungsmethoden**
  1. Fallstudie und Benchmarking
  2. Feldstudie
  3. **Kontrolliertes Experiment**
  4. **Umfrage**
  5. Metastudie

Forschungsmethoden: Eigenschaften (1)

- Forschungsmethoden lassen sich bzgl. **Qualitäts- und Eignungsaspekten** beschreiben (d.h. wann kann eine Methode im Einzelfall günstiger als eine andere sein):
  1. Aufwand und Kosten
  2. Stärke des Eingriffs
  3. Eher für qualitative oder eher für quantitative Forschung geeignet?
  4. Verlässlichkeit der Resultate
  5. Beschreibbarkeit und Verstehbarkeit der Forschung
  6. Reproduzierbarkeit der Forschung und der Resultate
  7. Verallgemeinerbarkeit der Resultate

Forschungsmethoden: Eigenschaften (2)

- **Forschungsmethoden oder einzelne Studien lassen sich bzgl. Strukturaspekten beschreiben (Eigenschaften die die unterschiedliche Durchführung charakterisieren)**
  1. Kontrolle
     - Kein Einfluss auf Bedingungen , aus denen das Beobachtete entsteht -> beliebige Manipulation
  2. Eingriffsstärke
     - Keine merkliche Veränderung der Arbeitsbedingungen der Versuchsperson -> totales Diktat
  3. Genauigkeit
     - Direkte, objektive, vollständige und genaue Beobachtung und Messung -> die Beobachtungen sind subjektiv, unvollständige Eindrücke aus zweiter Hand in qualitativer Form
  4. Relativität
     - Beobachtung nur einer Sorte von Bedingungen -> Vergleich unterschiedlicher Bedingungen
  5. Replikation
     - Einmalig -> häufige Durchführung einer beobachteten Tätigkeit

Forschungsmethoden: Kontrolliertes Experiment (1)

Ein **kontrolliertes Experiment** ist eine Studie, bei der

- alle voraussichtlich für das Ergebnis relevanten Umstände (**(Stör-)Variablen**) konstant gehalten werden (**Kontrolle**),
- mit Ausnahme von einem oder wenigen, die den Gegenstand der Untersuchung bilden (**Experimentvariablen**).
- Die Beobachtungen (**abhängigen Variablen**) für verschiedene gezielt ausgesuchte Werte der Experimentvariablen (**unabhängige Variablen**) werden miteinander verglichen, um so zu reproduzierbaren Aussagen zu kommen, die eine vor dem Experiment definierte Experimentfrage beantworten.
- Die **Experimentfrage** ist ein genügend enge Aspekt einer relevanten Forschungsfrage.

Forschungsmethoden: Kontrolliertes Experiment (2)

Beispiel Experimentvariable/ unabhängige Variablen

- Untersuchung des Unterschiedes der Verständlichkeit zw. Flussdiagrammen und Programmcode für drei unterschiedliche Programmgröße:
  - Betrachtung von zwei unabhängigen Variablen (Darstellungstyp und Programmgröße) in 6 verschiedenen Versuchsbedingungen

Beispiel abhängige Variable

- Untersuchung welches von zwei Entwurfsverfahren für eine gegebenes Entwurfsproblem das besserer ist:
  - Messen Entwurfsaufwand (Zeit) und Qualität (Anzahl Fehler) der entstehenden Entwürfe.

Forschungsmethoden: Kontrolliertes Experiment (3)

Aspekte von kontrollierten Experimenten

- **Verstehbarkeit/Vertrauen:**

  Hoher Grad von *Kontrolle* über die Beobachtungsbedingungen (Festlegung und Überwachung Variablen und Arbeitsbedingungen)

  => Höchster Grad an Vertrauen in die Beobachtung

- **Kosten:**

  Ausgleich von individuellen Variationen in der Arbeitsweise und Leistung der Teilnehmer führt zur Vergabe der *selben* Aufgabe an mehrere Personen (Versuchs- und Experimentgruppe) und nachfolgender statistischer Mittelwertbildung

  => Kann in eine SW-Prozess selten produktiv genutzt werden
  => relativ teuer!

Forschungsmethoden: Kontrolliertes Experiment (4)

Aspekte von kontrollierten Experimenten

- **Reproduzierbarkeit:**

  Genaue Festlegung aller Randbedingungen bzw. Konstanthalten vieler Aspekte (zur Kontrolle der übrigen Variablen)

  => hohe Reproduzierbarkeit der Ergebnisse
- **Stärke des Eingriffs**

  Fast immer werden die beteiligten Softwareingenieure total aus ihrer normalen Arbeit herausgenommen (künstliche Umgebung)

  => starker Eingriff

- Und was ist mit **Verallgemeinerbarkeit**?

Forschungsmethoden: Umfrage (1)

Bei einer **Umfrage**

- beantworten die Mitglieder (**Umfrageteilnehmer**) einer
- geeignet ausgewählten Personengruppe (**Zielgruppe**)
- mehrerer von den Forschern sorgfältigformulierte Fragen.
- Die Fragen können sich auf **subjektive oder objektive Sachverhalte** beziehen, sind selbst aber immer subjektiv und nur begrenzt überprüfbar.
- Die Beantwortung kann schriftlich (**Fragebogen**) oder mündlich (**strukturiertes Interview**) erfolgen.
- Die Antworten können zur **quantitativen Ausrichtung** der Umfrage schematisch erfolgen (z.B. Multiple-Choice-Verfahren) oder aber in der Form frei sein, was zur **qualitativen Forschung** führt.

Forschungsmethoden: Umfrage (2)

Beispiele für Umfragen:

- Herbsleb und Goldenson:

  Umfrage zur Bewertung einer Methode: Befragung von Benutzern des Prozessreifemodells CMM nach Erfahrungen und erzielten Ergebnissen.

  Ziel war es festzustellen, wie erfolgreich CMM in der Praxis ist.

- Daly:

  Befragung über die Benutzung von Vererbung in objektorientierten Sprachen.

  Die Auswertung ergab, dass Vererbung grundsätzlich hilfreich ist aber ab tiefe fünf zu Problemen führt.

  Daly und Kollegen führten dann zur Prüfung ein kontrolliertes Experiment durch.

Forschungsmethoden: Umfrage (3)

Umfragen sind für Querschnittsuntersuchungen geeignet:

- Wie verbreitet ist eine gewisse Vorgehensweisen?
- Welche Vor- und Nachteile habe bestimmte Vorgehensweisen (zumindest subjektiv) und welche Effekte treten bei ihrer Anwendung auf?
- Wie erfolgreich sind gewisse Vorgehensweise?
- Wie häufig sind bestimmte Randbedingungen und Anforderungen?

Forschungsmethoden: Umfrage (4)

Aspekte von Umfragen

- **Aufwand und Kosten:**

  Geringer Aufwand für die Teilnehmer und einfache Infrastruktur

  => relativ einfach und billig
- **Verlässlichkeit/Vertrauen:**

  Die Antworten dürfen nur sehr vorsichtig interpretiert werden, da sie meist *ungenau*, immer *subjektiv* und manchmal sogar *absichtlich verfälscht* sind.

  =>Verlässlichkeit ist unklar

Forschungsmethoden: Umfrage (5)

Weitere Problem bei Umfragen:

- Die in den Fragen benutzen Begriffe müssen von allen Teilnehmern und von den Forschern auf die gleiche Weise ausgelegt werden.
- Oft liegt der Rücklauf bei weit unter 100%, so dass sich die Frage der Repräsentativität stellt.

Inhalt

1. Empirie in der SWT
2. Überblick empirische Forschungsmethoden
3. **Zentrale Begriffe und Prinzipien**
4. Rollen kontrollierter Experimente
5. Wann ist ein Experiment gut?

Zentrale Begriffe: Störvariable (1)

Zweck eines kontrollierten Experimentes besteht darin, *Beobachtungen zu machen, deren Ursachen eindeutig festliegen*:
Wenn etwas zweimal gemacht wird und dabei alle Umstände bis auf einen gleich sind, dann müssen evtl. Unterschiede in den Ergebnissen von der Änderung dieses einen Umstandes herrühren.
Die einzelnen Parameter werden dabei wie folgt bezeichnet**:**

- **Abhängige Variable:** Die im Laufe des Experimentes gemessenen Größe**.**
- **Unabhängige Variable:** Die im Laufe des Experimentes manipulierte Größe**.**
- **Störvariable:** Die im Experiment konstant gehaltenen Variablen**.**

Zentrale Begriffe: Störvariable (2)

Für Störvariablen gilt:

- Im Prinzip sollen alle Störvariablen in einem kontrollierten Experiment *gleichgehalten* werden. Dieser Anspruch ist aber unrealistisch, denn:
  - Man kann niemals den genau gleichen Zustand der ganzen Welt herstellen.
- Das Festhalten aller Störvariablen ist aber auch nicht wichtig. Festgehalten werden müssen nur die die *relevanten Störvariablen*.
  - Nur welche Störvariablen sind die relevanten?
- Glücklicherweise ist das Bekanntsein gar nicht notwendig, da sich fast alle Störvariablen mit derselben (aus zwei Teilen bestehenden) Technik kontrollieren lassen und diese Technik in einem softwaretechnischen Experiment ohnehin angewendet wird.

Zentrale Begriffe: Störvariable (3)

Replikation und Randomisierung:

- Es wird nicht das Verhalten einer einzelnen Versuchsperson für jeden Wert der unabhängigen Variablen verglichen, sondern immer eine **ganze Gruppe unter Betrachtung deren mittleren Verhaltens**.
- Die Mitglieder jeder Gruppe werden aus den verfügbaren Versuchspersonen **zufällig ausgewählt**.
  - Somit wird ein möglicher systematischer Fehler (die Störvariable X verändert unbekannterweise das beobachtete Ergebnis) in einen statistischen Fehler verwandelt (zufällig war die Störvariable X für die verglichenen Gruppen nicht genau gleich), der mit statistischen Techniken quantifiziert und beherrscht werden kann.
- Jede solche Gruppe heißt **Versuchsgruppe** oder **Experimentgruppe**. Zusätzlich kann es eine zweite Gruppe geben, die auf „normale" Weise arbeitet: **Kontrollgruppe** oder **Vergleichsgruppe**.

Zentrale Begriffe: Störvariable (4)

Wichtiger Aspekt der Replikation und Randomisierung:

- Eine der wichtigsten Störvariablen in softwaretechnischen Experimenten ist die unterschiedliche Kompetenz der Versuchspersonen (**individuelle Variation**).
- Die Kontrolle der individuellen Variation ist die wichtigste Voraussetzung für ein gelungenes Experiment.

Zentrale Begriffe: Störvariable und innere Gültigkeit (1)

Innere Gültigkeit:

- Die **innere Gültigkeit** (interne Gültigkeit, internal validity) eines kontrollierten Experimentes ist der Grad, in dem die Änderungen in den Werten der abhängigen Variablen tatsächlich wie gewünscht nur auf Änderungen in den unabhängigen Variablen zurückzuführen sind.
  - D.h. wie gut letztendlich alle relevanten Störvariablen kontrolliert wurden.
- Kurz: Die innere Gültigkeit beschreibt, wie gut die Kontrolle der Störvariablen in einem Experiment war.

Zentrale Begriffe: Störvariable und innere Gültigkeit (2)

Bedrohung der innere Gültigkeit:

- **Reifung** (maturation)
  - Veränderungen im Verhalten einer Versuchsperson über die Zeit hinweg (z.B. Ermüdung oder Lerneffekte)
- **Instrumentation** (instrumentation)
  - Veränderungen im Verhalten des Experimentators oder des Versuchsaufbaus über die Zeit hinweg (z.B. Verlangsamung von Antwortzeiten von Programmierwerkzeugen durch Plattendefragmentierung)
- **Historie** (history)
  - Das Vergehen von Zeit kann Wirkungen haben, die außerhalb des eigentlichen Experimentes liegen (bei über viele Woche laufenden Experimenten z.B. neue beeinflussende Nachrichten aus der Fachpresse)

Zentrale Begriffe: Störvariable und innere Gültigkeit (3)

Bedrohung der innere Gültigkeit:

- **Auswahl** (selection)
  - Oft kann wegen Vorkenntnissen die Auswahl von Versuchspersonen nicht wirklich zufällig erfolgen. Hier muss sichergestellt werden, dass sich nicht die Kriterien für die Gruppeneinteilungen auf die Ergebnisse auswirken.
- **Regression** (regression)
  - Wenn eine Versuchperson (z.B. in einem Vortest) eine für ihre Verhältnisse besonders gute oder besonders schlecht Leistung erbracht hat, so ist zu erwarten, dass bei einer späteren Messung die Leistung der selben Versuchsperson schlechter bzw. besser ist.
  - Wird das Ergebnis des Vortests dann zur Einteilung in Gruppen verwendet und dann die Auswirkung einer bestimmten Methode zur Verbesserung gemessen, kann es zu Verfälschungen kommen.

Zentrale Begriffe: Störvariable und innere Gültigkeit (4)

Bedrohung der innere Gültigkeit:

- **Sterblichkeit** (mortality)
  - Wenn Versuchspersonen während des Experimentes auf eigenen Wunsch ausscheiden (vor allem wenn das Ausscheiden in verschiedenen Gruppen unterschiedliche nicht zufällige Gründe hat).
- **Anforderungscharakteristik** (demand characteristics)
  - Durch die Art, wie eine Aufgabe gestellt wird, kann es zu einer unbeabsichtigten Bevorzugung einer Gruppe kommen (z.B. Experimentator nicht neutral)
- **Verarbeitungsfehler**
  - Nicht korrekte Messung von abhängigen Variablen (z.B. unzuverlässige Messvorrichtungen, schwankenden Urteile bei subjektiven Bewertungen).

Zentrale Begriffe: Äußere Gültigkeit

Äußere Gültigkeit:

- Die **äußere Gültigkeit** (externe Gültigkeit, external valitity) eines kontrollierten Experimentes ist der Grad, in dem sich seine Resultate korrekt auf andere Anwendungsfälle übertragen lassen – insbesondere auf solche, die in der Praxis häufig vorkommen.
- Dies betrifft zum Beispiel
  - Qualifikation der Versuchspersonen,
  - die Art und Größe der Software,
  - die Art und Form der Arbeitsaufgaben,
  - sowie Randbedingungen wie sonstige softwaretechnische Methoden, technisches und räumliches Arbeitsumfeld, Nervenzustand, Arbeitszeiten, Zeitdruck, Qualitätsanforderungen und Ähnliches.

Zentrale Begriffe: Experimententwurf (1)

In einem Experimententwurf wird festgelegt:

- Welche Gruppe
- in welcher Reihenfolge
- welche Aufgabe erledigt bzw. welchen Behandlungen unterworfen wird und
- welche abhängigen Variablen dabei beobachtet werden.

Zentrale Begriffe: Experimententwurf (2)

Ein Experimententwurf kann faktoriell sein:

- Jede in einem Experiment manipulierte unabhängige Variable heißt **Faktor** (factor) und jeder ihrer Werte heißt ein **Niveau** (level). Die Veränderung in der abhängigen Variablen zw. zwei Niveaus eines Faktors heißt **Effekt** (effect).
- Weist ein Experimentplan mehr als einen Faktor auf, spricht man von einem **faktoriellen Entwurf** (factorial design) oder **Faktorentwurf**.
- Treten im Plan alle denkbaren Kombinationen von Niveaus tatsächlich auf, nennt man ihn einen **vollständig faktoriellen Entwurf** (full factorial design), anderenfalls einen **unvollständigen** (partialfactorial design).

Zentrale Begriffe: Experimententwurf (3)

Beispiel für faktoriellen Entwurf:

Untersuchung der Produktivität von 2 Programmiersprachen:

		Faktor 1	
		Java	C
Faktor 2	Aufgabe A	A_{Java}	A_C
	Aufgabe B	B_{Java}	B_C

- Falls nicht viele Versuchspersonen zur Verfügung stehen, kann es günstig sein, dass jede zwei Datenpunkte liefert:
  - G_1: A_{Java} und B_C
  - G_2: A_C und B_{Java}
- Achtung: Durch diesen Entwurf entsteht eine dritte unabhängige Variable: Reihenfolge der Sprache

Zentrale Begriffe: Experimententwurf (4)

Intra-Subjekt vs. Extra-Subjekt-Design

- **Extra-Subjekt-Design (extra-subject design):**
  - Experimentplan, in dem jede Person nur eine Aufgabe zu bearbeiten hat
  - Miteinander verglichenen Gruppen sind in jedem Falle disjunkt sind.
  - Vorteil:
    - Der Umfang der einzelnen Aufgaben kann größer gewählt werden.
- **Intra-Subjekt-Design (intra-subject design):**
  - Jede Versuchsperson löst mehrere Aufgaben
  - Folglich werden mehrere separate Messungen vorgenommen (wiederholte Messungen, repeated measurements).
  - Vorteil:
    - Bei gleicher Zahl von Versuchspersonen können mehr Daten erhoben werden.
    - Zufällige Unterschiede zwischen Versuchsgruppen können sich ausgleichen.

Zentrale Begriffe: Experimententwurf (5)

- **Achtung:** Bei Intra-Subjekt-Entwürfen hat häufig die Reihenfolge, in der die Versuchspersonen ihre Aufgaben erledigen, einen Einfluss auf das Ergebnis.
- **Beispiel:**
  - Untersuchung der Produktivität von 2 Programmiersprachen wie oben gesehen mit zwei Gruppen nacheinander an zwei Aufgaben:
    - G_1: A_{Java} und B_C
    - G_2: A_C und B_{Java}
  - Angenommen Ergebnis zeige, bei Aufgabe A sei Java besser, bei Aufgabe B aber C und G_1 ist sicher nicht versehendlich leistungsstärker als G_2. Dann sind mögliche Gründe:
    - Sprachvorteil ist problemabhängig
    - Reihenfolgeeffekt: Java-Programmieren in erster Aufgabe motiviert für zweite Aufgabe, C-Programmieren frustriert.

Zentrale Begriffe: Experimententwurf (6)

- Zum *Entdecken* von Reihenfolgeeffekten oder Lerneffekten oder zum *Neutralisieren* von Reihenfolgeeffekten kann der **gegenbalancierte Experimententwurf** (counterbalancing) verwendet werden:
  - Im Experiment werden alle verschiedenen Reihenfolgen der Niveaus eines Faktors untersucht,
  - in dem man jede Versuchsgruppe in entsprechende Untergruppen aufteilt.
  - Durch eine solche Anordnung kann man anschließend Reihenfolgeeffekte wahlweise analysieren (durch Vergleich der Untergruppen)
  - oder neutralisieren (durch zusammenwerfen der Untergruppen).
  - Werden alle möglichen Reihenfolgekombinationen aller Faktoren untersucht, so heißt der Experimententwurf **vollständig gegenbalanciert**, anderenfalls **teilweise gegenbalanciert**.

Zentrale Begriffe: Experimententwurf (7)

- Im angegebenen Beispiel resultieren also vier Gruppen:
  - G_{1a}: A_{Java} und B_C
  - G_{1b}: B_C und A_{Java}
  - G_{2a}: A_C und B_{Java}
  - G_{2b}: B_{Java} und A_C

Zentrale Begriffe: Experimententwurf (8)

- Selbst bei perfekt balancierten Gruppen weiteres Problem:
  - Große Unterschiede in der Leistung zw. einzelnen Versuchpersonen
  - Es ist ja leider in der SWT nicht möglich eine Person mit sich selbst zu vergleichen (Lerneffekt!)
- Lösung: **Paarbildung**
  - Jede Person wird nicht mit sich selbst, sondern mit einer anderen ihr ähnlichen verglichen.
  - Diese beiden bilden ein Paar, von dem wir so tun, als seien beide dieselbe Person.
  - Paarbildung kann auf Basis von *Vortests* erfolgen.
  - Für die Analyse können dann statische Methoden angewendet werden (z.B. gepaarte Hypothesentests)

Zentrale Begriffe: Experimententwurf (9)

- Ein **Vortest**
  - im engeren Sinne ist eine Aufgabe, der die Teilnehmer aller Versuchsgruppen unterworfen werden,
  - und zwar vor Beginn des eigentlichen Kernexperimentes
  - und ohne Unterschied in der Behandlung der Versuchsgruppen.
  - Die im Vortest bearbeitete Aufgabe sollte idealerweise von gleicher Art sein wie die im Experiment zu bearbeitenden, und es sollten auch die gleichen abhängigen Variablen beobachtet werden.
  - Ein kann
    - Lerneffekte im Experiment vermindern
    - Grundlage für eine Paarbildung sein
    - …

Inhalt

1. Empirie in der SWT
2. Überblick empirische Forschungsmethoden
3. Zentrale Begriffe und Prinzipien
4. **Rollen kontrollierter Experimente**
5. Wann ist ein Experiment gut?

Rollen kontrollierter Experimente

Ein kontrolliertes Experiment kann unterschiedliche Arten von Wirkungen (in Form einer Rolle) haben:

- **Hypothesentest**
- **Erkundung**
- Quantifikation von Effekten
- **Modellbildung**
- Absicherung von Erkenntnissen
- Verbreiterung von Erkenntnissen

Rollen kontrollierter Experimente

1. **Hypothesentest**: Bislang häufigste Rolle kontrollierter Experimente in der SWT ist die Prüfung einer konkreten Vermutung, in der Regel von der Form „A ist besser als B bezüglich Eigenschaft X“. Dies erfolgt meist mit Hilfe eines **statistischen Hypothesentests**.
   - Vorherrschen des Hypothesentests in der SWT ist ein Zeichen für die *geringe Reife des Feldes*:
     - Denn er besagt nur, ob zumindest mit einer gewissen Wahrscheinlichkeit ein Unterschiede besteht , aber nicht wir groß dieser ist oder auf welchen Mechanismen er beruht.
   - Beispiel: Flußdiagramme benötigen weniger Zeit für das Verständnis als Pseudocode [Scanlan 1989].

Rollen kontrollierter Experimente

2. **Erkundung**: Erkundung ist das allgemeine Untersuchen kaum verstandener Aspekte ohne vorbestimmte Erwartungen.
   - Zweck: Das *Bilden von Hypothesen und Modellvorstellungen* über Wirkung und Wirkungsweise einer Technik.
   - Meist nicht Hauptziel eines kontrollierten Experimentes, da hierzu Fallstudien viel billiger.
   - Kontrollierte Experimente haben als Erkundungsmethode zwei Vorteile vor Fallstudien: *Replikation* und *Kontrolle*.
     - Wenn in einer Fallstudie ein bestimmter Effekt beobachtet wird, weiß man nur, dass er überhaupt auftreten kann (nicht dass er evtl. nur sehr selten auftritt).
     - Nur die Kontrolle (Kontrollgruppe) ermöglicht es, denn Effekt auf den erkundeten Tatbestand zurückzuführen.
   - Erkundung neuer Aspekte tritt bei Experimenten meist als Nebenprodukt auf (*stolpern* über interessante Beobachtung)

Rollen kontrollierter Experimente

- Beispiel Erkundung:
  - Porter und Kollegen 1995
  - Vergleich unterschiedlicher Teamgrößen für Inspektionen
  - Sie stellten fest: Zahl der beim Inspektionstreffen aufgedeckten Fehler, die zuvor keiner der Inspekteure gefunden hatte, war im Mittel nicht größer als die zuvor zwar gefundenen dann aber im Treffen wieder verschütteten Fehler.
  - Resultierende Frage: Sind Treffen wirklich sinnvoll?

Rollen kontrollierter Experimente

3. **Modellbildung**: Sobald eine Einflussgröße durch einen Hypothesentest als solche nachgewiesen ist und evtl. die Größe ihrer Wirkung für konkrete Einzelfälle quantifiziert ist, stellen sich weiterführende Fragen:
   - Wie kommt die Wirkung zustande (Struktur) und
   - wie würde sie sich unter anderen Umständen verändern (Wechselwirkungen)?
   - Gesucht: Quantitatives Modell, dass den Einfluss und das Zusammenspiel aller beteiligten Variablen bei der Softwareentwicklung beschreibt!
     - Eigenschaften der Menschen
     - Entwicklungsmethoden
     - Projektorganisation …
   - Leider: Zahl der Einflussgrößen in der SWT sehr groß und ihre Wirkungen und Wechselwirkungen wenig bekannt.

Rollen kontrollierter Experimente

- Die meisten Modelle in der SWT entnehmen ihre Daten Feldstudien (z.B. Kostenschätzungsmodelle)
- Es ex. aber auch Ausnahmen, in denen Modelle aufgrund von kontrollierten Experimenten gebildet wurden.

Inhalt

1. Empirie in der SWT
2. Überblick empirische Forschungsmethoden
3. Zentrale Begriffe und Prinzipien
4. Rollen kontrollierter Experimente
5. **Wann ist ein Experiment gut?**

Wann ist ein Experiment gut?

Beim Erstellen und Ausführen von Studien sollte so vorgegangen werden, dass folgende Punkte maximiert werden:

- **Die Fehlerfreiheit der Interpretation**: Dies ist in dem Sinne zu verstehen, dass das resultierende Ergebnis nicht das Ergebnis eines unbekannten Einflusses ist.
- **Relevanz**: Das Ergebnis liefert wichtige Informationen zur SWT.
- **Auswirkung**: Das Ergebnis beeinflusst den praktischen Einsatz der oder die Forschung in der SWT.

Als begrenzende Faktoren fungieren dabei die zur Verfügung stehenden Recourcen.

Wann ist ein Experiment gut?

Diese Sicht auf Experiment wird durch zwei Begriffe geprägt:

- **Glaubwürdigkeit:**
  - **Gültigkeit** (validity)
  - **Verlässlichkeit** (reliability)
- **Relevanz**

- Die **Verlässlichkeit** bestimmt , wie reproduzierbar die ermittelten Daten sind.
- Die **Relevanz** der vom Experiment bearbeiteten Fragestellung ergibt sich aus zwei Aspekten:
  - Wie bedeutsam ist die Forschungsfrage, zu deren Beantwortung das Experiment durchgeführt wird?
  - Wie nützlich ist die konkret bearbeitete Experimentfrage zur Beantwortung der Forschungsfrage?

Danke!

Vorträge

1	**Structured Flowcharts Outperform Pseudocode: An Experimental Comparison +** **Experimental investigations of the utility of detailed flowcharts in programming**
2	**Formal Methods Application: An Empirical Tale of Software Development.** + **Comments on "Formal Methods Application: An Empirical Tale of Software Development +** **Response to "Comments on 'Formal Methods Application: An Empirical Tale of Software Development'**
3	**An initial assessment of aspect-oriented programming.**
4	**An experimental evaluation of assumption of independence in multiversion programming+** **A reply to the criticisms of the Knight and Leveson**
5	**Two controlled experiments assesing the usefulness of design pattern information during program maintenance.**
6	**Principles of Survey Research** (Part 1 – Part 6)
7	**A systematic survey of CMM experience and results +** **Software quality and the Capability Maturity Model**
8	**Customer-Developer Links in Software Development**
9	**A controlled experiment measuring the effects of Personal Software Process (PSP)**
10	**An internally replicated quasi-experimental comparison of checklist and perspective-based reading of code documents.**
11	**Evaluating inheritance depth on the maintainability of object-oriented software. +** **A controlled experiment on inheritance depth as a cost factor for maintenance.**
12	**Anywhere, anytime code inspections: Using the web to remove inspection bottlenecks in large-scale software development. +** **Assessing software review meetings: A controlled experiment study using CSRS**
13	**An Empirical Methodology for Introducing Software Processes**
14	**The empirical investigation of perspective-based reading.**
15	**Experimental Design and Analysis in Software Engineering** (Part 2 –Part 5)