Empirische Evaluation: Kontrollierte Experimente

Ilhan Aslan, Chi Tai Dang, Björn Bittner, Katrin Janowski, Elisabeth André

Human Centered Multimedia
Institute of Computer Science
Augsburg University
Universitätsstr. 6a
86159 Augsburg, Germany

Kontrollierte Experimente – Grundlagen

- Beantwortung offener Fragen über:
  - Performance des Systems
  - Zufriedenheit der Nutzer
  - …
- Wissen über generische Fragen zu Anwendungen:
  - Vergleich von Designstrategien
  - Vergleich von Eingabe- bzw. Ausgabegeräten (Interaktionsparadigmen und Präsentationsformen)
  - …
- Sammeln von Wissen für Guidelines
  - Was ist in welcher Situation gut oder schlecht?
  - Wie sollte man das System in welchen Situationen realisieren?
  - …

Kontrollierte Experimente – Grundlagen

- Grundlegende Idee
  - Spezifikation von Werten (Variablen), die gemessen und verglichen werden sollen
  - Ausgewählte Teilnehmer führen ausgewählte Tasks aus
  - Ergebnisse helfen Schlussfolgerungen zu treffen
- Typischerweise als Laborstudien durchgeführt

Kontrollierte Experimente –
Was wird getestet?

Universität Augsburg University

Mögliche Hypothesen:

- Das System funktioniert.
- Das System funktioniert besser mit Merkmal X.
  - ➢ Test mit und ohne X
- Merkmal X hat die Eigenschaften…
  - ➢ Verändere X und beobachte die Effekte
- System A funktioniert besser als B (weil es X hat).
  - ➢ Test mit System A und B

Wichtige Fragestellungen:

- Ist nur Merkmal X als Grund für die Verbesserung möglich?
- Was bedeutet „besser“?
  (z.B. effizienter, effektiver, zufriedenstellender)

Kontrollierte Experimente –
Schritte bei der Durchführung

Universität Augsburg University

1. Experimentelles Design:
   - **Variablendefinition:**
     - Manipulierbare Eigenschaften (z.B. „Merkmal X“)
     - Messbare Eigenschaften (z.B. Effizienz, Effektivität…)
     - Messverfahren:
       - Befragungs- und/oder Beobachtungstechniken
       - Qualitative und/oder Quantitative Messungen
   - **Hypothesen** (Annahmen über Variablen) **aufstellen**
   - **Gruppendesign**
     - Within-Groups
     - Between-Groups

- **Planung der Studie**
  - Wahl der Versuchspersonen (Welche? Wie viele?)
  - Festlegung des Ablaufs
    - Auswahl der Aufgaben (Tasks)
    - Sonstiges: z.B. Texte zur Einführung und Erläuterung / Aufklärung
  - Sonstiges: Beschaffung und Testen der Messverfahren

2. Durchführung der Studie
   - Pilottest
   - Eigentliche Studie
3. Statistische Auswertung der Ergebnisse

Universität Augsburg University

Zwei Typen von Variablen:

- **Unabhängige Variablen** werden im Experiment verändert und beeinflussen das Ergebnis (Manipulierbare Eigenschaften)
- **Abhängige Variablen** hängen von den unabhängigen Variablen ab und werden gemessen (Messbare Eigenschaften)

Einfache Experimente:

- Eine unabhängige Variable, ein bis zwei abhängige Variablen

Komplexere Experimente:

- Multivariate Experimente (mehrere abhängige Variablen)
- Multifaktorielle Experimente (mehrere unabhängige Variablen)

Unabhängige Variablen legen die Bedingungen im Experiment fest.

- Beispiele:
  - Anzahl der Elemente in einer Liste
  - Schriftgröße

Die Werte pro unabhängiger Variable heißen **Level**.

- Beispiel:
  - Unabhängige Variable: Schriftgröße
  - Level 1: 12
  - Level 2: 16
  - Level 3: 20

Frage:

Wie verändert die Manipulation der unabhängigen Variablen eine messbare Eigenschaft?

Vorgehen:

Durchlaufen aller Level der **unabhängigen Variablen** und Messen der entsprechende Effekte anhand der **abhängigen Variablen**.

Messverfahren:

- Objektive Datenerhebung
- Subjektive Datenerhebung

1. Experimentelles Design - Variablendefinition

- Objektive Datenerhebung **(Beobachtungstechniken)**
  - Beispiele:
    - Verhalten der Nutzer (audio-visuelle Aufzeichnungen)
    - Zeit um einen Task T durchzuführen (Log-Files)
    - Anzahl der Fehler oder durchgeführten Aktionen (Log-Files)
- Subjektive Datenerhebung **(Befragungstechniken)**
  - Beispiele:
    - Vorlieben der Nutzer
    - Schwierigkeitsgrad der Nutzung
- Ergebnisse:
  - quantitative Daten und / oder
  - qualitative Daten, die quantifiziert wurden (z.B. Annotationen)

Wichtig!! Störeinflüsse vermeiden!!

- Abhängige Variablen sollten nur von den unabhängigen Variablen beeinflusst werden!
- Beispiel: „Schriftgröße“ (Level: 12, 16 und 20)
  - Zwei abhängige Variablen: Zeit und Fehleranzahl
  - Zu jedem der drei Level werden die abhängigen Variablen gemessen und später in der Analyse verglichen
  - Beispiele für Ergebnisse:
    - Für die Schriftgröße gilt, dass:
      - der Level 16 am effizientesten (Zeit) ist.
      - der Level 20 am effektivsten (Anzahl Fehler) ist.
  - Mögliche Störeinflüsse: Licht, andere Personen im Raum …

1. Experimentelles Design - Hypothese

- Definition:
  Vermutung bzw. Vorhersage bzgl. der Beziehung zwischen abhängiger und unabhängiger Variable (Ergebnis des Experiments)
- Beispiele:
  - H1: Unterschiedliche Schriftgrößen verändern die Effizienz.
  - H0-1: Die Schriftgröße hat keinen Einfluss auf die Effizienz. (Gegenhypothese).
  - H2: Unterschiedliche Schriftgrößen verändern die Effektivität.
  - H0-2: Die Schriftgröße hat keinen Einfluss auf die Effektivität. (Gegenhypothese).
- Hinweis: Hypothesen können auch eine gerichtete Annahme enthalten. (z.B. Größere Schriftgröße erhöht die Effektiviät.)

1. Experimentelles Design - Hypothese

- Mit Hilfe des Experiments wird die Hypothese entweder bestätigt oder widerlegt
- Vorgehen:
  1. Beginn mit einer Null-Hypothese (Gegenteil der Annahme):
     - „Die unabhängige Variable hat keinen Effekt auf die abhängige.“
  2. Durchführung des Experimentes und Nutzung statistischer Methoden um die Null-Hypothese zu widerlegen
  3. Wenn die Statistik einen signifikanten Unterschied für die Ergebnisse des Experiments zeigt, ist der Effekt kein Zufall und die tatsächliche Annahme damit bewiesen!
     („Signifikanz“ siehe Foliensätze Datenanalyse)

Universität Augsburg University

Versuchspersonen sind:

- **teuer und schwer zu finden**
  - meist zu wenige Versuchspersonen,
  - bei hoher Variabilität Probleme mit statistischer Analyse
- **sehr variabel** bzgl. Vorwissen, Fähigkeiten, Reaktionszeiten, Einstellung zum Versuchsleiter...
  - Heterogenität statt Homogenität
  - Störvariablen beim Vergleich der Ergebnisse
  - Nur Personen, die zu den Personas passen!
- **lernfähig**
  - Trainingseffekt: Aufeinanderfolgende, gleiche Versuche sind nicht unabhängig, da die Versuchsperson mit jedem Versuch dazulernt.
  - Reihenfolgeeffekt: Wird eine Versuchsperson in unterschiedl. Levels getestet, kann die Reihenfolge der Level von Bedeutung sein.

Universität Augsburg University

Within-Groups

Alle Teilnehmer führen alle Levels aus

- Werte aus jedem Level direkt miteinander vergleichbar
- kein Personeneffekt

ABER: Oft starke Lerneffekte!

- Reduzierung von Lerneffekten durch **Beachtung von Reihenfolgen**
- Beispiel:

Gruppe 1	Gruppe 2
Level 1	Level 2
Level 2	Level 1

gleiche Personenanzahl in beiden Gruppen

Universität Augsburg University

Between-Groups

Gruppe 1

Level 1

Gruppe 2

Level 2

Unabhängige Gruppen testen immer genau einen Level

- Werte von Person X aus Gruppe „Level 1“ werden mit Werten von Person Y aus Gruppe „Level 2“ verglichen.
- kein Lerneffekt

ABER: Personeneffekt, wenn Gruppen nicht homogen ist!

- Sind X und Y wirklich vergleichbar?
  - Reduzierung von Personeneffekten durch:
    - **sehr viele Testpersonen**
    - **bewusste Eingruppierung von Personen** (z.B. gleich viele Männer und Frauen)

2. Durchführung des Experimentes

1. Begrüßung und Aufklärung über die Grundlagen
2. Erlaubnis der Aufzeichnung einholen
3. Starten bzw. Kalibieren der Aufzeichnungsgeräte
4. Eigentlicher Versuch:
   - Versuchspersonen die Aufgaben mitteilen.
   - Abarbeitung der Aufgaben. (Wird aufgezeichnet.)
   - Within-Groups: Erneute Abarbeitung der Aufgaben mit geänderten unabhängigen Variablen bis alle Level durchlaufen wurden.
5. Aufzeichnungen beenden und bedanken
6. Eventuell Aufklärung über Details der Studie (z.B. gewünschte Erkenntnisse…)

3. Auswertung des Experimentes

- Auswertung der Ergebnisse z.B. anhand von Annotationen der Aufzeichnungen oder der quantitativen Daten (z.B. Click-Stream)
  - Beispiele:
    - Benötigte Zeit oder Fehleranzahl
    - Bewertung der Zufriedenheit anhand einer Skala von 1-5.
- Ergebnis:
  - Hypothese entweder belegt oder widerlegt.

- **Objektivität:**
  - Ergebnisse sind unabhängig davon welche Person das Experiment durchgeführt, ausgewertet oder die Ergebnisse interpretiert hat und wie sie sich dabei verhalten hat.
  - **Gibt es das selbe Ergebnis, wenn ein anderer Versuchs-leiter den Versuch durchführt, auswertet und interpretiert?**
- **Reliabilität bzw. Zuverlässigkeit:**
  - Grad der Genauigkeit, mit der ein bestimmtes Ergebnis bei einer Wiederholung des Experiments erneut gemessen wird (unabhängig davon, ob man dieses Ergebnis mit dem Test überhaupt messen wollte).
  - **Gibt es das selbe Ergebnis, wenn ich den Versuch wiederhole?**

- **Validität:**
  - Grad der Genauigkeit mit der ein Test tatsächlich das misst, was er messen soll.
  - Beispiel: Klausur

    Alles verstanden oder nur gut auswendig gelernt?
  - **Kann man mit dem Ergebnis wirklich eine Aussage über die Hypothese treffen?**
  - Arten von Validität:
    - Konstruktvalidität
    - externe Validität
    - interne Validität

1. **Konstruktvalidität:**
   - Gibt es Korrelationen zwischen unabhängiger Variablen A und abhängiger Variablen B?
   - Repräsentieren die unabhängigen und abhängigen Variablen tatsächlich die erwarteten Konstrukte?
   - Schlechtes Beispiel: Erfassung der Intelligenz durch Kopfumfang (keine Korrelation)
   - Stichwort: sinnvolle Variabilität
2. **Externe Validität (*Allgemeingültigkeit)*:**
   - Lassen sich die Schlussfolgerungen bzw. Entscheidungen auf andere Populationen, Situationen oder Zeitpunkte generalisieren bzw. übertragen?

3. Interne Validität:

- Sind die Ergebnisse auf eine Kausalbeziehung zwischen unabhängigen Variablen (UV) und abhängigen Variablen (AV) zurückführbar?
- Ist die AV einzig von der UV abhängig oder gab es während der Durchführung weitere Variablen bzw. Effekte, die die AV beeinflusst haben (= Störvariablen)?
- Beispiel: Verbesserte Effizienz der Nutzung nur mit der Änderung der Schriftgröße erklärbar oder hat sich noch etwas anderes geändert?
  - z.B. Lichtverhältnisse, Lautstärke?
  - Gab es Personen- oder Lerneffekte?
    - Lerneffekt, der Nutzer weiß jetzt wie die Aufgabe zu erfüllen ist und ist deshalb schneller
    - Personeneffekt, der eine Nutzer ist geübter bei der Aufgabenerfüllung als der andere Nutzer

- **Generalisierung bezüglich**
  - experimenteller Einheiten (sprich: Stichprobe)
    - Auswahl der Versuchspersonen (VP):
      - Entsprechen die VP den Personen, für die die Ergebnisse gelten sollen?
      - ➢ Repräsentative Stichprobe
    - **Ist das Ergebnis auch mit anderen Versuchs-personen der Zielgruppe zu erwarten?**
  - experimenteller Umgebung (sprich: Setting)
    - Situationsmerkmale
    - Künstlichkeit der experimentellen Situation
    - **Ist das Ergebnis auch in einem anderen Setting zu erwarten (z.B. im Feld)?**

Ziele:

1. Äquivalenz aller experimentellen Einheiten (Personen, Gruppen...) hinsichtlich aller Merkmale, die nicht zur Erklärung der Unterschiede der abhängigen Variablen herangezogen werden sollen.
   - ➢ Veränderungen der abhängigen Variable können allein auf die **Variation der unabhängigen Variablen** zurückgeführt werden.

2. Variation bzw. Konstant-Haltung der Variablen
   - – Kontrolle der unabhängigen Variablen durch planmäßige Variation
   - – Kontrolle von Störvariablen durch Konstant-Haltung ihres Einflusses

Gütekriterien für Ergebnisse -
Sicherung der internen Validität

Problem:

Versuchsteilnehmer können nicht „konstant gehalten“ werden.

Verfahren zur Sicherung der internen Validität:

- Randomisieren
- Ausbalancierung
- Parallelisieren

Sicherung der internen Validität - Randomisierung

- Zufallszuweisung von:
  - Personen zu Gruppen
  - Gruppen zu experimentellen Bedingungen
- Ziel:
  - Zufällige Verteilung (unkontrollierbarer) Personenmerkmale (z.B. Motivation, Erfahrung, Stimmung...) auf verschiedene Gruppen
  - Systematische Unterschiede, die zu systematischen Fehlern in den Daten führen, sollen sich über die gesamte Gruppe hinweg „rausmitteln“.
- Wann?
  - Unverzichtbar, wenn Störvariablen wirksam sind und keine andere Kontrolle möglich ist
  - Funktioniert nur bei **großen Stichproben** optimal

Sicherung der internen Validität - Ausbalancierung

- **Vollständige Ausbalancierung**

  Jeder Level soll:

  - gleich häufig vorkommen
  - gleich häufig vor und nach jedem anderen Level vorkommen
  - jeder VP gleich oft dargeboten werden (bei Messwiederholung (= Within-Group Design))

- Beispiel:

  A B C
  A C B
  B A C
  B C A
  C B A
  C A B

  3! = 3*2 Möglichkeiten

Sicherung der internen Validität - Ausbalancierung

- **Unvollständige Ausbalancierung**

  Jeder Level steht:

  - gleich oft an 1., 2., ...k-ter Stelle
  - gleich oft vor jeder anderen Bedingung

- Beispiel:

  A B C D
  B D A C
  C A D B
  D C B A

  4 aus 24 (4!) möglichen Reihenfolgen wurden realisiert!

Sicherung der internen Validität - Parallelisierung

- „Gleichmachung“ verschiedener Gruppen bezüglich eventuell beeinflussender Merkmale
- Wann?
  - bei **kleinen Stichproben**
  - bei sehr engem Zusammenhang zwischen AV/UV und der Störvariablen.
  - Besonders bei Between-Groups Design, wenn es zu Störung durch Personeneffekte kommen kann.
- Wie?
  - Gruppen werden so gebildet, dass sie sich in Mittelwert und Standardabweichung der Störvariablen nicht unterscheiden
  - Verteilung von Personenmerkmalen auf verschiedene Gruppen durch *matching („Statistische Zwillinge“)*