
statistical
behavioral
data
analysis
statistics
Statisticians
techniques
horrified
may
many
perhaps
scientists
problems
subjects
statisticians
simple
often
routines
open
little
design
fiddled
fitted
However
nonstatistician
applications
multivariate
merit
straightforward
software
put
measurements
employing
extravagance
ignored
regression
covariance
powerful
analyses
folklore
advice
awareness
misuse
particularly
reallife
less
horror
pitfalls
Quintessentially
stresses
solving
similar
access
plot
graphs
p-value
balanced
although
complain
provide
sometimes
upset
axes
times
examples
lie
feel
investigator
statistician
graph
inappropriate
one
Indeed
sciences
significance
lack
insecure
use
part
just
routine
average
become
tests
answers
interpretation
nonstandard
stories
approach
nonlinear
cell
relevant
methods
used
collection
working
way
questionable
useful
observations
example
collected
Many
abuse
undue
faith
four
place
wait
meaning
apparent
linear
misleading
methodology
favor
questions
für
relatively
research
demand
path
features
unwary
complex
apply
practice

27.04.2017


statistical
behavioral
data
analysis
statistics
Statisticians
techniques
horrified
may
many
perhaps
scientists
problems
subjects
statisticians
simple
often
routines
open
little
design
fiddled
fitted
However
nonstatistician
applications
multivariate
merit
straightforward
software
put
measurements
employing
extravagance
ignored
regression
covariance
powerful
analyses
folklore
advice
awareness
misuse
particularly
reallife
less
horror
pitfalls
Quintessentially
stresses
solving
similar
access
plot
graphs
p-value
balanced
although
complain
provide
sometimes
upset
axes
times
examples
lie
feel
investigator
statistician
graph
inappropriate
one
Indeed
sciences
significance
lack
insecure
use
part
just
routine
average
become
tests
answers
interpretation
nonstandard
stories
approach
nonlinear
cell
relevant
methods
used
collection
working
way
questionable
useful
observations
example
collected
Many
abuse
undue
faith
four
place
wait
meaning
apparent
linear
misleading
methodology
favor
questions
für
relatively
research
demand
path
features
unwary
complex
apply
practice

27.04.2017

Multivariate Statistik

Inhaltliche Abstimmung innerhalb des Moduls

Vorlesung

- Grundlegende Konzepte der Datenanalyse
- Lineare Regression
- Allgemeines Lineares Model

Übung

- Konkrete Einführung in R
- Praktische Übung der Analyse von Daten mit R

Multivariate Statistik

Lernziele

- Verständnis der Konzepte der Datenanalyse
- Verständnis des Allgemeinen Linearen Models
- Anwendung der Konzepte auf konkrete Daten

Leistungsanforderungen

- Vor- und Nachbereitung der Vorlesung
- Lesen der Kapitel im Buch
- Anwendung der Konzepte in den Übungen

Multivariate Statistik

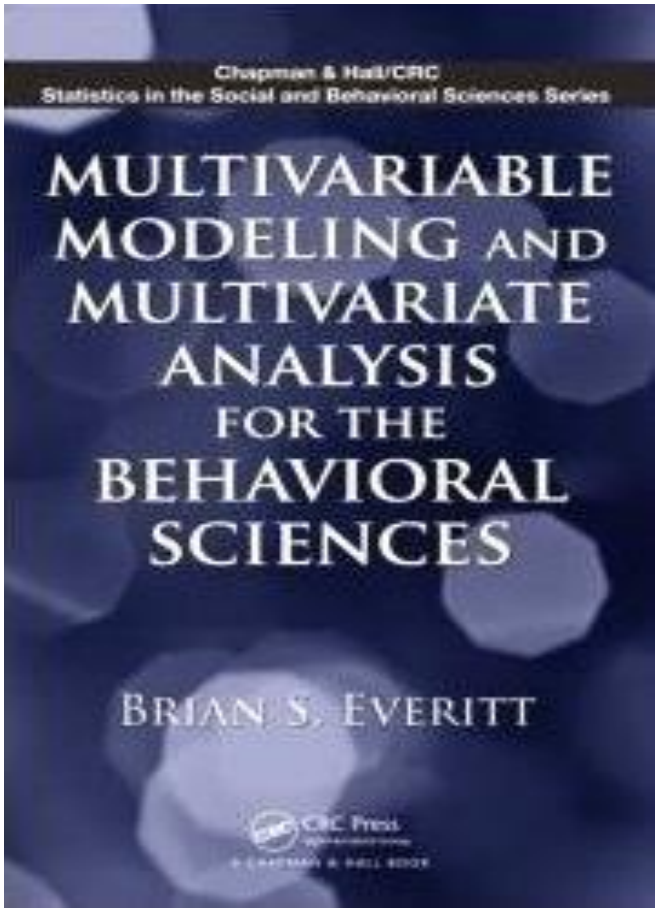
Lernhilfen

- Vorlesung
- Buch zur Vorlesung
- Skript der Vorlesung auf der Homepage
- Blockkurs zur Einführung in R am 08./09.4. & 22./23.4.2017
- Anleitung zur Umsetzung in R
- Buch zur Umsetzung in R
- Übungsaufgaben in R
- Permanent ansprechbarer Tutor

Multivariate Statistik

Literatur

Everitt (2010)

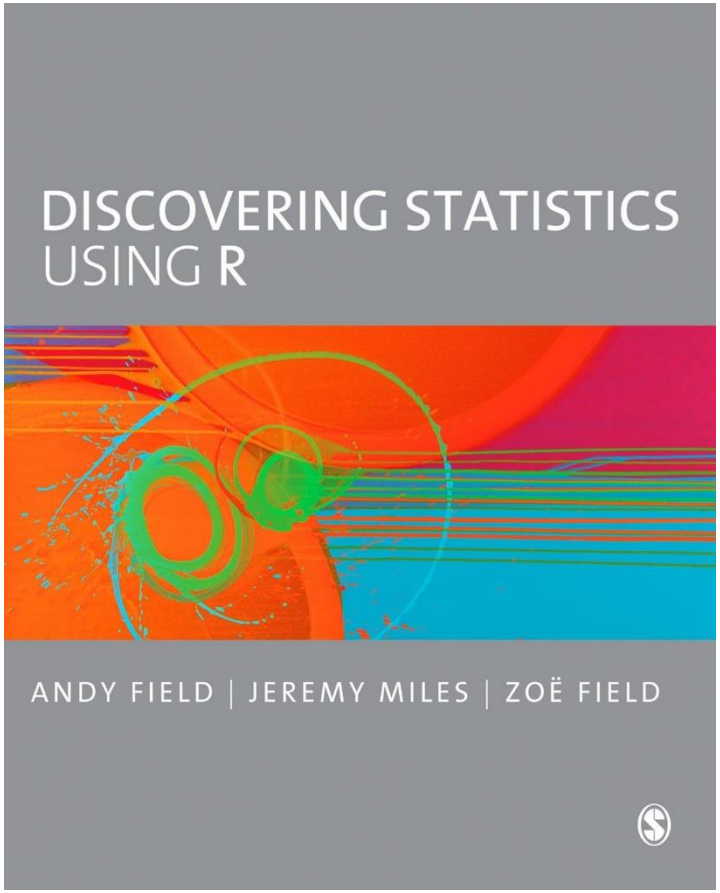


48 €

Multivariate Statistik

Literatur

Field (2012)



62 €

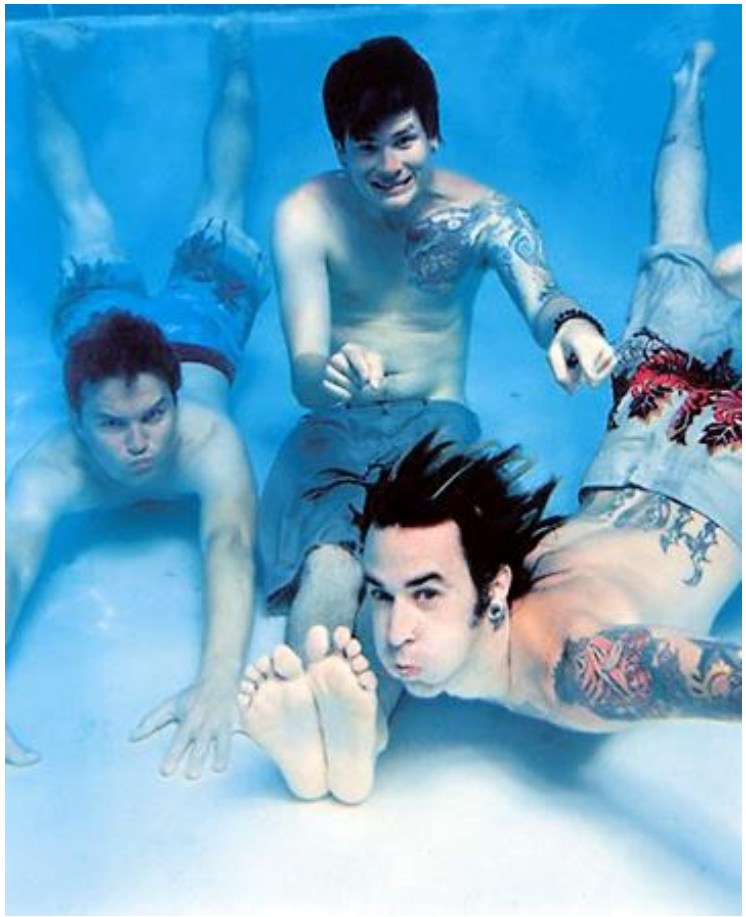
Multivariate Statistik

Zeiten

Termine zu Übungen im Pool

Donnerstag 10:15 - 11:45

Donnerstag 12:15 - 13:45



Heinrich Düker Weg 12

Raum 0.481

joshua.driesen@stud.uni-goettingen.de

Multivariate Statistik

Prüfung

Anwendung der Konzepte auf einen Datensatz

Modifikation vorliebender Daten-Analysen

Visualisierung von Daten

Interpretation von Analyseergebnissen

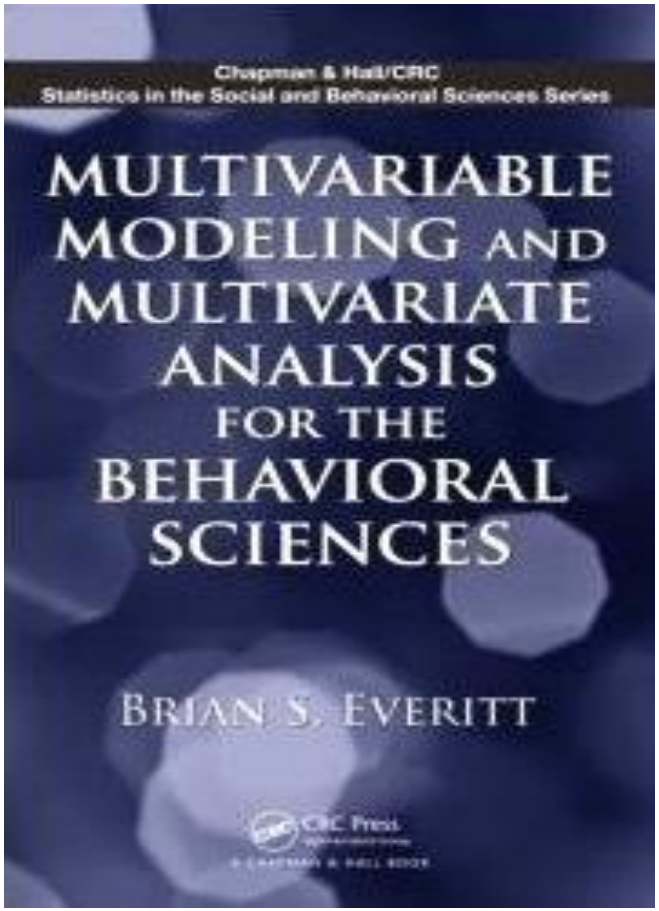
Termine

SS 2016	Sa 19. 8. 2017, 10:00 – 16:00
WS 2016/2017	Sa 03.02. 2018, 10:00 – 16:00

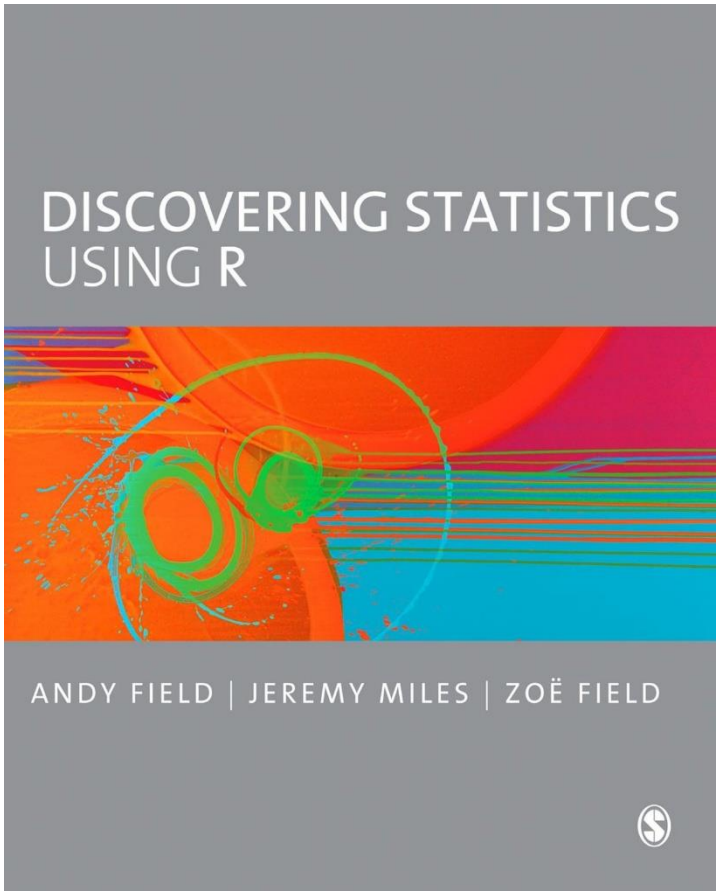
Daten und Modelle

- **Überblick**
  - Einleitung
  - Typen von Studien
  - Skalen von Messungen
  - Fehlende Messwerte
  - Datenanalyse mit Modellen
  - Stichprobengröße
  - Signifikanztests und Konfidenzintervalle

Literatur



Kapitel 1



Kapitel 1 & 2

Einleitung

Einleitung

Einleitung

Statistics is a general intellectual method that applies wherever data, variation, and chance appear. It is a fundamental method because data, variation and chance are omnipresent in modern life. It is an independent discipline with its own core ideas, rather than, for example, a branch of mathematics .... Statistics offers general, fundamental and independent ways of thinking.

Journal of the American Statistical Association

Einleitung

- **Vorkenntnisse**
  - Einfaches Hypothesen testen
  - Verwendung von Signifikanz-Tests
  - Einfache lineare Regression
  - Einfache Varianzanalyse

Einleitung

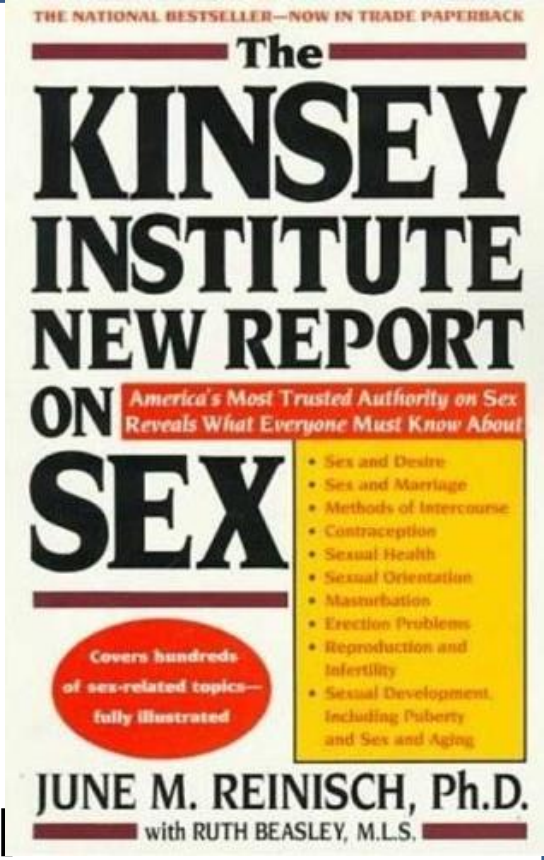
- **Alpträume eines Statistikers**
  - Irreführender Mittelwert
  - Irreführende Graphik
  - Unangemessener p-Wert
  - Unangemessene Analysen
  - **…**

- **Minimal-Ziele**
  - Etwas mehr über Statistik wissen
  - Statistik angemessen verwenden können

Einleitung

- **Themen**
  - Rekapitulation
  - Abbildungen
  - Lineare Regression
  - Varianzanalyse und ALM
  - Logistische Regression
  - Zeitreihenanalyse
  - Multivariate Analyse
  - PCA
  - Faktoren-Analyse
  - Cluster-Analyse

Typen von Studien



- **Umfragen**
  - Kinsey-Report
- **Experimente**
  - Experimentator variiert UV und ko
- **Beobachtungsstudien / Quasi-Ex**
  - Natürliche Gruppen: Rauchen + Blutdruck
- **Quasi-Experimentelle / Kombinierte Studie**
  - Natürliche Gruppen als Quasi-UV + variierte UV

Typen von Messungen

- **Qualität von Messungen**
  - Objektiv
  - Präzise
  - Reproduzierbar

- **Präzision von Messungen**
  - Größe, Geschwindigkeit, Zeit
  - IQ
  - Angst, Depression, Schmerz

Typen von Messungen

- **Nominale / Kategoriale Messungen**
  - Wechselseitig exklusiv – ohne Ordnung
- **Ordinale Messungen**
  - Exklusiv – Geordnet
- **Intervallskalierte Messungen**
  - Exklusiv - Geordnet – bedeutsame Abstände
- **Verhältnisskalen**
  - Exklusiv - Geordnet – bed. Abstände + Verhältnisse

Typen von Messungen

- **Nominale / Kategoriale Messungen**
  - Wechselseitig exklusiv – ohne Ordnung
- **Ordinale Messungen**
  - Exklusiv – Geordnet
- **Intervallskalierte Messungen**
  - Exklusiv - Geordnet – bedeutsame Abstände
- **Verhältnisskalen**
  - Exklusiv - Geordnet – bed. Abstände + Verhältnisse

Typen von Messungen

- **Multivariable Data**
  - Abhängige Variable = Response = Outcome
  - Unabhängige Variable = Explanatory Variable
- **Multivariate Data**
  - Verschiedene Zufallsvariablen

Fehlende Messwerte

- **Daten fehlen weil …**
  - Probanden nicht weitermachen wollen
  - sie nach der Erhebung verloren gehen
  - eine von vielen Messungen nicht funktioniert hat
  - …
- **Vier mögliche Reaktionen**
  - Unvollständige Datensätze weglassen
  - Plausible Ersatzdaten schätzen
  - Monte-Carlo Simulationen
  - Auf die Statistik verzichten

Fehlende Messwerte

- **Unvollständige Datensätze weglassen?**
  + Kein Aufwand zur Ergänzung der Werte
  - bei großen Datensätzen ineffizient
  - Reduktion der statistischen Power
  - Verzerrter Datensatz, weil Missing Values nicht unbedingt zufällig entstanden sind

  $\Rightarrow$ Gut, wenn nur selten Daten fehlen (5%)

Fehlende Messwerte

- **Plausible Ersatz-Daten schätzen**

  Mittelwert aus den andern Daten verwenden

  \+ ändert das Mittel nicht

  \- reduziert Variabilität

  \- reduziert Korrelationen in den Daten

  Mittels Regression geschätzte Werte verwenden

  \+ ändert das Mittel nicht

  \- erhöht Korrelationen in den Daten

  \- Unsicherheit in den Daten wird unterschätzt

  -> falsche p-Werte

Fehlende Messwerte

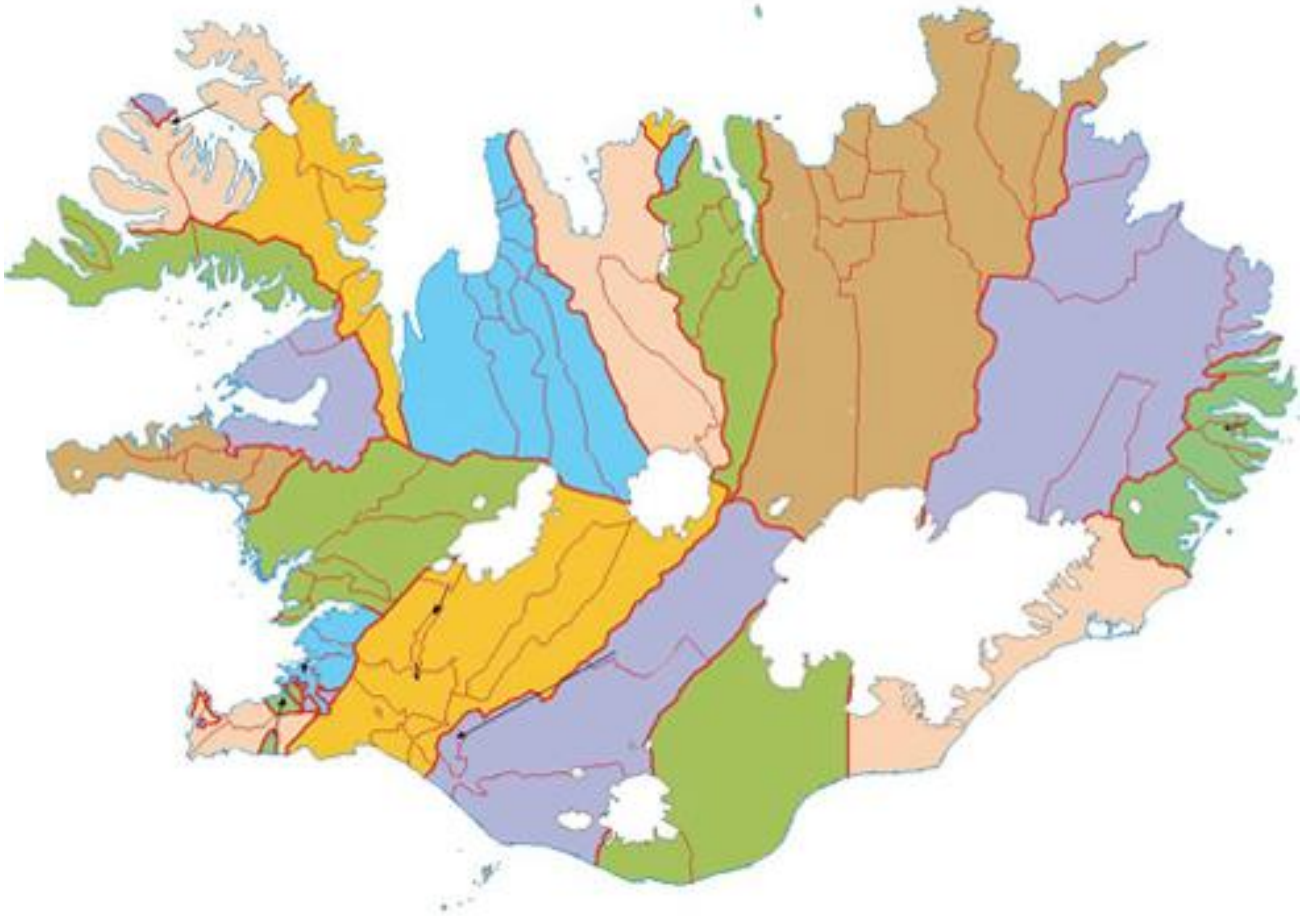
- **„Multiple Imputation“ (Rubin, 1987)**
  - Monte-Carlo-Simulationen
  - 3-10 komplette Datensätze mit simulierten Werten
  - Normale Analyse von jedem Datensatz
  - Zusammenführen der Analysen

- **Keine Statistik machen**

  Aus Nix kann Nix werden !

Datenanalyse mit Modellen

- **Modelle**



Datenanalyse mit Modellen

- **Modelle**
  - vereinfachte Imitationen wirklicher Objekte
  - folgen den relevanten Merkmalen möglichst genau
  - sind einfach zu handhaben
  - erlauben Schlussfolgerungen über das Objekt
  - bilden beobachtbare Daten ab

- **Beispiel: Lerneffekt**

Datenanalyse mit Modellen

- **Beispiel: Lerneffekt**

  1. Testergebnis = 20
  2. Testergebnis = 24

  Lerneffekt = 4

Datenanalyse mit Modellen

- **Modell 1**

$$x_1 = \gamma + \varepsilon_1$$

$$x_2 = \gamma + \delta + \varepsilon_2$$

$$\hat{\delta} = x_2 - x_1$$

Datenanalyse mit Modellen

- **Modell 1**

$$x_1 = \gamma + \varepsilon_1$$

$$x_2 = \gamma + \delta + \varepsilon_2$$

$$\hat{\delta} = x_2 - x_1$$

- **Modell 2**

$$x_1 = \gamma\varepsilon_1$$

$$x_2 = \gamma\delta\varepsilon_2$$

$$\hat{\delta} = \frac{x_2}{x_1}$$

Datenanalyse mit Modellen

- **Modell 1**

$$x_1 = \gamma + \varepsilon_1$$

$$x_2 = \gamma + \delta + \varepsilon_2$$

$$\hat{\delta} = x_2 - x_1$$

- **Modell 2**

$$x_1 = \gamma\varepsilon_1$$

$$x_2 = \gamma\delta\varepsilon_2$$

$$\hat{\delta} = \frac{x_2}{x_1}$$

- **Modell 3**

$$x_1 = \gamma + \varepsilon_1$$

$$x_2 = \gamma + (\lambda - \gamma)\delta + \varepsilon_2$$

Datenanalyse mit Modellen

- Modelle versuchen Änderungen in Messungen zu erklären

- **Die Wahl des Modells**
  - Additive lineare Modelle, weil Statistik möglich ist
  - Vorwissen
  - Empirische Daten
  - Post-Hoc Modelle + Test an neuen Daten
  - Ockham‘s Rasiermesser
  - „Alle Modelle sind falsch, manche sind nützlich“

Stichprobengröße

- **Praktische Aspekte**
  - Zeit
  - Verfügbarkeit der Probanden
  - Finanzielle Ressourcen

- **Statistische Aspekte**
  - Signifikanz-Niveau
  - Variabilität der Messungen
  - Erwünschte Power der Untersuchung
  - Größe des Effekts, der noch entdeckt werden soll „clinically relevant difference“

Stichprobengröße

$$n = \frac{2\left(Z_{\alpha|2} + Z_{\beta}\right)^2 \sigma^2}{\Delta^2}$$

- **Statistische Aspekte**
  - Signifikanz-Niveau α
  - Variabilität der Messungen σ
  - Erwünschte Power der Untersuchung β
  - Größe des Effekts, der noch entdeckt werden soll „clinically relevant difference“ Δ

Stichprobengröße

$$n = \frac{2\left(Z_{\alpha|2} + Z_{\beta}\right)^2 \sigma^2}{\Delta^2}$$

- **Beispiel: Therapie von Anorexie**
  - Signifikanz-Niveau = 0.05
  - Variabilität der Messungen = 4 kg
  - Erwünschte Power der Untersuchung = 90%
  - „clinically relevant difference“ = 1kg

Stichprobengröße

$$n = \frac{2\left(Z_{\alpha|2} + Z_{\beta}\right)^2 \sigma^2}{\Delta^2}$$

$$n = \frac{2 \times (1.96 + 1.28)^2 \times 4^2}{1} = 336$$

- **Beispiel: Therapie von Anorexie**
  - Signifikanz-Niveau = 0.05
  - Variabilität der Messungen = 4 kg
  - Erwünschte Power der Untersuchung = 90%
  - „clinically relevant difference“ = 1kg

Stichprobengröße

- **Reduktion der Stichprobengröße**
  „cynically relevant difference“

  „a guess masquerading as mathematics“
  „a game that can produce any number you wish“

- **Warnung**
  „Absence of evidence is not evidence of absence“

- **Und**
  „Some evidence is better than none“

Signifikanztests und Konfidenz-Intervalle

- **Oakes (1986): Was bedeutet der P-Wert?**

  Zwei Gruppen Versuchsplan a 20 Probanden

  t-Test ergibt t=2.7, df = 18, p=0.01

1. Die Nullhypothese ist absolut wiederlegt.
2. Die Wahrscheinlichkeit, dass die H0 wahr ist.
3. Die Experimental-Hypothese ist absolut bewiesen.
4. Die Wahrscheinlichkeit für die Experimental-Hypoth.
5. Die Wahrscheinlichkeit, dass das Zurückweisen der H0 falsch ist.
6. Eine Replikation wird zu 99 % signifikant werden.

Signifikanztests und Konfidenz-Intervalle

- **Oakes (1986): Was bedeutet der P-Wert?**

  Zwei Gruppen Versuchsplan a 20 Probanden

  t-Test ergibt t=2.7, df = 18, p=0.01

Frequencies and Percentages of "True" Responses in Test of Knowledge about p-Values

Statement	Frequency	Percentage
1. The null hypothesis is absolutely disproved.	1	1.4
2. The probability of the null hypothesis has been found.	25	35.7
3. The experimental hypothesis is absolutely proved.	4	5.7
4. The probability of the experimental hypothesis can be deduced.	46	65.7
5. The probability that the decision taken is wrong is known.	60	85.7
6. A replication has a 0.99 probability of being significant.	42	60.0

Signifikanztests und Konfidenz-Intervalle

- **Oakes (1986): Was bedeutet der P-Wert?**

  Zwei Gruppen Versuchsplan a 20 Probanden

  t-Test ergibt $t=2.7$, $df = 18$, $p=0.01$

- **Lösung**

  p gibt an, wie wahrscheinlich es ist den beobachteten oder einen noch größeren Unterschied zu finden, wenn die H0 wahr ist.

Signifikanztests und Konfidenz-Intervalle

- **P-Wert für eine einfache Ja-Nein-Antwort**
  - P = 0.049 $\Rightarrow$ Nullhypothese verwerfen
  - P = 0.051 $\Rightarrow$ Nullhypothese beibehalten
- **Konfindenz-Intervalle**
  - Plausibler Bereich der Werte des Parameters
- **P-Wert zur Beurteilung …**
  - der Möglichkeit, dass ein Effekt existiert
  - Patient tot?


statistical
behavioral
data
analysis
statistics
Statisticians
techniques
horrified
may
many
perhaps
scientists
problems
subjects
statisticians
simple
often
become
methodology
linear
misleading
questions
unwary
complex
apply
practice
für
relatively
research
demand
path
features
meaning
apparent
wait
relevant
observations
tests
answers
interpretation
nonstandard
nonlinear
cell
methods
example
collection
working
collected
Many
abuse
undue
faith
four
place
used
approach
stories
way
questionable
useful
routine
Indeed
inappropriate
sciences
significance
feel
lie
examples
complain
pitfalls
Quintessentially
stresses
solving
similar
access
plot
graphs
balanced
p-value
sometimes
provide
statistician
investigator
graph
upset
one
horror
reallife
less
misuse
awareness
covariance
powerful
analyses
folklore
advice
particularly
regression
fitted
fiddled
However
software
put
measurements
employing
ignored
extravagance
design
open
little
routines
nonstatistician
applications
multivariate
merit
straightforward
axes
times
lack
insecure
use
part
just
average
However
although
Many
Quintessentially