universität wien

Einführende Statistik
1. Allgemeine Einführung

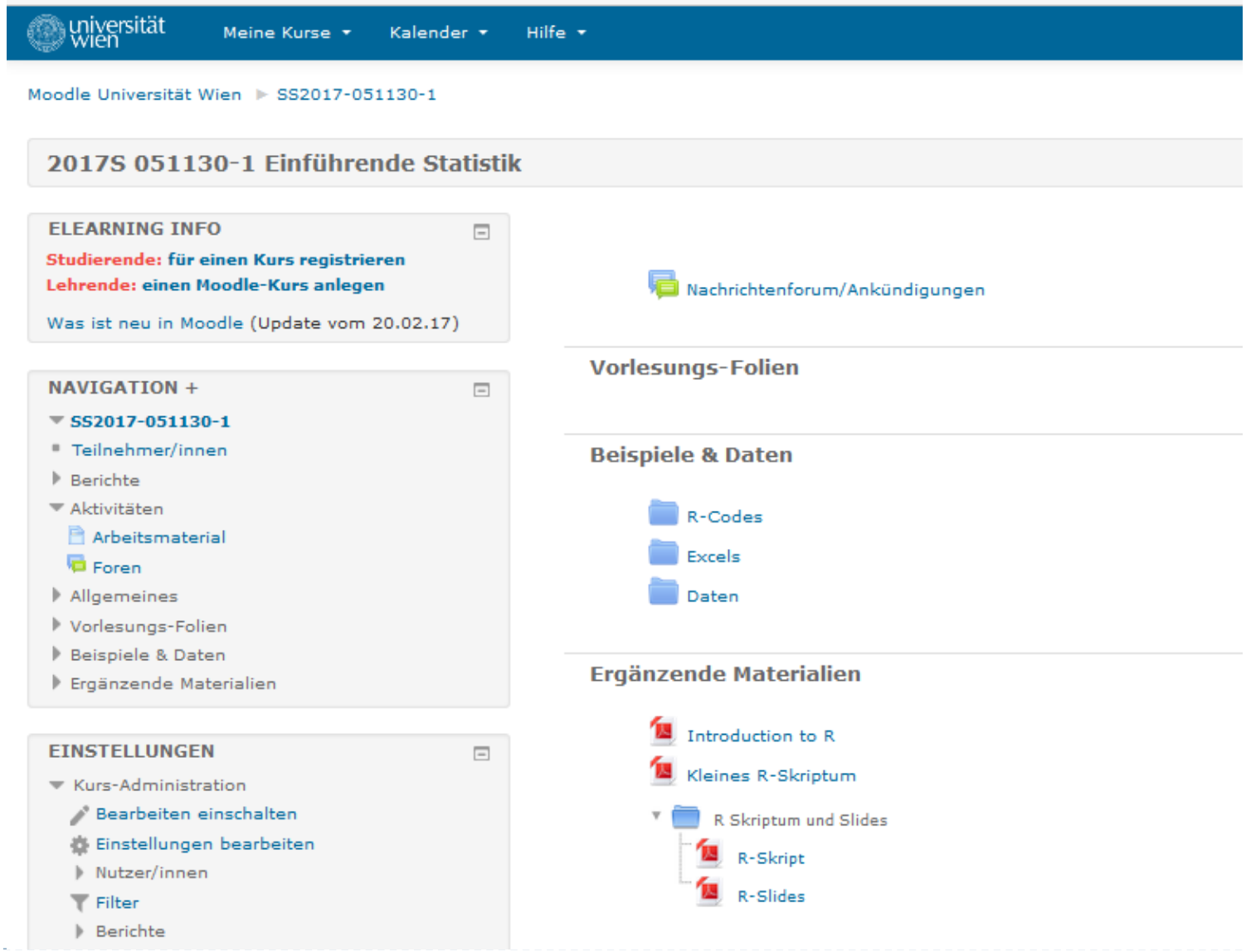
Marcus Hudec

Informationen

- Infos zu dieser Lehrveranstaltung: www.marcushudec.at

- Unterlagen zur Vorlesung: https://moodle.univie.ac.at/

Unterlagen auf Moodle



Ziele der Lehrveranstaltung

Ziele der Lehrveranstaltung

Die Studierenden verfügen über Fähigkeiten empirische Sachverhalte mittels statistischer Basistechniken zu beschreiben und graphisch korrekt zu repräsentieren; sowie über ein prinzipielles Verständnis für die grundlegenden Konzepte der Wahrscheinlichkeitstheorie und der inferenzstatistischen Modellierung und Methodik.

Die Studierenden sind in der Lage inhaltliche Fragestellungen in statistische Modelle zu übersetzen und diese mittels adäquater Techniken der Inferenzstatistik korrekt zu beantworten. Dabei können Sie moderne Softwarewerkzeuge für Analytik und Visualisierung zur Beantwortung datenanalytischer Fragestellungen erfolgreich anwenden.

Inhaltliche Gliederung

Inhalte

Deskriptive und Explorative Statistik
- Darstellung von Verteilungen
- Empirische Verteilungsfunktion und Quantile
- Deskriptive Maßzahlen der Lage und Streuung
- Weitere Maßzahlen (Schiefe, Wölbung)
- Assoziation, Korrelation

Wahrscheinlichkeitsrechnung
- Grundlagen der Wahrscheinlichkeitsrechnung
- Ereignisalgebra, Grundaufgaben der Kombinatorik
- Bedingte Wahrscheinlichkeit und Unabhängigkeit
- Satz von der Totalen Wahrscheinlichkeit
- Theorem von Bayes
- Zufallsvariablen
- Wichtige Diskrete Verteilungen
- Wichtige Stetige Verteilungen
- Ungleichung von Tschebyscheff
- Gesetz der großen Zahlen
- Zentraler Grenzwertsatz

Techniken der Inferenzstatistik
- Punktschätzer
- Intervallschätzer
- Hypothesentesten
- Klassische Tests bei Normalverteilung
- Einfache Varianzanalyse
- Test auf Unabhängigkeit
- Überprüfung von Verteilungsannahmen
- Nichtparametrische Testverfahren

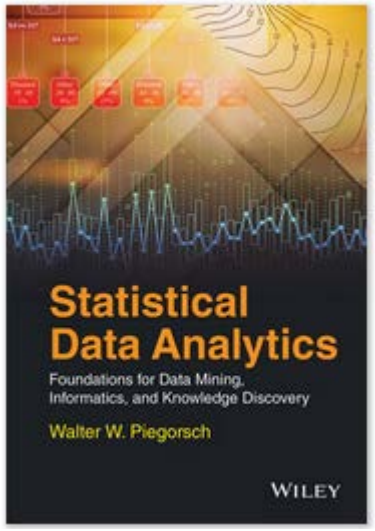
Regression
- Lineare Einfach-Regression
- Inferenz über die Parameter
- Konfidenz- und Prognoseintervalle
- Residuenanalyse

Weitere Infos zur LV

Literatur

- Statistics & Data with R: An Applied Approach Through Examples. Y. Cohen & J Y. Cohen, Wiley 2008.
- Introductory Statistics with R. P.Dalgaard, Springer 2002.
- Discovering Statistics Using R. A.Field, J. Miles, and Z. Field, Sage 2014.
- R Einführung durch angewandte Statistik. R.Hatzinger, K.Hornik, H.Nagel, Pearson Studium2011.
- Statistical Data Analytics. W.Piegorsch, Wiley 2015.
- Grundlagen der Datenanalyse mit R: Eine anwendungsorientierte Einführung. D. Wollschläger, Springer 2010.

R ist eine freie Software-Umgebung für statistische Datenanalyse, Simulation und graphische Visualisierungen.



Statistical Data Analytics: Foundations for Data Mining, Informatics, and Knowledge Discovery

Walter W. Piegorsch

ISBN: 978-1-118-61965-0

464 pages
August 2015

Prüfungsmodus

Prüfungsmodus

Ausschließlich schriftliche Prüfung über den Vorlesungsstoff nach dem Prüfungsterminraster der SPL Informatik. Erster Termin am Ende der Lehrveranstaltung (Ende Juni oder Anfang Juli), drei weitere Termine im folgenden Semester.

Die Prüfung enthält Multiple-Choice Fragen, einfache Rechenaufgaben sowie Fragen zur Ergebnisinterpretation.

Der Stoff umfasst alle Themen, die in der Vorlesung vorgetragen wurde.

Bei der Prüfung darf jeder Teilnhemer ein selbst gestaltetes, handschriftliches A4-Blatt mit Formeln, Notizen etc. mitbringen. (Collagen, Leporellos u.ä. ist nicht erlaubt). Die Nutzung von darüberhinaus gehenden Unterlagen (Bücher, Skripten) ist bei der Prüfung nicht erlaubt.

Taschenrechner dürfen bei der Prüfung verwendet werden. Untersagt ist aber die Verwendung von PDAs, Notebooks und ähnlichen elektronischen Geräten sowie die Nutzung von Smartphones.

Anmeldung für die Prüfung über UNIVIS ONLINE oder u:SPACE ist unbedingte Voraussetzung!

Jeder Teilnehmer muss einen Studentenausweis oder einen amtlichen Lichtbildausweis zur Prüfung mitbringen!

http://cran.r-project.org/

CRAN
Mirrors
What's new?
Task Views
Search

About R
R Homepage
The R Journal

Software
R Sources
R Binaries
Packages
Other

Documentation
Manuals
FAQs
Contributed

The Comprehensive R Archive Network

Download and Install R

Precompiled binary distributions of the base system and contributed packages, **Windows and Mac** users most likely want one of these versions of R:

- Download R for Linux
- Download R for (Mac) OS X
- Download R for Windows

R is part of many Linux distributions, you should check with your Linux package management system in addition to the link above.

Source Code for all Platforms

Windows and Mac users most likely want to download the precompiled binaries listed in the upper box, not the source code. The sources have to be compiled before you can use them. If you do not know what this means, you probably do not want to do it!

- The latest release (2013-09-25, Frisbee Sailing) R-3.0.2.tar.gz, read what's new in the latest version.
- Sources of R alpha and beta releases (daily snapshots, created only in time periods before a planned release).
- Daily snapshots of current patched and development versions are available here. Please read about new features and bug fixes before filing corresponding feature requests or bug reports.

http://www.r-project.org/

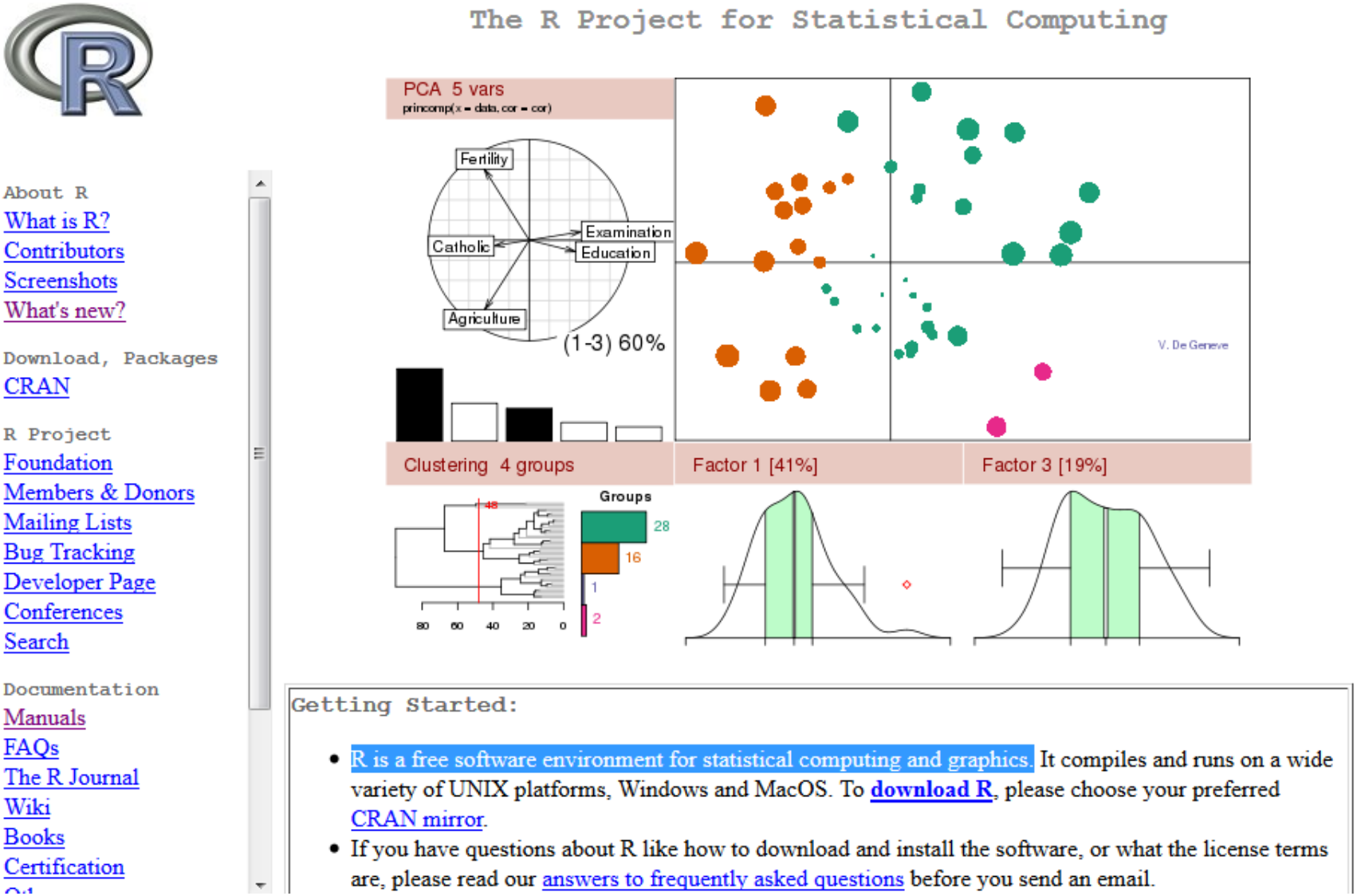


Image der Statistik

„Mit Statistik kann man alles beweisen!“

„Traue keiner Statistik außer der,
die du selbst gefälscht hast!“

„There are three kinds of lies:
lies, damned lies, and statistics“

Relevanz der Statistik

Wir leben im Jahrhundert der Statistiken.
Ein Trommelfeuer von Daten, Zahlen, Fakten,

Tabellen, Kurven, Trends und Tests decken den modernen Medienkonsumenten ein.

Hier sind Übersicht und ein klarer Kopf gefragt.

W. Krämer

Artikel aus der New York Times 08/2009

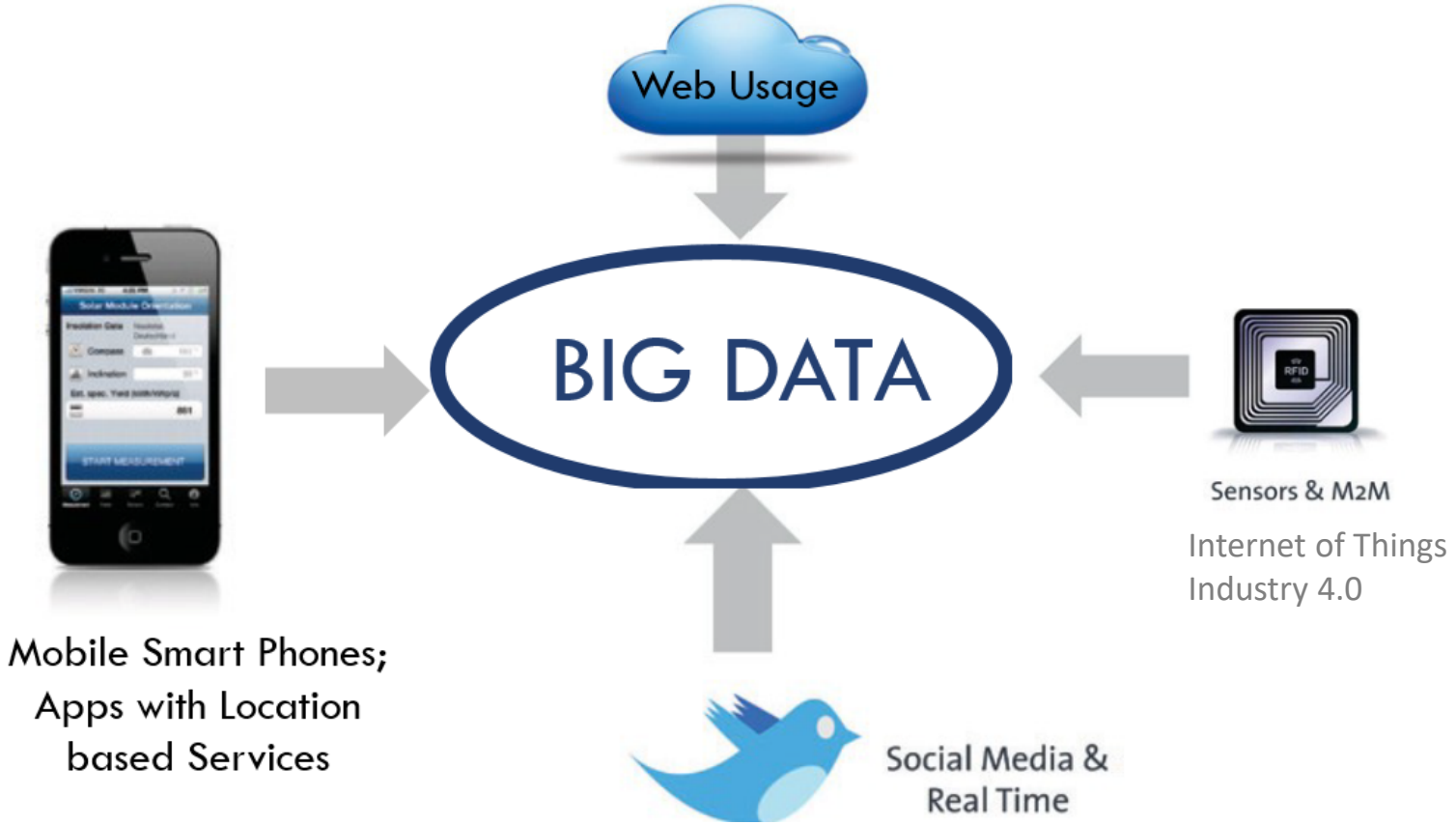
- "I keep saying that the sexy job in the next 10 years will be statisticians," said Hal Varian, chief economist at Google. "And I'm not kidding."
- The rising stature of statisticians is a by-product of the recent explosion of digital data. In field after field, computing and the Web are creating new realms of data to explore — sensor signals, surveillance tapes, social network chatter, public records and more.
- Yet data is merely the raw material of knowledge. "We're rapidly entering a world where everything can be monitored and measured," said Erik Brynjolfsson, an economist and director of the Massachusetts Institute of Technology's Center for Digital Business. "But the big problem is going to be the ability of humans to use, analyze and make sense of the data."

Wir leben in einer digitalen Welt

- Alle zwei Tage werden 5 Exabytes (= 5 Milliarden Gigabytes) an Daten produziert.
- Das ist etwa dieselbe Menge, die seit Beginn unserer Zivilisation bis ins Jahr 2003 entstanden ist.

The Data Revolution

Availability of data is dramatically increasing in industry



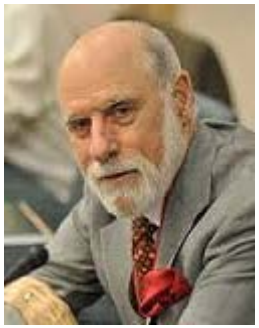
Entstehung von Big Data

- Forschungsdatenbanken
- Gen-Forschung
- Satellitenbilder in der Meteorologie
- CERN (Brutto 60 Terabyte pro Sekunde; 1GB p sec permanent)
- Prozessdaten aus Real-Time-Applikationen
- Mobilfunknetze
- Industrielle Fertigungsstraßen
- Transaktionen auf Finanzmärkten
- Webdaten
  - Content
  - Usage

Daten wie noch nie:

"We never, ever in the history of mankind have had access to so much information so quickly and so easily."

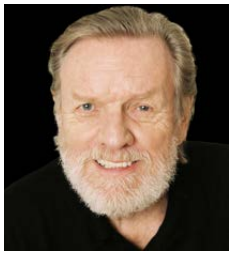
Vinton G. Cerf, „Vater des Internets"

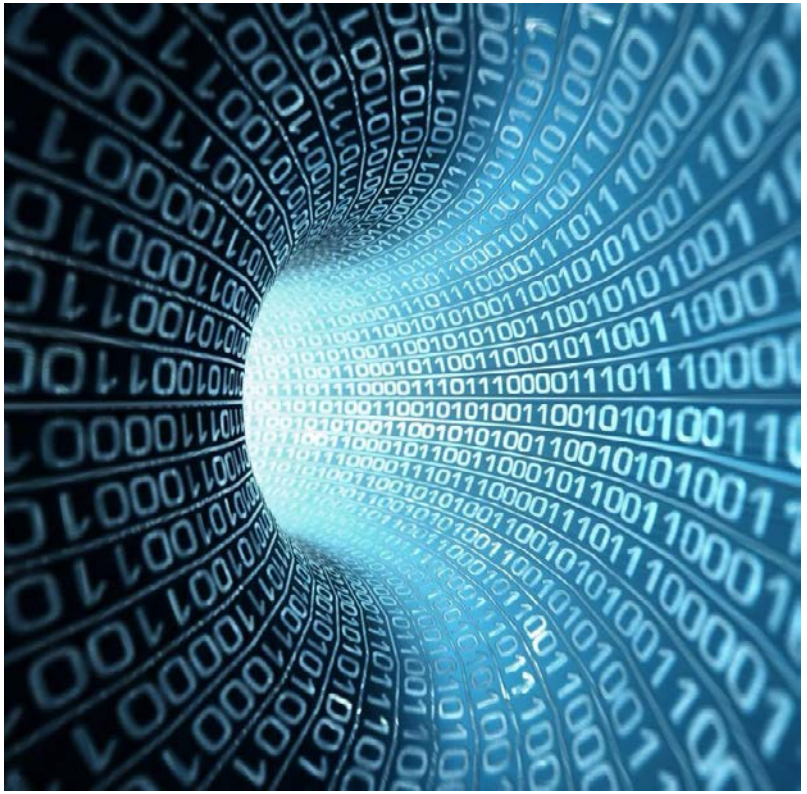


Need for Data Scientists

„We are drowning in information and starving for knowledge“

John Naisbitt, Trendforscher





Herausforderung

- Aus diesen gewaltigen, komplexen Datenmengen Informationen zu extrahieren, die sinnvoll nutzbar sind
  - Neue wissenschaftliche Erkenntnisse
  - Kommerziell verwertbares Wissen über Märkte und Kundenverhalten

Zitat aus www.futurezone.at

- Dass diese ungeheuren Datenmengen wertvolle Informationen beinhalten, liegt auf der Hand. Die Frage ist nur, wie aus dem teils unstrukturierten Datenhaufen nützliche Informationen herausgefiltert werden können. **Data Scientists**, oder Datenanalysten, wie sie auf Deutsch genannt werden, sollen mit Hilfe von Algorithmen aus dem vermeintlichen Datenmüll nützliche Informationen extrahieren.
- Was in Europa bestenfalls erste Gehversuche unternimmt und noch in den Kinderschuhen steckt, hat sich in den USA bereits zu einem großen Trend entwickelt. **Data Scientists sind heißbegehrte Spezialisten**, die von zahlreichen Unternehmen umworben werden. Mit ihnen wird die Hoffnung verbunden, aus Big Data Ideen für neue Erkenntnisse, Innovationen und Geschäftsmodelle herauszufiltern sowie bestehende Produktionsprozesse zu rationalisieren und Risiken etwa für Erdbeben oder Epidemien vorherzusagen.

Big Data ist heute ein Management-Thema

- **"Data als Rohstoff"**
  Datenangebot – als Rohstoff für vorausschauende Entscheidungen, Aktionen und Maßnahmen
- **"Big-Data-Technologien"**
  Rechnercluster, neuartige Datenbanken und skalierbar verteilte Verarbeitung
- **"Prädiktive Analytik"**
  extrahiert Strukturen und Zusammenhänge aus den Daten. Voraussagemodelle lassen sich automatisch trainieren und passen sich ständig an
- **"Data-driven Enterprises"**
  erschließen systematisch das Potenzial aus allen verfügbaren Daten. Sie gewinnen an Effizienz, schaffen individualisierte Angebote und smartere Produkte.

Statistische Kultur

- Fähigkeit statistische Inhalte (Diagramme, statistische Größen, Statistiken) verstehen zu können
- Kritisches Bewerten statistischer Ergebnisse mit denen wir im täglichen Leben konfrontiert werden
- Einschätzen der Bedeutung der Statistik
  - im privaten Bereich
  - im beruflichem Umfeld
  - im öffentlichen/politischen Bereich
- Quantitative Literacy
- Korrektes Verständnis quantitativer Größen, elementarer Rechenkonzepte (z.B. Veränderungsraten), Risiko- und Wahrscheinlichkeiten

"As our society is driven increasingly by science and technology, the need to establish levels of quantitative literacy becomes ever more important."
NSF Director Rita Colwell

Statistik & Demokratie

Die Statistik hat eine erhebliche Bedeutung für eine staatliche Politik. Wenn die ökonomische und soziale Entwicklung nicht als unabänderliches Schicksal hingenommen, sondern als permanente Aufgabe verstanden werden soll, bedarf es einer umfassenden, kontinuierlichen sowie laufend aktualisierten Information über die wirtschaftlichen, ökologischen und sozialen Zusammenhänge.
Erst die Kenntnis der relevanten Daten und die Möglichkeit, die durch sie vermittelten Informationen für die Statistik zu nutzen, schafft die für eine am Sozialstaatsprinzip orientierte Politik unentbehrliche Handlungsgrundlage.

Erkenntnis des deutschen Bundesverfassungsgerichtes 1983

Statistik schafft Handlungsgrundlage

Zuteilung von Budgetmitteln in der EU

Objektive Daten zur Umweltproblematik

Prognose der Bevölkerungsdynamik

Risikomanagement auf Finanzmärkten

Zulassung neuer Medikamente

Erheben politischer Stimmungslage

Quantifizieren sozialer Phänomen

Analyse von Kaufverhalten

Historische Quellen

- „Lehre von der Zustandsbeschreibung des Staates“
- John Graunt — Bills of Mortality (1662)
- Edmond Halley — Auswertung der Kirchenbücher von Breslau (1693)
- John Arbuthnot — Knaben/Mädchengeburten (1667-1735)
- Peter Süßmilch — „politische Arithmetik“ (1707-1767)
- Adolphe Quetelet — „ideale Histogramm“ (1796-1874)
- In Österreich: Wilhelm Winkler
- Glücksspiel - Wahrscheinlichkeitsrechnung
- Blaise Pascal (1623-1662), Pierre de Fermat (1601-1665), Thomas Bayes (1702-1761), Pierre Simon de Laplace (1749-1827), Andrei Kolmogoroff (1903-1987)
- Astronomische Messungen
- Tycho Brahe — Prinzip der Mittelwertsbildung (1546-1601)
- Carl Friedrich Gauß — Kleinste Quadrate Prinzip, Normalverteilung (1777-1855)

Was ist Statistik ?

- Statistik wird hier als Hilfswissenschaft aufgefasst. Sie ist eine der Methoden, mit der die Verbindung zwischen Theorie und Erfahrung (Empirie) systematisch reflektiert wird.

  Außer den reinen Formalwissenschaften wie Mathematik und Logik hat jede Wissenschaft "theoretische" und "empirische" Bestandteile.

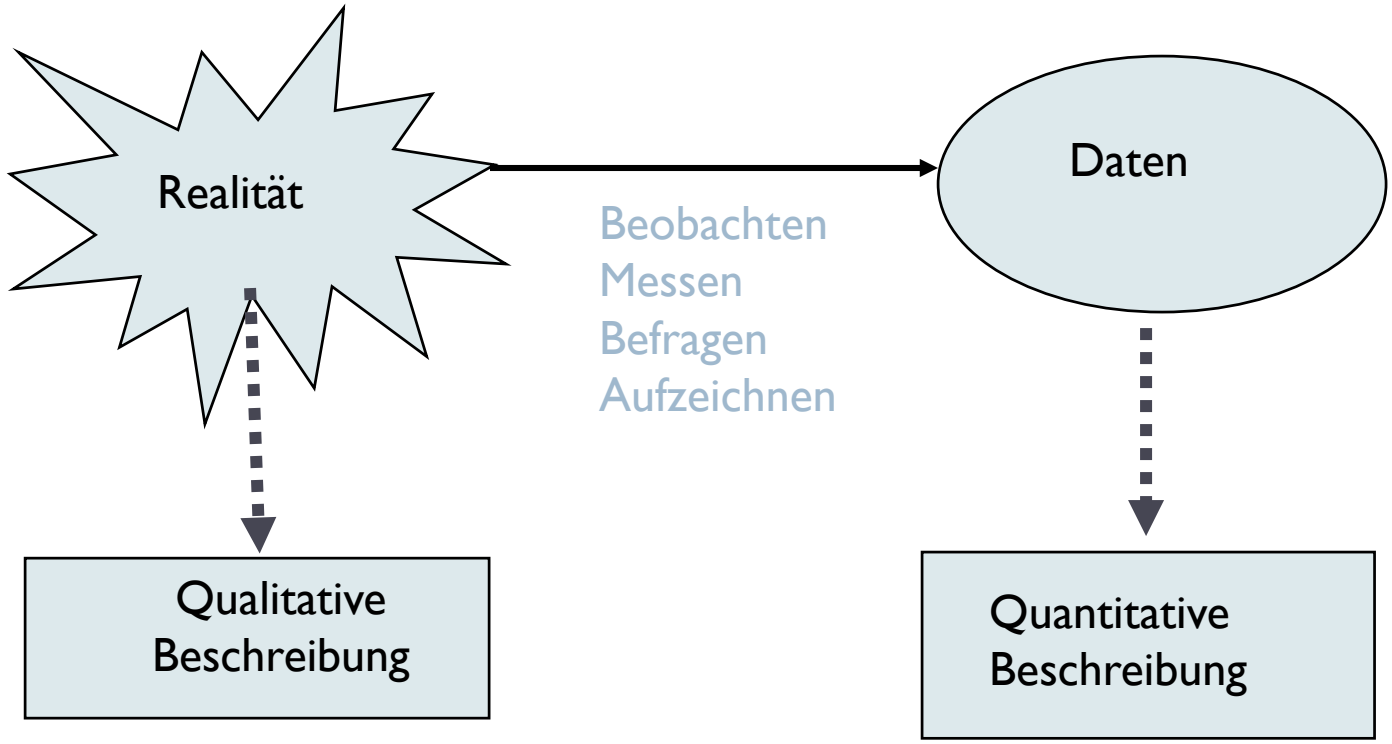
  Die Einsatzmöglichkeit der statistischen Methoden reicht demnach von Naturwissenschaften wie Physik, Astronomie, Biologie bis zu den Gesellschafts- und Geisteswissenschaften wie Nationalökonomie, **Soziologie**, Linguistik, Geschichte usw.

Franz Ferschl

Teilgebiete der Statistik

- Deskriptive Statistik - Inferenzstatistik
- Explorative Statistik - Konfirmatorische Statistik
- Angewandte Statistik - Theoretische Statistik

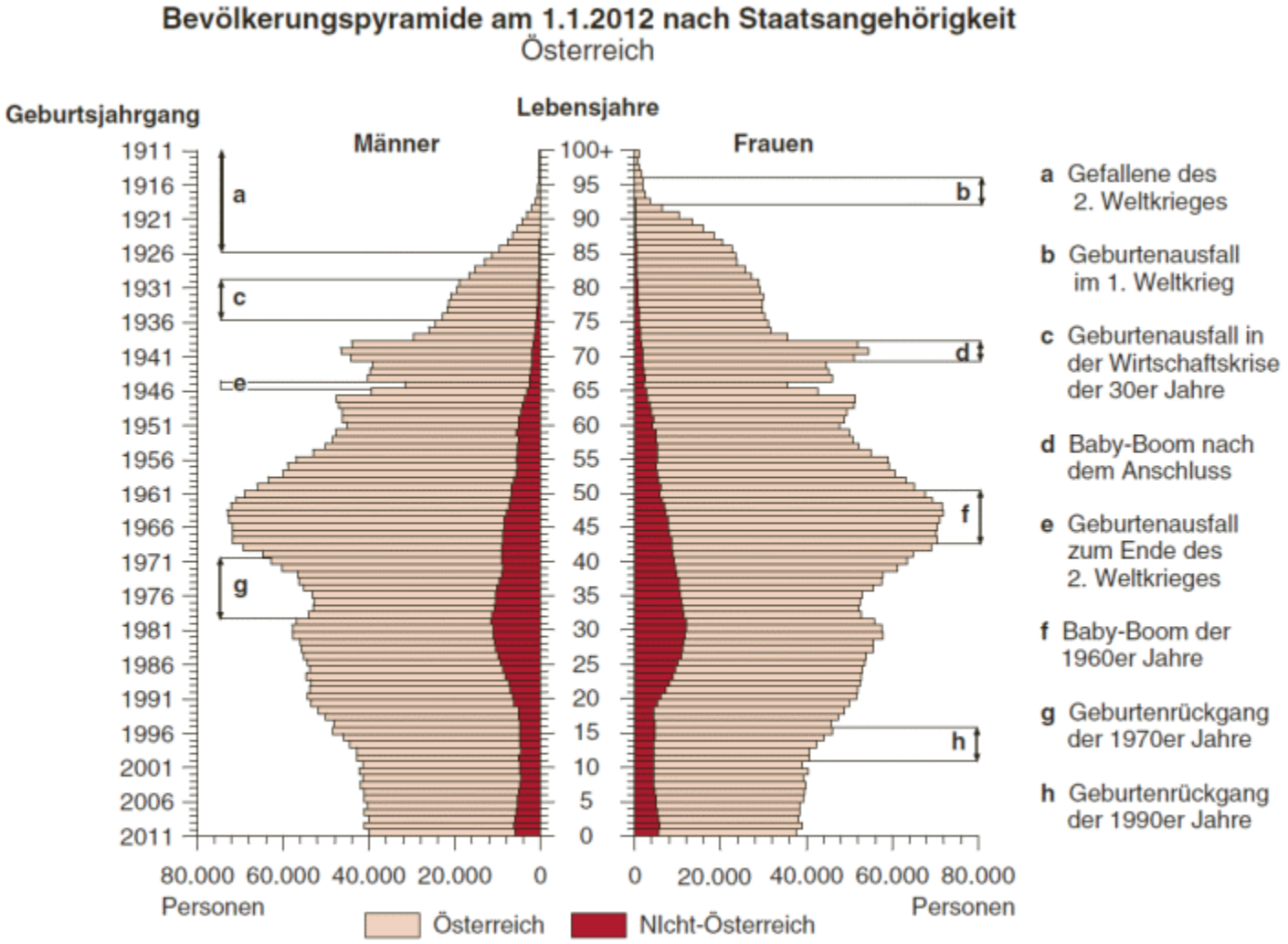
Grundprinzip der deskriptiven Statistik



Deskriptive Statistik

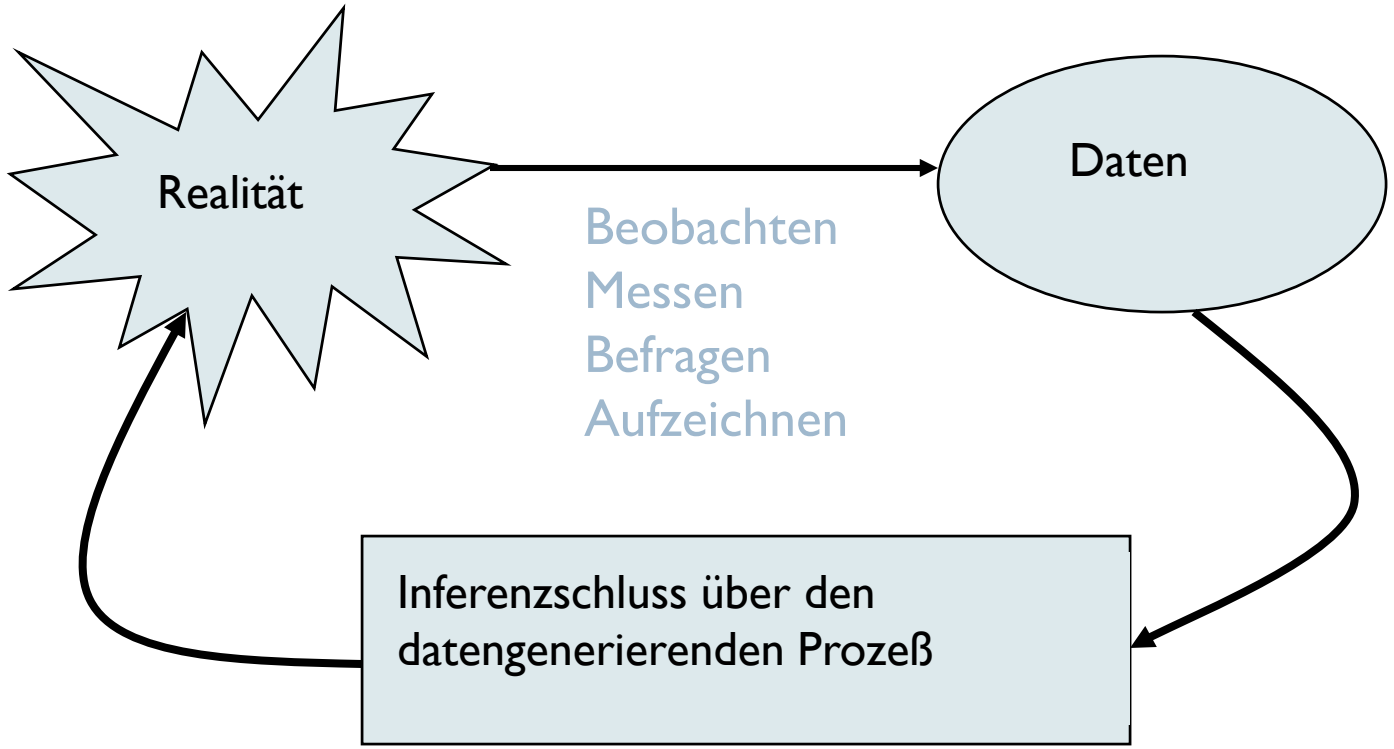
- Erhebung und Betrachtung der Daten per se
- Anwendung von beschreibenden Methoden
- Instrumentarium zur Beschreibung und Präsentation von Datenmaterial
- Graphische Aufbereitung von quantitativen Daten
- Erkennen und Beschreiben typischer Muster in Datenkörpern
- Ermittlung von Kennwerten
- Vorstufe zur schließenden Statistik
- Quantitative Erfassung und überschaubare Aufbereitung von massenhaft auftretenden Einzelerscheinungen
- Deskriptive Statistik ist der Sammelbegriff für Methoden zur Beschreibung von Daten in Form von Tabellen, Graphiken oder einzelnen Kennwerten (Indikatoren)

Beschreibung demographischer Entwicklungen



Q: STATISTIK AUSTRIA, Statistik des Bevölkerungsstandes. Erstellt am: 14.05.2012.

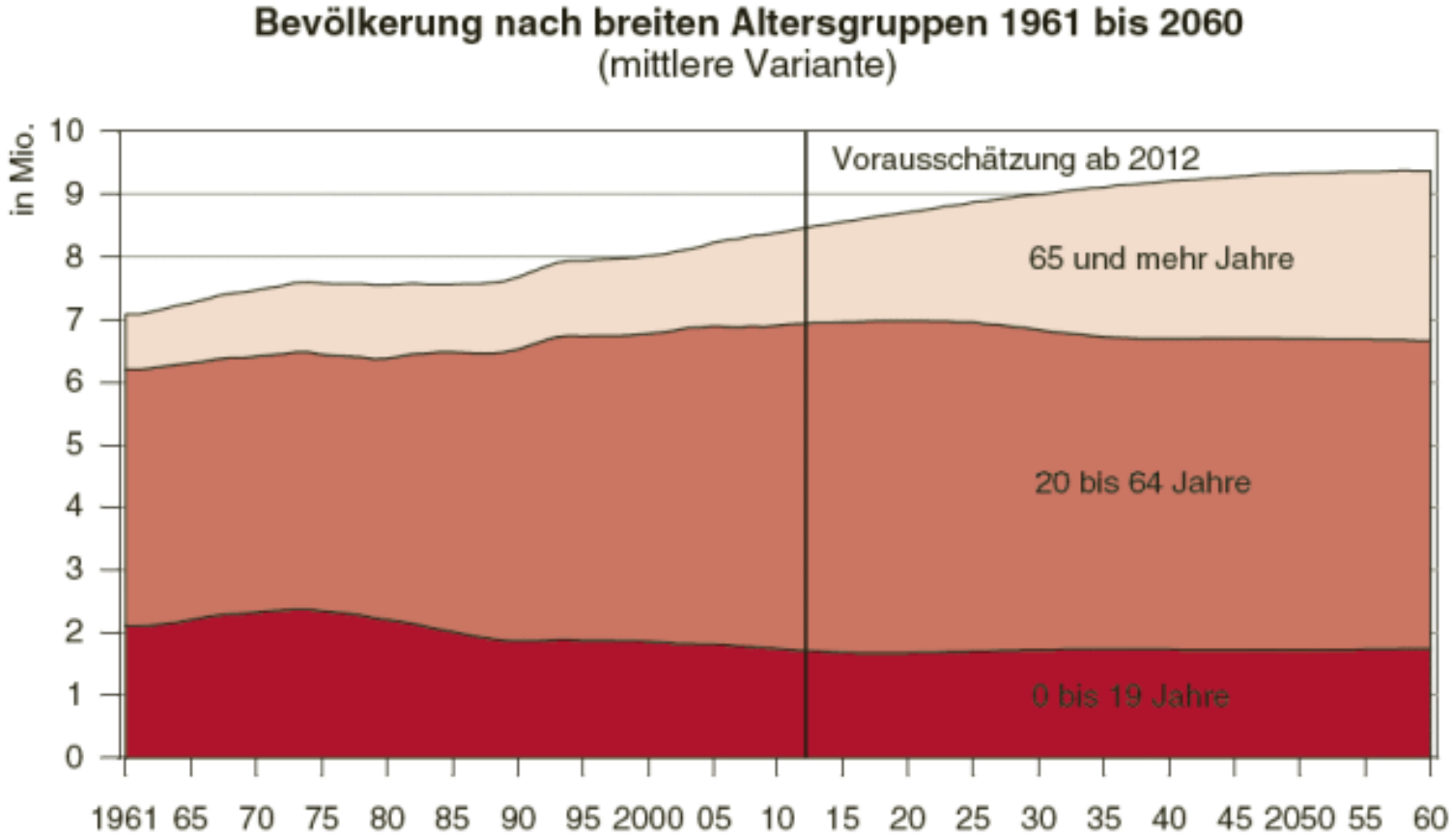
Grundaufgabe der Inferenz-Statistik



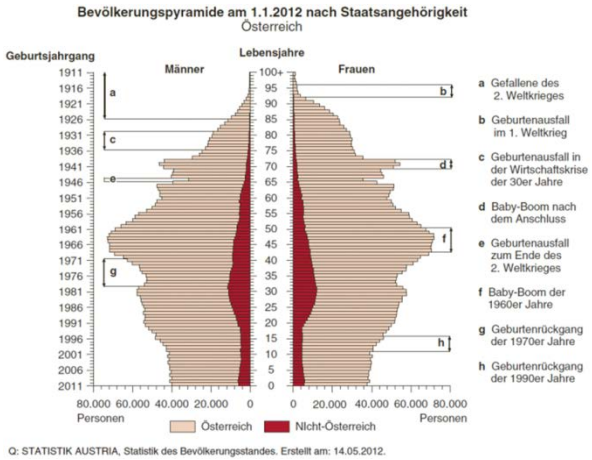
Inferenzstatistik

- Induktive (schließende, beurteilende) Statistik versucht aus den erhobenen Daten Schlüsse auf Ursachenkomplexe zu ziehen, die diese Daten produziert haben
- Schluss von Teilgesamtheiten auf Grundgesamtheiten
- Erkennen von allgemeinen Gesetzmäßigkeiten, die über den Beobachtungsbereich hinaus Gültigkeit besitzen; Prognose und Vorhersagemethoden (statistisches Schätzen)
- Überprüfung von Hypothesen - statistische Entscheidungstheorie (statistisches Testen)
- Modell des Zufalls "Sicherheit über Unsicherheit zu gewinnen“ - „Der Gezähmte Zufall“

Demographische Prognose



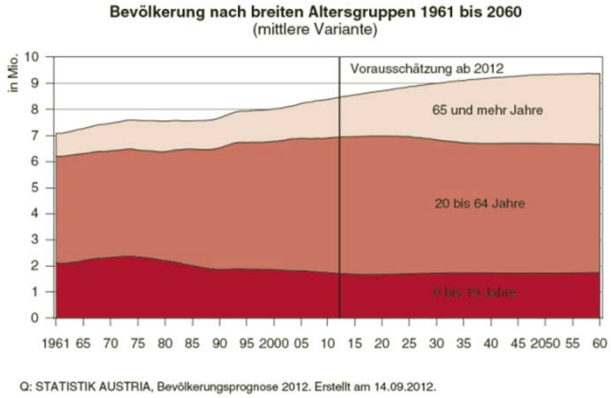
Q: STATISTIK AUSTRIA, Bevölkerungsprognose 2012. Erstellt am 14.09.2012.



Q: STATISTIK AUSTRIA, Statistik des Bevölkerungsstandes. Erstellt am: 14.05.2012.

Bild zur quantitativen Beschreibung der Ist-Situation

DESKRIPTIVE Beschreibung des IST-Zustandes



Q: STATISTIK AUSTRIA, Bevölkerungsprognose 2012. Erstellt am 14.09.2012.

PROGNOSTIK von Entwicklungen

Mittels eines multiregionalen Anwendung der Kohorten-Methode basiert werden unter Annahmen über die künftige Entwicklung der demographischen Indikatoren zu Fertilität, Mortalität und Migration Prognoseszenarien entwickelt

Typische Fragestellungen

- **Moderne Umfrageforschung**
  Mit welcher Präzision kann aus der Befragung einer Stichprobe auf die Meinung der Grundgesamtheit geschlossen werden?

- **Wirksamkeit von neuen Substanzen in der Medizin**
  Unterstützen die empirisch gewonnen Daten eine theoretische Hypothese oder widersprechen sie dieser Hypothese?

Was ist Statistik ?

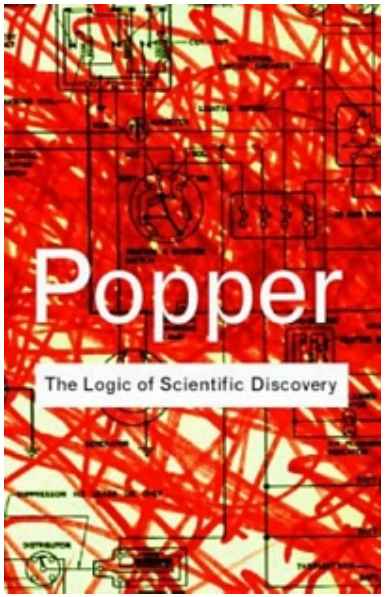
Statistik ist eine Zusammenfassung von Methoden, die uns erlauben, vernünftige optimale Entscheidungen im Falle von Ungewissheit zu treffen.

A. Wald

Teilgebiete der Statistik

- Deskriptive Statistik - Inferenzstatistik
- Explorative Statistik - Konfirmatorische Statistik
- Angewandte Statistik - Theoretische Statistik

Explorative und konfirmatorische Sicht auf Daten



Die Tätigkeit des wissenschaftlichen Forschers besteht darin, Sätze oder Systeme von Sätzen aufzustellen und systematisch zu überprüfen; in den empirischen Wissenschaften sind es insbesondere die Hypothesen, Theoriensysteme, die aufgestellt und an der Erfahrung, durch Beobachtung und Experiment überprüft werden.

K. Popper (1934)

Hypothesenerstellung <===> Hypothesenprüfung

explorative Sicht <===> konfirmatorische Sicht

Prinzipien konfirmatorischer Statistik

- Vor Durchführung der empirischen Studie ist es erforderlich, dass eine theoretisch gut begründete Hypothese formuliert wird.
- Die Planung der Datenauswertung muss vor Ansicht der empirischen Daten erfolgen
- Die datengesteuerte Spezifikation verringert den Erkenntniswert
- Bei der wahllosen Anwendung statistischer Methoden können wertlose Zufallsbefunde entstehen

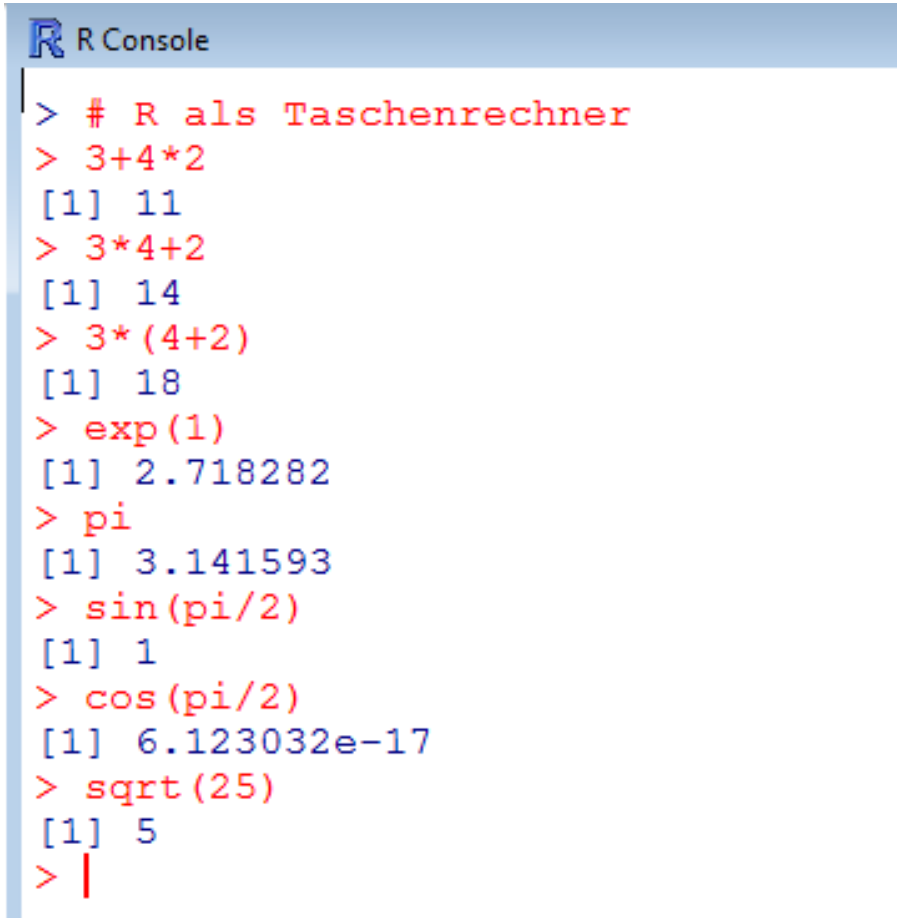
Phasen konfirmatorischer empirischer Forschung

- Erkundungsphase
- Theoretische Phase
- Planungsphase
  - Auswahl der Merkmale, Art der Datenerhebung, Planung der Stichprobe
- Planung der statistischen Auswertung
- Untersuchungsphase
- Auswertungsphase
- Entscheidungs- bzw. Beurteilungsphase

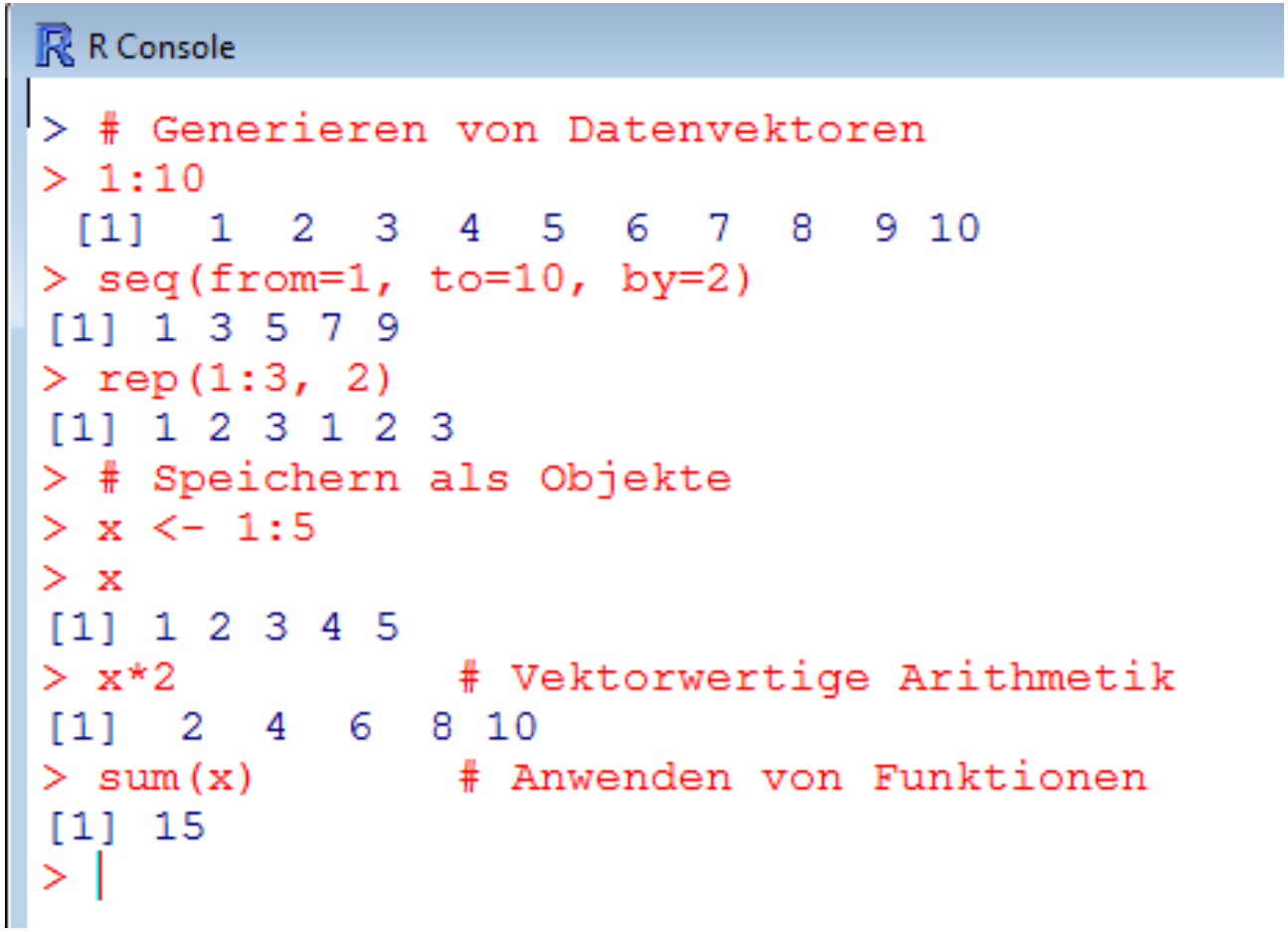
Teilgebiete der Statistik

- Deskriptive Statistik - Inferenzstatistik
- Explorative Statistik - Konfirmatorische Statistik
- Angewandte Statistik - Theoretische Statistik

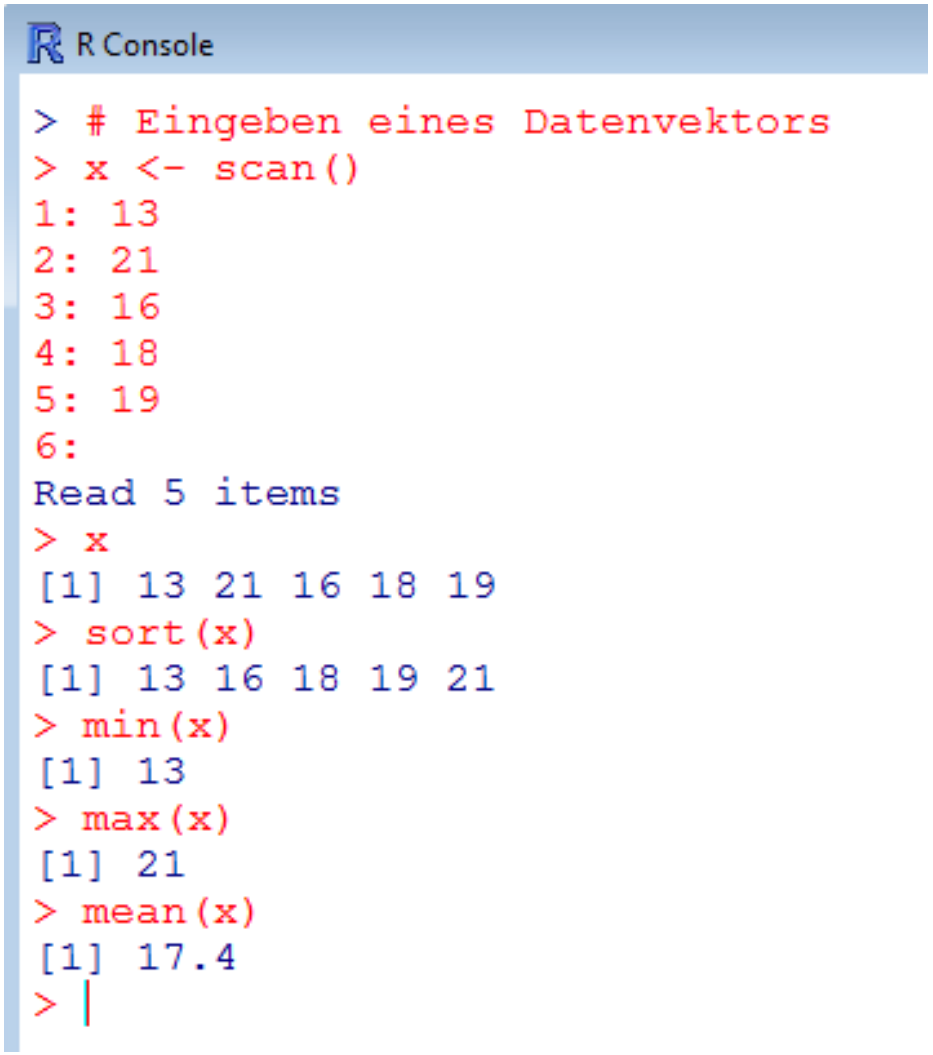
Eine erste R-Session



Anlegen von Datenvektoren



Eingeben eines Datenvektors

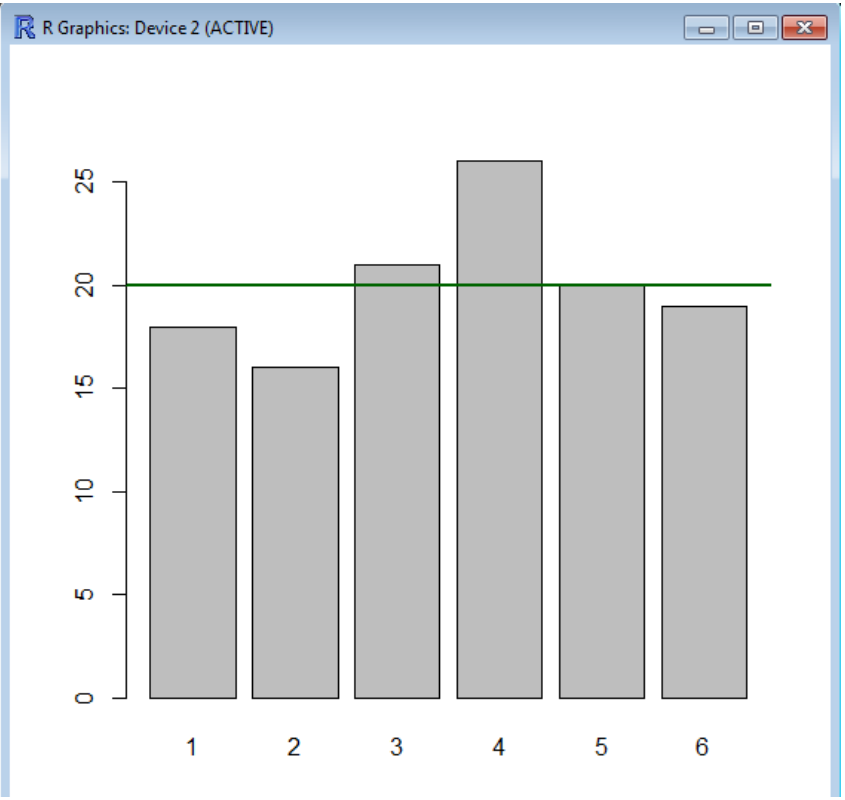


Simulation eines Würfelwurfs

R Console

```
> # Simulation eines Würfelwurfs
> ergebnis <- sample(1:6, 120, replace=T)
> ergebnis
  [1] 3 4 2 1 4 5 4 3 1 4 6 6 4 3 6 6 1 1 5 1 1 2 4 2 2 5 2
 [28] 5 3 4 4 2 3 5 4 5 6 4 6 1 5 6 1 3 1 3 2 4 6 4 6 6 5 4
 [55] 4 6 6 3 2 2 1 4 5 2 6 6 2 6 3 6 2 2 5 1 1 5 3 4 2 5 2
 [82] 4 3 6 3 5 6 5 1 3 3 1 3 1 2 4 3 4 5 3 3 1 4 5 5 4 3 5
[109] 4 6 1 3 3 1 4 5 5 4 4 4
> table(ergebnis)
ergebnis
 1  2  3  4  5  6
18 16 21 26 20 19
> mean(ergebnis)
[1] 3.591667
> barplot(table(ergebnis))
> abline(h=20, col="darkgreen", lwd=2)
>
```

R Graphics: Device 2 (ACTIVE)

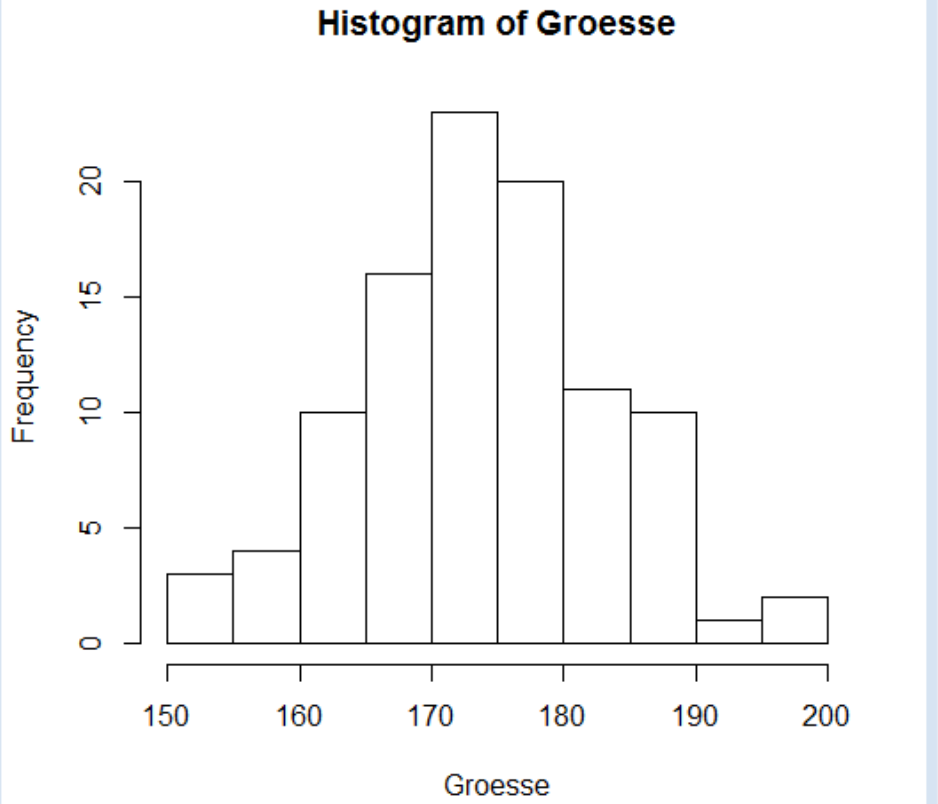


Einlesen von Daten

R Console

```
> # Analyse von Daten
> daten <- read.csv(file="C:\\work\\data\\studenten.csv")
> # Erzeugt einen Data Frame
> head(daten)
  Groesse
1     170
2     183
3     154
4     186
5     165
6     180
> mean(daten$Groesse)
[1] 174.4
> # Einfacher
> attach(daten)
> mean(Groesse)
[1] 174.4
> hist(Groesse)
>
```

R Graphics: Device 2 (ACTIVE)



Datenquellen und Erhebungsarten

- Art der Erhebung
  - Beobachtung versus Experiment
  - Labor- oder Felduntersuchung
  - Befragung (persönlich/schriftlich/telefonisch/webbasierend)
- Umfang der Erhebung
  - Vollerhebung (Totalerhebung) "census"
  - Teilerhebung(Stichprobenerhebung) "sample"
- Quelle der Erhebung
  - Primärstatistik
  - Sekundärstatistik

Quelle der Erhebung

- Primär- versus Sekundärstatistik
- Primärstatistik bedeutet, dass die Daten eigens für den Untersuchungszweck erhoben werden
  - Nachteil: hoher Aufwand an Geld und Zeit
  - Vorteil: Daten können optimal auf den Forschungszweck abgestimmt werden
- Sekundärstatistik bedeutet, dass bereits bestehende Daten zur Untersuchung herangezogen werden.
  - Nachteil: keine spezifische Ausrichtung am Forschungsdesign
  - Vorteil: geringere Kosten, rasche Verfügbarkeit

Arbeitsmarktforschung

- Beispiel:

STATISTIK AUSTRIA
Die Informationsmanager

Offene-Stellen-Erhebung - Wichtigste Eckpunkte	
Gegenstand der Statistik	Offene sowie besetzte Stellen
Grundgesamtheit	Rund 249.000 Unternehmen mit mind. einem unselbständig Beschäftigten in Österreich
Statistiktyp	Primärstatistische Stichprobenerhebung der offenen Stellen, Sekundärstatistik der unselbständig Beschäftigten laut Hauptverband

Basierend auf der Anzahl der offenen Stellen und der Anzahl der Beschäftigten wird die Offene-Stellen-Quote berechnet.
Die Offene-Stellen-Quote ist einer der wichtigsten europäischen ökonomischen Indikatoren (Principal European Economic Indicators).

Begriffstrukturierung

- Statistische Masse (Gesamtheit, Grundgesamtheit, Population)
  - sachlich, örtlich und zeitlich abgegrenzte Menge von Merkmalsträgern (statistischen Einheiten)
- Merkmalsträger (statistische Einheiten, Untersuchungseinheiten)
  - Personen, Objekte oder Ereignisse oder Zusammenfassungen davon (z.B. Haushalte), die einer statistischen Untersuchung zugrunde liegen und durch bestimmte Eigenschaften gekennzeichnet sind
- Merkmale (statistische Variable)
  - Messbare Eigenschaften eines Merkmalsträgers

Messen & Merkmalstypen

- Messen:
  Zuordnung von Symbolen oder Zahlen (Merkmalsausprägungen) zu den Merkmalsträgern
- Merkmalstypen
  - Merkmale können aufgrund der Anzahl der Ausprägungen typisiert werden
  - Merkmale können aufgrund der Eigenschaften der zugrundeliegenden Skala typisiert werden

Typologie von Merkmalen

	Informationsqualität
Nominalskalierte Merkmale (klassifikatorische, nominale) z.B. Geschlecht, Religionsbekenntnis	*Unterscheidung*
Ordinalskalierte Merkmale (komparative, ordinale, rangskalierte) z.B. Schulnoten, Güteklassen	*Rangfolge*
Intervallskalierte Merkmale z.B. Jahreszahlen, Temperatur in ◦C	*Abstand*
Verhältnisskalierte Merkmale z.B. Alter, Monatseinkommen	*Verhältnis*
Absolutskalierte Merkmale z.B. Kinderanzahl, Verkehrstote pro Jahr	*absolute Einheit*

Sinnvolle Operationen

Skalenart	Auszählen von Häufigkeiten	Ordnen, Bilden von Rangfolgen	Bilden von Differenzen	Bilden von Quotienten
Nominalskala	ja	nein	nein	nein
Ordinalskala	ja	ja	nein	nein
Intervallskala	ja	ja	ja	nein
Verhältnisskala	ja	ja	ja	ja

Informationserhaltende Transformationen

Nominalskala: jede bijektive Transformation (eindeutige Umbenennung)

$$g: x \to y \qquad x_1 \neq x_2 \Leftrightarrow g(x_1) \neq g(x_2)$$

Ordinalskala: jede rangerhaltende (streng monotone) Transformation

$$g: x \to y \quad x_1 > x_2 \Rightarrow g(x_1) > g(x_2)$$

Intervallskala: lineare Transformationen (Verschiebung und Streckung/Stauchung)

$$g: x \to y \quad y = a + bx$$

Verhältnisskala: Ähnlichkeitstransformationen (Streckung/Stauchung)

$$g: x \to y \quad y = ax$$

Absolutskala: keine

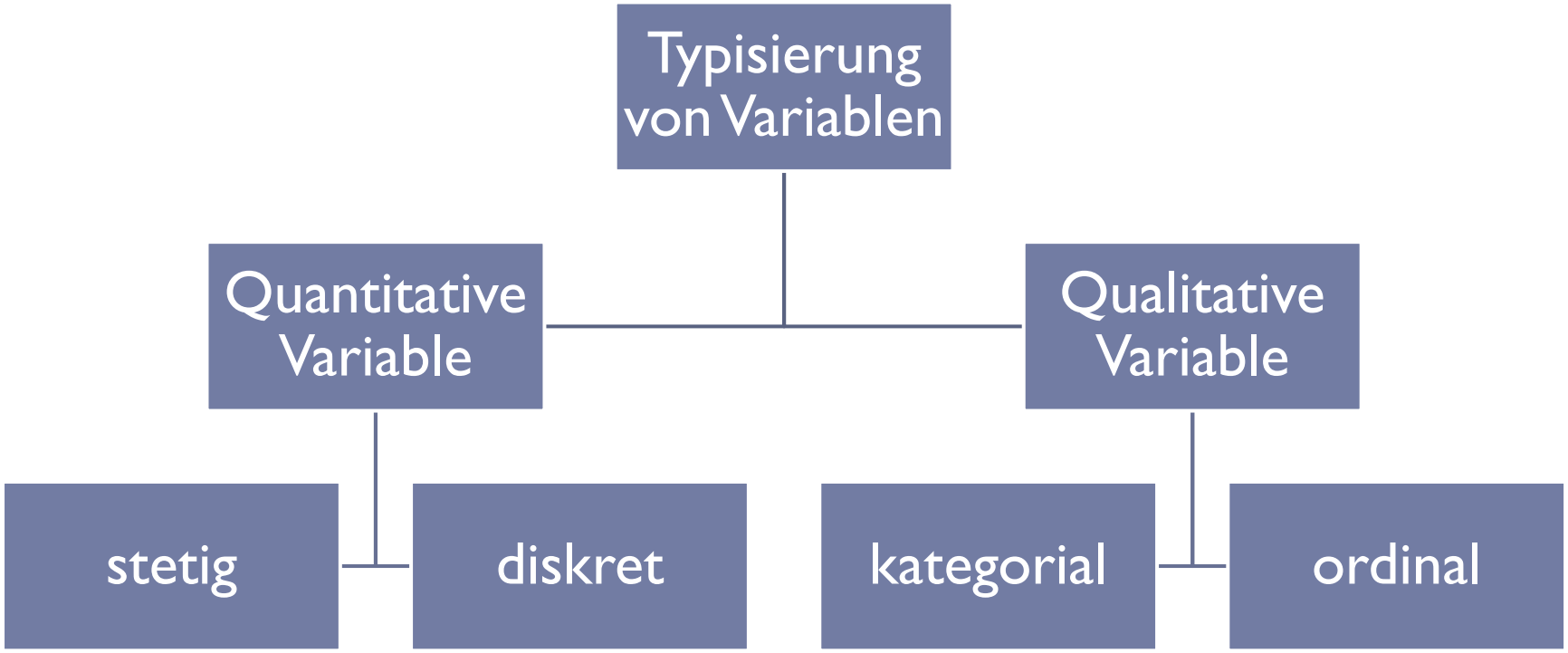
Quantitativ versus Qualitativ

- Quantitative Merkmale
  Ausprägungen unterscheiden sich durch die Größe
  ~ „Ausprägungen sind Zahlen“
  Intervallskala, Verhältnisskala, Absolutskala
- Qualitative Merkmale
  Ausprägungen unterscheiden sich durch die Art
  Ausprägungen zeigen die Zugehörigkeit zu einer Gruppe an
  ~ „Ausprägungen sind keine Zahlen"
  Nominalskala, Ordinalskala

Typisierung quantitativer Merkmale

- DISKRETE versus STETIGE Merkmale
- STETIGE MERKMALE (continuous variables)
- Merkmale, bei denen sich die Merkmalswerte selten wiederholen, d.h. viele Merkmalsträger weisen unterschiedliche Werte auf
- Beispiel: Körpergröße in mm
- DISKRETE MERKMALE (discrete variables)
- Merkmale, mit wenigen unterschiedlichen Ausprägungen
- Beispiel: Ausbildungszeit in Jahren
- Beachte: Stetige Merkmale können durch Vergröberung immer diskretisiert werden (Kategorienbildung)

Schematische Typisierung von Variablen



Terminologie

Objekt

- Untersuchungs-einheit
- Merkmalsträger
- Fall (case)

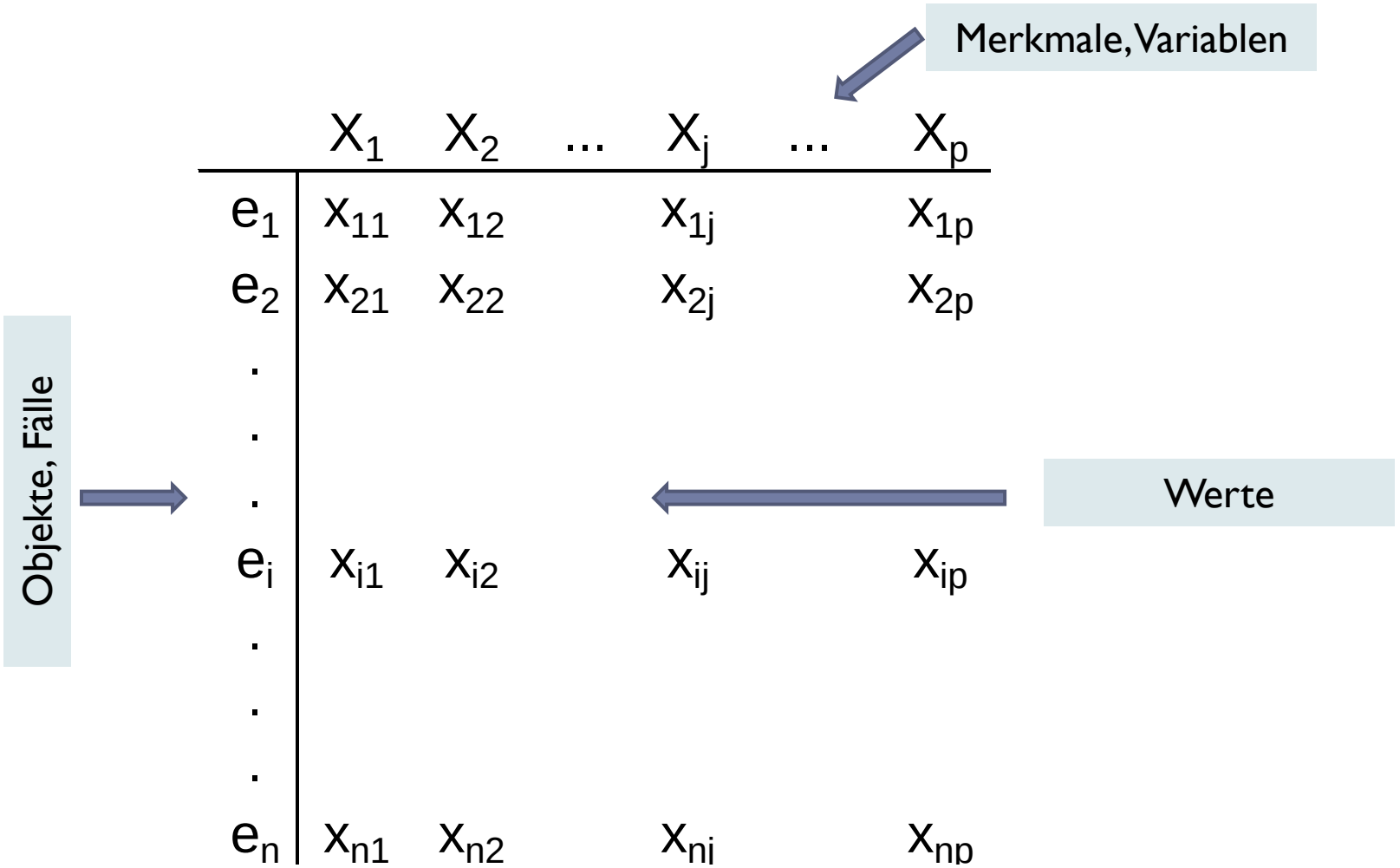
Eigenschaft

- Attribut
- Merkmal
- Variable (variable)

konkrete Eigenschaft

- Wert des Attributs
- Ausprägung des Merkmals
- Wert (value)

Datenmatrix



Erfassen einer Datenmatrix

R Console

```
> # Erfassen einer Datenmatrix
> Daten <- data.frame()
> fix(Daten)
```

Dateneditor

	var1	var2	var3	v
1				
2				
3				
4				
5				
6				
7				

Dateneditor

	Name	Alter	Taschengeld	va
1	Franz	13	10	
2	Martin	14	20	
3	Karl	14	10	
4	Sabine	13	20	
5	Josef	13	5	
6	Anna	14	20	
7				
8				

Arbeiten mit dem Dataframe

R Console

```
    Name Alter Taschengeld
1  Franz    13          10
2 Martin    14          20
3   Karl    14          10
4 Sabine    13          20
5  Josef    13           5
6   Anna    14          20
> mean(Taschengeld)          # Keine Namensauflösung möglich
Fehler in mean(Taschengeld) : Objekt 'Taschengeld' nicht gefunden
> mean(Daten$Taschengeld)    # Ansprechen einer Spalte
[1] 14.16667
> search()
[1] ".GlobalEnv"        "package:stats"     "package:graphics"
[4] "package:grDevices" "package:utils"     "package:datasets"
[7] "package:methods"   "Autoloads"         "package:base"
> attach(Daten)              # Erweitern der Searchliste
> search()
 [1] ".GlobalEnv"         "Daten"              "package:stats"
 [4] "package:graphics"   "package:grDevices" "package:utils"
 [7] "package:datasets"   "package:methods"   "Autoloads"
[10] "package:base"
> mean(Taschengeld)          # Namensauflösung möglich
[1] 14.16667
>
> tapply(Taschengeld, Alter, mean)
      13       14
11.66667 16.66667
```

Verwenden von vorhandenen Daten

R Console

```
> # Verwenden vorhandener Daten
> data()
> help(faithful)
> attach(faithful)
> plot(waiting, eruptions)
> abline(lsfit(waiting, eruptions), col="red", lwd=2)
> title(main="Beispiel für eine lineare Trendrechnung")
>
```

R data sets

```
euro.cross (euro)      Conversion Rates of Euro Currencie
eurodist               Distances Between European Cities
faithful               Old Faithful Geyser Data
fdeaths (UKLungDeaths)
                       Monthly Deaths from Lung Diseases
freeny                 Freeny's Revenue Data
freeny.x (freeny)      Freeny's Revenue Data
freeny.y (freeny)      Freeny's Revenue Data
infert                 Infertility after Spontaneous and
                       Abortion
iris                   Edgar Anderson's Iris Data
iris3                  Edgar Anderson's Iris Data
islands                Areas of the World's Major Landmas
ldeaths (UKLungDeaths)
```

faithful {datasets}

R Documentation

Old Faithful Geyser Data

Description

Waiting time between eruptions and the duration of the eruption for the Old Faithful geyser in Yellowstone National Park, Wyoming, USA.

Usage

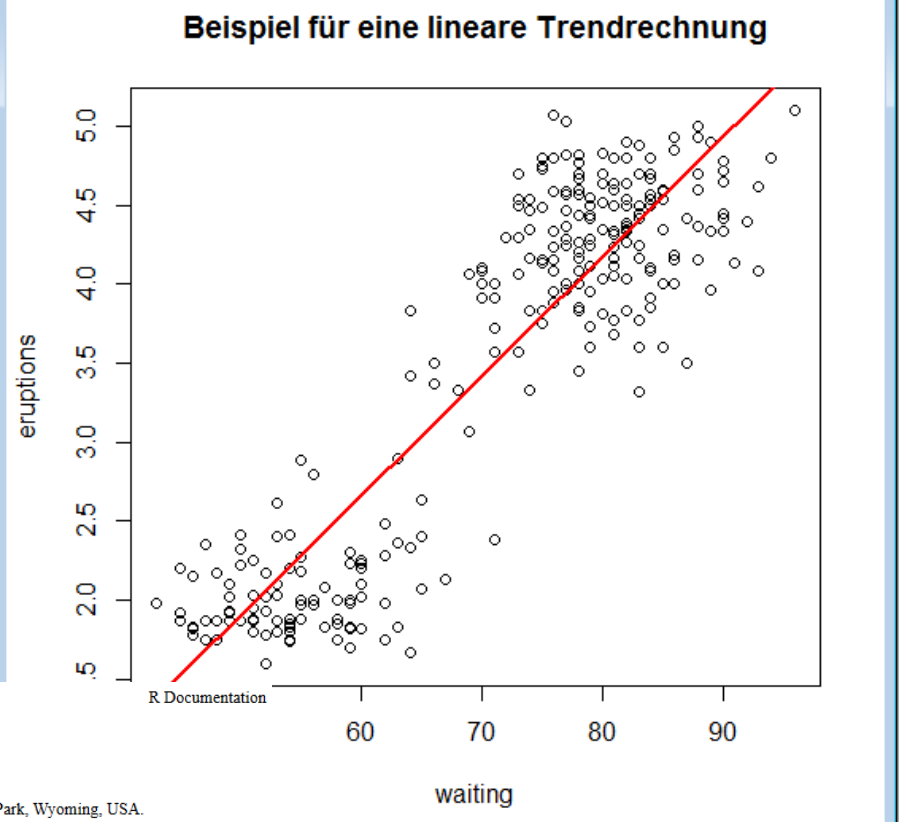
```
faithful
```

Format

A data frame with 272 observations on 2 variables.

[,1] eruptions numeric Eruption time in mins
[,2] waiting numeric Waiting time to next eruption (in mins)

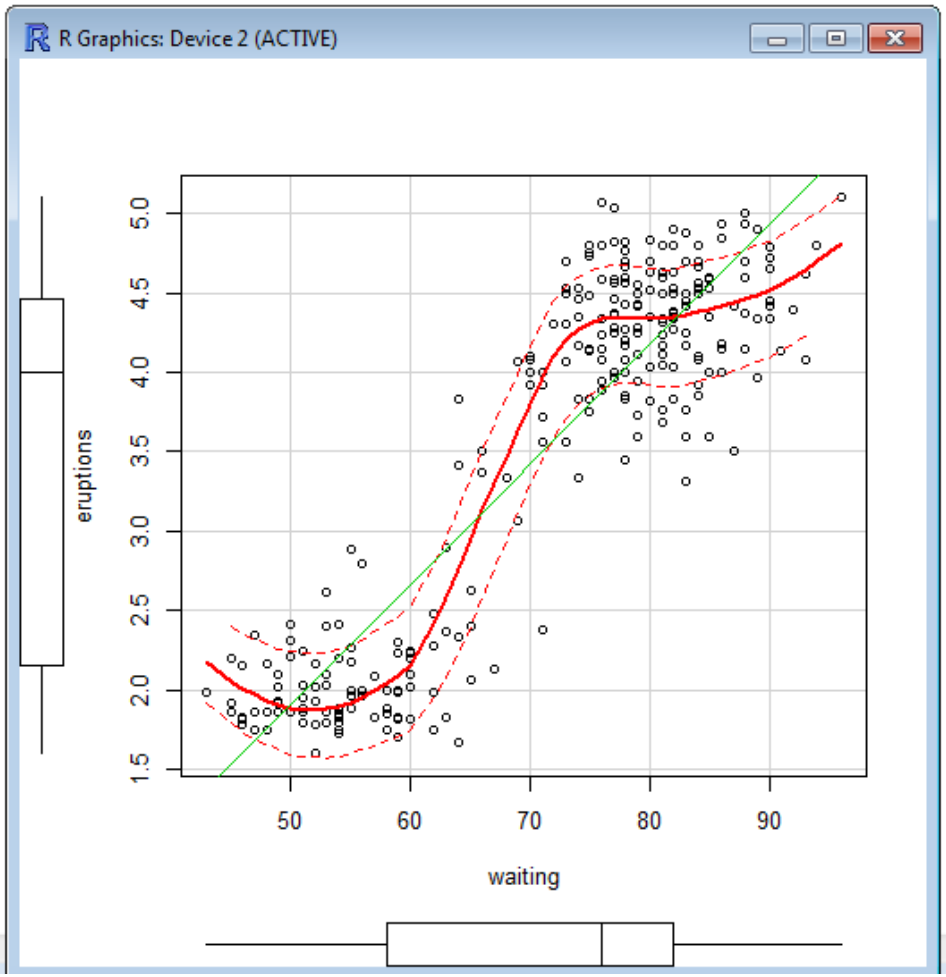
R Graphics: Device 2 (ACTIVE)



More Sophisticated Plot

R Console

```
> library(car)
> scatterplot(eruptions ~ waiting, data=faithful)
>
```



Statistical thinking will one day be as necessary for efficient citizenship as the ability to read and write.

H.G. Wells