

Statistik 1 für Studierende der Soziologie, des Nebenfachs Statistik, der Medieninformatik und der Cultural and Cognitive Linguistics

Thomas Augustin & Georg Schollmeyer

WS 2016/2017

Bitte beachten Sie: Dieses zur Verfügung gestellte Material ist kein Skript. Es soll ihnen beim Mitschreiben helfen, indem es schreibintensivere Teile getippt zur Verfügung stellt. Inhaltlich wesentliche Erläuterungen und Beispiele werden langsam in der Vorlesung entwickelt.

Besonderer Dank gilt Prof. Dr. Thomas Kneib (jetzt Georg-August-Universität Göttingen), Prof. Dr. Carolin Strobl (jetzt Universität Zürich), Prof. Dr. Helga Wagner (Johannes Kepler Universität Linz) und Dr. Fabian Scheipl, die im Rahmen ihrer Vorlesungen an der LMU im WS 07/08, WS 08/09, WS 10/11 und WS 12/13 das ursprüngliche Material weiterentwickelt haben.

1 Einführung und erste Grundbegriffe

1.1 Vorbemerkungen zur Organisation, Bedeutung und Struktur der Veranstaltung

1.1.1 Die Dozenten

- Vorlesung durch Thomas Augustin (erster Teil) und Georg Schollmeyer (Institut für Statistik, AG Method(olog)ische Grundlagen der Statistik)
- Büro: Ludwigstrasse 33, 2. Stock, Zi. 250; Zi. 248
- Kontakt Augustin:
 - Sprechstunde Mittwochs 17.15-18.15 Uhr (um Voranmeldung wird im eigenen Interesse dringend gebeten; kurze Fragen natürlich auch nach der Vorlesung)
 - Email: augustin@stat.uni-muenchen.de
- Homepage der Arbeitsgruppe <http://www.statistik.lmu.de/institut/ag/agmg/index.html>

1.1.2 Organisatorisches I

Veranstaltungshomepage: http://www.statistik.lmu.de/institut/ag/agmg/lehre/2016_WiSe/StatISoz/index.html

1.1.3 Zum Rahmen der Veranstaltung I: Vorbemerkungen und Ihre Erwartungen

Was erwarten Sie von einer Statistik Vorlesung?

“I keep saying that the sexy job in the next 10 years will be statisticians. And I’m not kidding.”(Hal Varian, Chief Economist at Google, McKinsey Quarterly (Jan. 2009).)

Ausführlicher lautet das Zitat:

“I keep saying the sexy job in the next ten years will be statisticians. The ability to take data - to be able to understand it, to process it, to extract value from it, to visualize it, to communicate it that's going to be a hugely important skill in the next decades, not only at the professional level but even at the educational level for elementary school kids, for high school kids, for college kids. Because now we really do have essentially free and ubiquitous data. So the complimentary scarce factor is the ability to understand that data and extract value from it.”

- Wells/Wilks on Statistical Thinking

<http://www.causeweb.org/resources/fun/db.php?id=105> Aufruf 13.10.16

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

Quote from the presidential address in 1951 of mathematical statistician Samuel S. Wilks (1906 - 1964) to the American Statistical Association found in JASA, Vol. 46, No. 253., pp. 1-18. Wilks bezieht sich auf 'Mankind in the Making.' H.G. Wells schreibt darin:

“The great body of physical science, a great deal of the essential fact of financial science, and endless social and political problems are only accessible and only thinkable to those who have had a sound training in mathematical analysis, and the time may not be very remote when it will be understood that for complete initiation as an efficient citizen of one of the new great complex worldwide States that are now developing, it is as necessary to be able to compute, to think in averages and maxima and minima, as it is now to be able to read and write.”

Aus: Samuel S. Wilks, Presidential Adress, Journal of the American Statistical Association (1951), 46, 1-18, hier S.5

1.1.4 Wie tickt ein(e) Statistiker(in)? Ein erster Versuch

Cox & Snell (Applied Statistics, 1981, Chapman & Hall, p. 3):

„Statistical Analysis deals with those aspects of the analysis of data that are not highly specific to particular fields of study. That is, the object is to provide concepts and methods that will, with suitable modification, be applicable in many different fields of application; indeed one of the attractions of this subject is precisely this breadth of potential applications.“

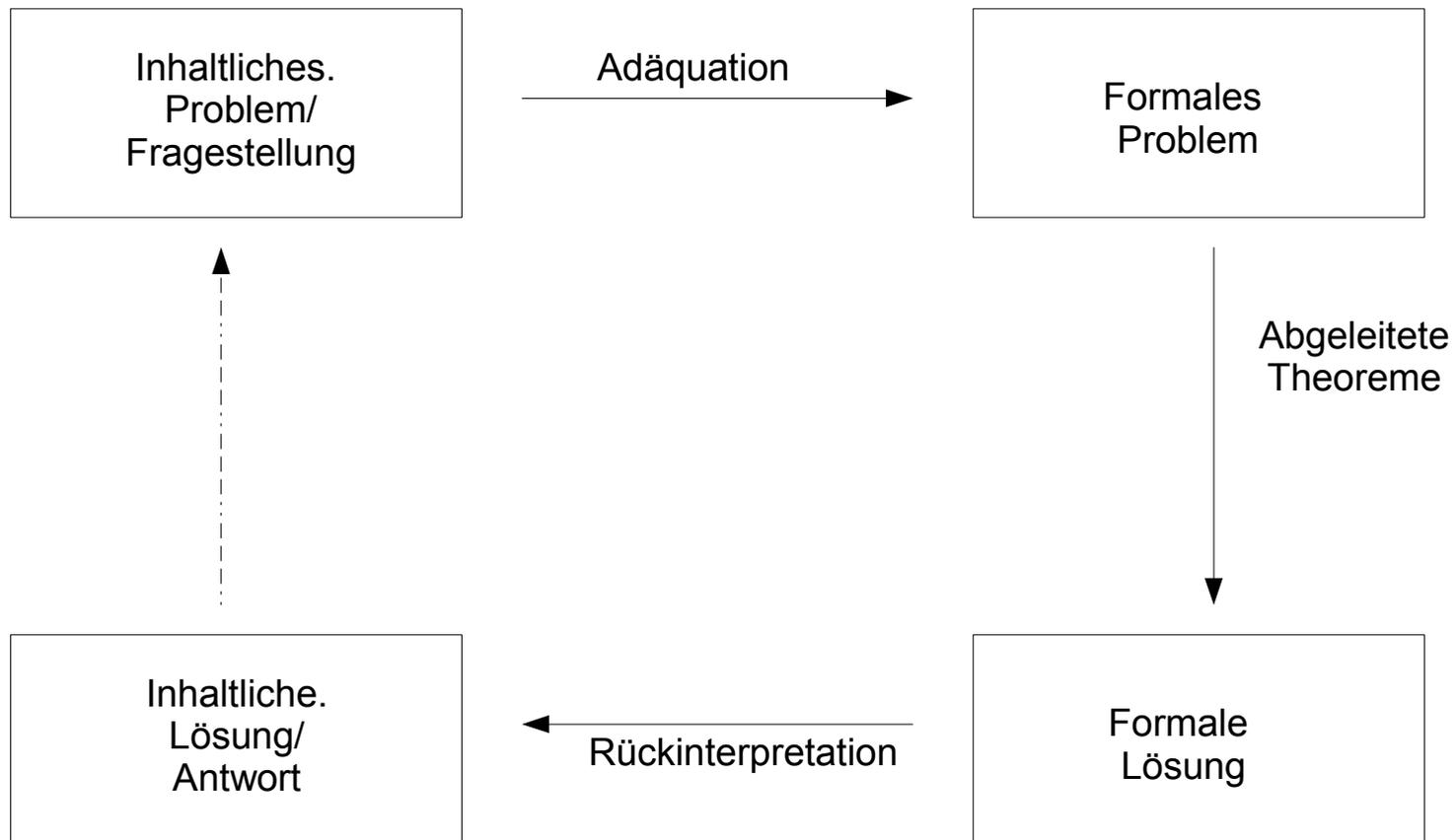
- Abstrahiere von der konkreten Fragestellung: Ordne konkretes Problem in eine *Klasse* von Problemstellungen ein \Rightarrow abstrakte Fragestellung
- Wähle bzw. entwickle für diese Klasse von Problemstellungen das richtige Verfahren \Rightarrow abstrakte Lösung
- Adäquationsproblem: Die Einordnung in eine Klasse muss *strukturerhaltend* (“nicht inhalts-zerstörend“) sein.
Die adäquate Abstraktion heißt nicht nur Vernachlässigung von Unwesentlichem, sondern auch Herausarbeiten der grundlegenden Struktur des Problems. Gehen wesentliche formale Strukturen des Ausgangsproblems verloren, so schwächt bzw. zerstört dies natürlich die Aussagekraft der Ergebnisse der Analyse.

Um einen ersten Eindruck zu erhalten, ein paar komplexe Beispiele. Welche lassen sich als eine gemeinsame Klasse von Problemstellungen zugehörig zusammenfassen?

- 1.) Welche Bedeutung haben Ernährungsgewohnheiten für das Auftreten von Herz-Kreislaufproblemen?
- 2.) Welche Bedeutung für das Scheitern von Paarbeziehungen haben starke Rollenkonflikte am Arbeitsplatz des/der Hauptverdienenden?
- 3.) Welche Rolle spielt die formale Bildung für das durchschnittliche Erwerbseinkommen?
- 4.) Welcher Anteil an abhängig Beschäftigten erzielte letztes Jahr ein Einkommen $\tilde{A}_{\frac{1}{4}}$ über 100.000 Euro?
- 5.) Welche Rolle spielt bei Arbeitslosen die erfolgreiche Teilnahme an Fortbildungsmassnahmen?

- 6.) Welche Rolle spielen bei den heute 40 bis 60 Jährigen Stigmatisierungserfahrungen in der eigenen Jugend bei der aktuellen Partnerpräferenz?
- 7.) Welche Rolle spielen Rabattaktionen bei der Kundenbindung?
- 8.) Wieviele Diebstähle wurden während des letzten Oktoberfests polizeilich gemeldet?
- 9.) Reduziert eine begleitende Gesprächstherapie die Rückfallgefahr nach einem erfolgreichen Drogenentzug?
- 10.) Ist das neue Medikament wirksamer zur Blutdrucksenkung als das alte?

Das Grundschema der Formalisierung



Ziel statistischer Methodik leistungsfähige Auswertungsregeln für Klassen von Problemstellungen entwickeln

- 'Klassen': Problemstellungen mit gleicher Struktur (und nur diese) zusammenfassen
- 'Regeln': (Hauch von) Objektivität, Konsens über Ergebnisse, zumindest intersubjektive Überprüfbarkeit
- 'leistungsfähig': Vergleich der Leistungsfähigkeit von prinzipiell geeigneten Verfahren und Entwicklung von auf bestimmte Spezifika zugeschnittene Verfahren ist wesentlicher Gegenstand fortgeschrittener Veranstaltungen
- Hier werden nur Standardmethoden vorgestellt. Beurteilung, ob geeignet oder nicht, Eignung heißt v. a. Zuordnung zur richtigen Klasse von Problemstellungen, Abstraktionsschritt muss inhaltsverdichtend sein, ohne inhaltszerstörerisch zu sein

1.1.5 Lernziele

Grundsätzliches Ziel der Veranstaltung ist es, dass Sie in die Lage versetzt werden, statistisch argumentieren zu können und auf einem elementaren Niveau aus Daten die richtigen substanzwissenschaftlichen Schlüsse zu ziehen.

Hierzu gehören insbesondere die Fähigkeiten,

1. die Anwendung statistischer Methodik und die gewonnenen Ergebnisse in der Literatur kritisch zu hinterfragen,
2. selbst korrekte statistische und probabilistische (=wahrscheinlichkeitsbezogen) Argumentationen aufbauen zu können,
3. grundlegende Auswertungen selbst durchzuführen und die Ergebnisse korrekt zu interpretieren,
4. sich bei Bedarf eigenständig weiterführende Verfahren anzueignen.

1.1.6 Organisatorisches II

- Vorlesung: Dienstag und Donnerstag
- Vorschlag: Di 16.10 – 17.55 Uhr, Do 14.10 – 15.55 Uhr, dafür letzte Stunden ausfallen lassen
- Übungen: Mittwoch 10 – 12 oder 12 – 14 Uhr
- Tutorium Studierende der Medieninformatik: Mittwoch 12 – 14 Uhr
- Als Zusatzangebot: Tutorium v. a. für Erstsemester und Wiederholende mit Schwierigkeiten: Montag 16 – 18 Uhr ¹

¹Bitte melden Sie sich mit Begründung, wenn Sie am Tutorium sehr gerne teilnehmen möchten, aber wegen einer Überschneidung verhindert sind. augustin@stat.uni-muenchen.de

1.1.7 Zur Rolle der Statistik in Ihrem Studium

- BA Soziologie
- Statistik als Nebenfach für Studierende der Soziologie
- Nebenfach Statistik für andere Studiengänge
- Studierende der Medieninformatik
- Masterstudiengang 'Cultural and Cognitive Linguistics (CCL)'

1.1.8 Einige Studien- und Überlebenstipps

Zum Verhältnis Vorlesung - Übung - Tutorium

Die Veranstaltungen (Vorlesungen, Übungen, Tutorium) sind ein ANGEBOT

Mathe-Auffrischung:

- Material zum Formalisierungspropädeutikum

Lerngruppen bilden

Unbedingt dabeibleiben (strikt linearer Aufbau)

1.1.9 Zur besonderen (?) Struktur der Veranstaltung

1.1.10 'Benutzervertrag'

1.1.11 Ein überraschend hilfreicher Vergleich: Statistik als Sprache mit Mathematik als Grammatik/Syntax

1.1.12 Material zur Veranstaltung

Nochmals Homepage:

http://www.statistik.lmu.de/institut/ag/agmg/lehre/2016_WiSe/StatISoz/index.html

Achtung: Wegen eines zentralen Umzugs der Hompages des Statistik-Instituts wird sich diese Adresse leider im Laufe des Semesters ändern.

- Begleitmaterial (kein Skript) zur Erleichterung des Mitschreibens
- Übungsblätter
- Formelsammlung

1.1.13 Studierende in besonderen Lebenslagen

Studierende mit einer Behinderung oder einer chronischen Erkrankung bzw. die betreuungsbedürftige Familienangehörige haben oder schwanger sind: Melden Sie sich bei uns, wenn Sie denken, dass wir etwas für Sie tun können, direkt, oder über

- Beratungsstelle der LMU für behinderte und chronisch kranke Studierende: LMU, Ludwigstr. 27/I, Zimmer G 122, 80539 München, Sprechzeiten nach Vereinbarung, 089 2180-2963

http://www.uni-muenchen.de/studium/beratung/beratung_service/beratung_lmum/barrierefrei_stud/index.html

- Beratungsstelle für Schwangere und Studierende mit Kind(ern), LMU, Ludwigstr. 27/I,

Montags-Mittwochs 9 bis 12 Uhr 089 2180-3124,

http://www.uni-muenchen.de/studium/beratung/beratung_service/beratung_lmum/schwangere_kind/index.html

1.2 Was soll Statistik (nicht)?

1.2.1 Beispiele statistischer Erhebungen und Fragestellungen

Münchener Mietspiegel (fortlaufendes Beispiel in Fahrmeir et al.)

- Mietspiegel bieten Mietern und Vermietern eine Übersicht zu den sogenannten „ortsüblichen Vergleichsmieten“.
- Ortsüblichen Vergleichsmiete: „die üblichen Entgelte, die in der Gemeinde X für nicht preisgebundenen Wohnraum vergleichbarer Art, Größe, Beschaffenheit und Lage in den letzten vier Jahren vereinbart (. . .) oder geändert worden sind.“²
- Statistische Fragestellung: Wie beeinflussen Merkmale einer Wohnung (Wohnfläche, Baujahr, Küchenausstattung, etc.) die Nettomiete (pro Quadratmeter)?
- Den aktuellen Mietspiegel für München finden Sie unter

<http://www.mietspiegel-muenchen.de>

²<http://www.mietspiegel-muenchen.de/2013/gesetzeslage.html>, aufgerufen am 17.10.16

Sozio-ökonomisches Panel (SOEP)

- Seit 1984 durchgeführte Befragung von deutschen Haushalten.
- 2015 waren etwa 19.000 (repräsentativ ausgewählte) Haushalte mit circa 37.000 Befragungspersonen beteiligt.
- Themenschwerpunkte: Haushaltszusammensetzung, Erwerbs- und Familienbiographie, Erwerbsbeteiligung und berufliche Mobilität, Einkommensverläufe, Gesundheit und Lebenszufriedenheit.
- Besonderheiten: Die gleichen Personen werden wiederholt befragt (Panelstudie).
Befragung auf Haushaltsebene
Freiwillige Teilnahme
Daten werden gegen Aufwandsentschädigung Forschern zur Verfügung gestellt.

Ausmaß des Mangels an quantitativ qualifizierten Absolventen sozialwissenschaftlicher Studiengänge

- Untersuchung von Rainer Schnell (Universität Duisburg-Essen)³
- Information aus dem Arbeitgeberinformationssystem (AIS): Daten über größten Teil der bundesweit arbeitslos gemeldeten Personen.
- Analyse von 1745 arbeitslosen Soziolog(innen) auf Beherrschung Statistik-Software (z.B. SPSS),
Spezielle Statistik-Kenntnisse,
Erfahrung bei der Durchführung quantitativer empirischer Projekte,
Erfahrung bei der Durchführung qualitativer empirischer Projekte,
“1“ bedeutet „ja“, “0“ bedeutet „nein“
- Qualifikationsprofile der am 1.6.01 arbeitslos gemeldeten Soziolog(inn)en:

³Schnell (2002): Ausmaß und Ursachen des Mangels an quantitativ qualifizierten Absolventen sozialwissenschaftlicher Studiengänge. In: Engel, Uwe (Hrsg.): Praxisrelevanz der Methodenausbildung. Informationszentrum Sozialwissenschaften, Bonn, S. 35–44

<i>N</i>	Prozent	SPSS	Statistik	Quantitativ	Qualitativ
1	0.06	1	1	1	1
1	0.06	1	1	0	1
2	0.11	0	1	0	1
3	0.17	0	0	1	1
7	0.40	1	0	0	1
13	0.74	1	0	1	0
18	1.03	1	1	1	0
26	1.49	0	1	1	0
28	1.60	1	1	0	0
34	1.95	0	0	0	1
80	4.58	1	0	0	0
93	5.33	0	1	0	0
97	5.56	0	0	1	0
1342	76.91	0	0	0	0

- Strenggenommen bräuchte man zum Vergleich noch die analoge Verteilung unter den erwerbstätigen Soziolog(inn)en; Zahlen sprechen aber trotzdem für sich

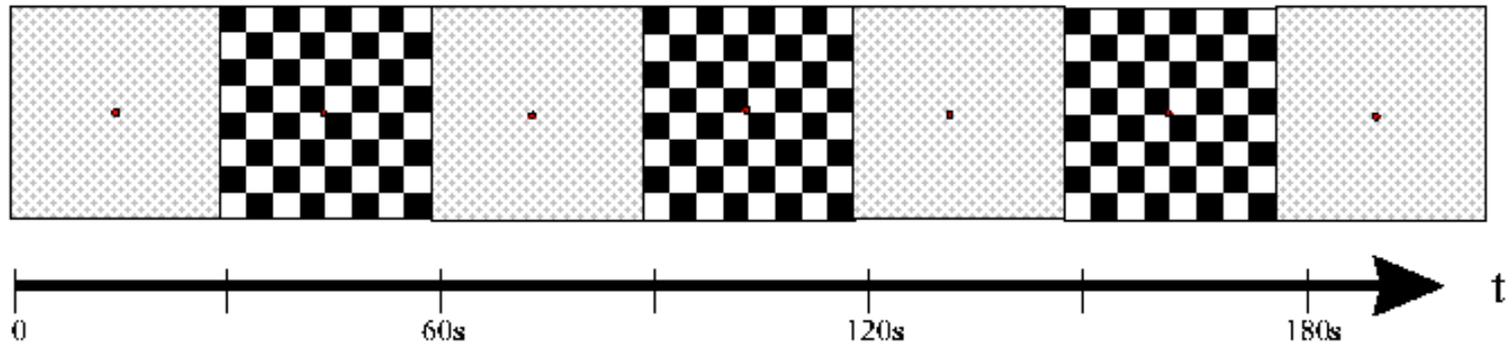
Beispiele statistische Untersuchungen (ausgewählte Projekte am Institut)

- Lebenszufriedenheit
- Waldschadensdaten
- Extreme Windgeschwindigkeiten an ICE-Strecken
- Auswirkungen von Luftverschmutzung
- Ernährungsgewohnheiten und Herz-Kreislaufkrankungen
- Strahlenbelastung (z.B. am Arbeitsplatz) und Krebs
- KfZ-Unfälle
- Gemeindestudien

- Bildungsbiographien
- Statistische Genetik
- Kreditwürdigkeitsprüfung
- Wirksamkeit eines Medikaments
- Qualitätskontrolle: Wann läuft ein Prozess (Schraube, Lungenfunktion) aus dem Ruder?
- Geschlechtsunterschiede bei innerbetrieblicher Mobilität
- Rückkehrchancen Erwerbsloser in den Arbeitsmarkt
- Human Brain Mapping
- Marktmechanismen bei illegalem Drogenhandel

Human Brain Mapping (Gehirnkartierung)

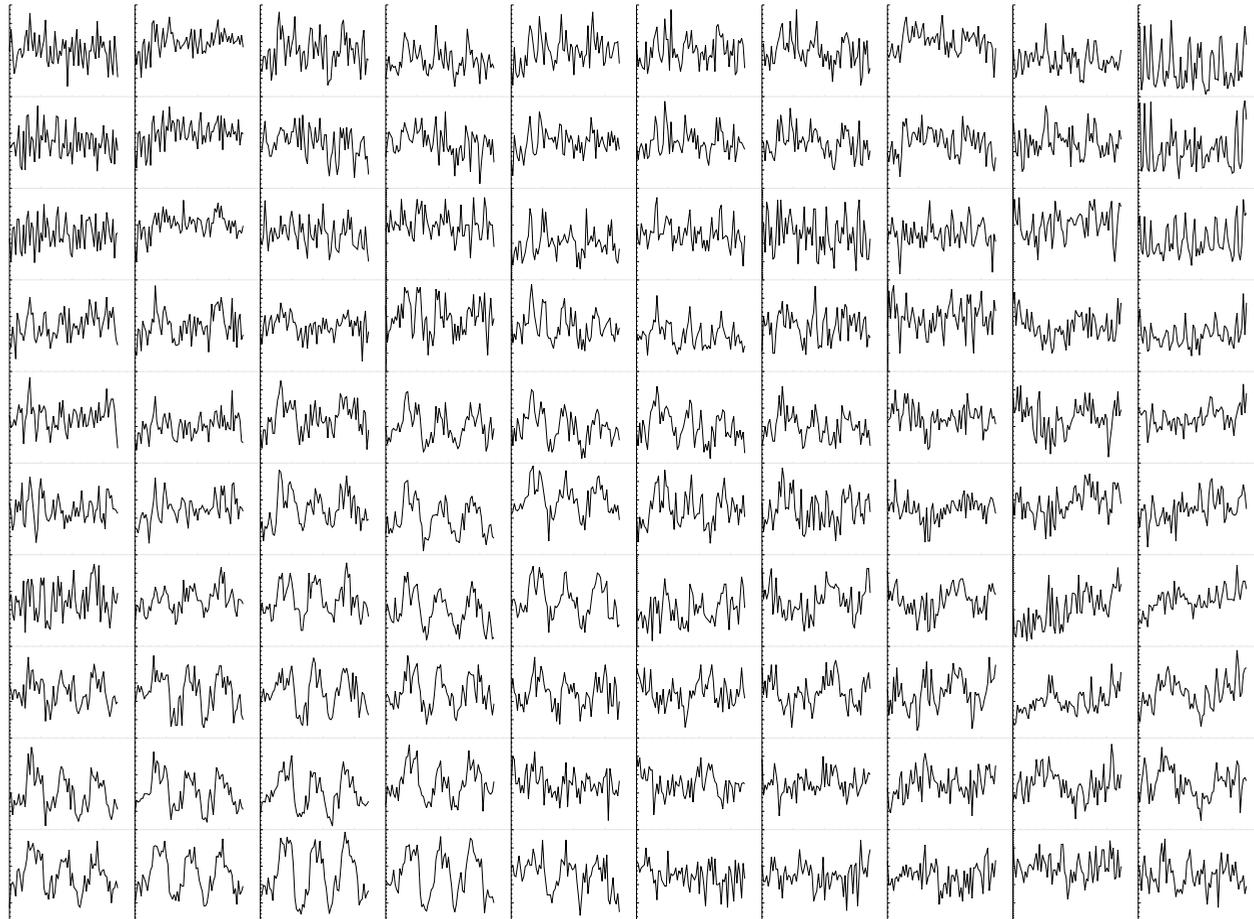
- Ziel: Identifikation von Regionen im Gehirn, die an der Erfüllung bestimmter Aufgaben beteiligt sind (z.B. das Sehzentrum).
- Experiment mit visuellem Stimulus:
 - Abwechselnd Phasen mit und ohne Stimulus.
 - Dauer einer Phase jeweils 30 Sekunden.
 - Die Gehirnaktivität wird alle drei Sekunden an $128 \times 128 \times 7$ Voxeln gemessen.

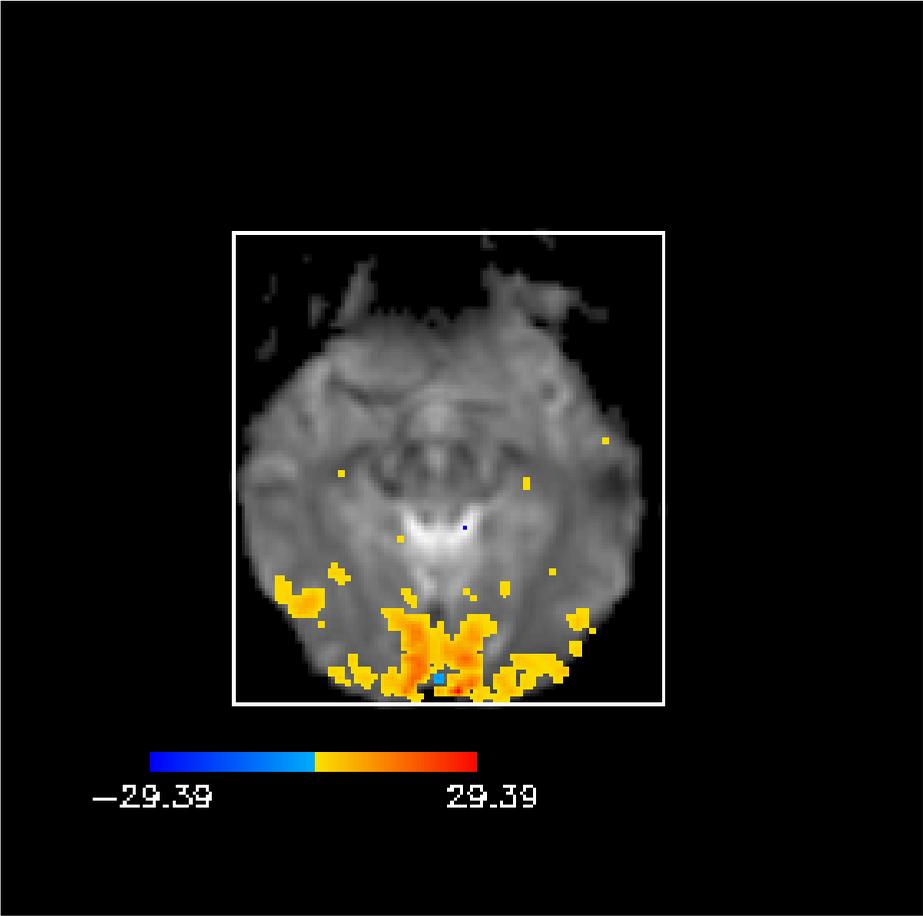


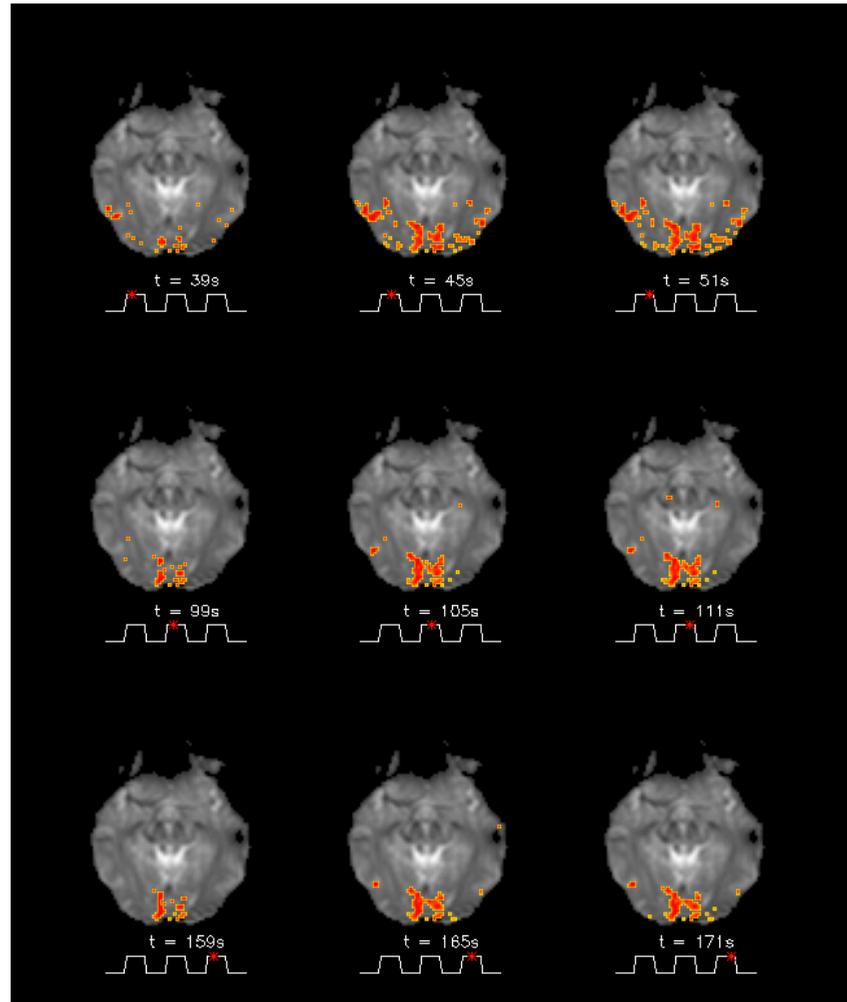
- Aktivierung wird durch funktionelle Magnetresonanztomografie (fMRT) gemessen.



- Die Messungen werden durch zufällige Fehler überlagert (Patient bewegt sich, ist unkonzentriert, Messungenauigkeit, . . .).
- Rolle der Statistik: Trennung von Signal und Rauschen.
- Aktivierung an einigen Voxeln:







1.2.2 Ein Definitionsversuch und einige weitere prizielle Bemerkungen

Statistik: (nach Brockhaus / dtv Lexikon (1992))

- „im materiellen Sinn“: Informationen in Form empirischer Zahlen [und Diagramme] („Statistiken“)
- „im instrumentellen Sinn“: „Statistische Methodenlehre“, Statistik als Wissenschaft; hier in erster Linie;

a) Die 7 Superlative: Statistik ist der Inbegriff

i) der Lüge

ii) des Beliebigen

iii) des Irrelevanten

iv) des Unsinnigen

v) des Reaktionären

vi) der Gleichmacherei

vii) des Langweiligen

b) Defensive Antwort

c) Offensive Antwort

- **Statistik ist die interdisziplinäre Wissenschaft von der datenbasierten Erkenntnisgewinnung.**
- „Statistik ist eine Wissenschaft, keine Instrumentenkunde (Menges)“
- Encyclopædia Britannica: „Statistics is the art and science of gathering, analyzing and making inferences from data. Originally associated with numbers gathered for governments, the subject now includes large bodies of method and theory.“
- Nochmals Erinnerung an: Cox & Snell (Applied Statistics, 1981, Chapman & Hall, p. 3):
„Statistical Analysis deals with those aspects of the analysis of data that are not highly specific to particular fields of study. That is, the object is to provide concepts and methods that will, with suitable modification, be applicable in many different fields of application; indeed one of the attractions of this subject is precisely this breadth of potential applications.“

Hauptgebiete der Statistik als Methodenlehre

- Deskriptiv / explorativ

Analyse der Daten der *konkret vorliegenden* Gesamtheit (*keine Verallgemeinerung beabsichtigt*)

Deskription = Beschreibung (durch Tabellen, Kennzahlen, . . .)

Informationsgewinn durch Verdichtung (Wald vor lauter Bäumen sehen!)

Aufspüren von Zusammenhängen, Hypothesen*generierung* (keine Prüfung!)

Data Mining („Graben in Datenschätzen“)

- Induktive Statistik = Statistische Inferenz

Schluss von einer Stichprobe auf die dahinterstehende Grundgesamtheit

Die Ergebnisse der Stichprobe sind nur Mittel zum Zweck des verallgemeinernden Schlusses

Solche Schlüsse vom Teil auf das Ganze sind zwangsläufig potentiell fehlerhaft. Man kann diesen Fehler („Inferenzfehler“) nicht ausschalten (Induktionsproblem,), *aber* unter Umständen kontrollieren.

„Trick“: Ziehe die Stichprobe zufällig (Wahrscheinlichkeitsauswahl), dann kann man die Wahrscheinlichkeit von groben Fehlschlüssen berechnen.

Beachte: Unter 'zufällig ist' hier - im Gegensatz zur Umgangssprache - nicht „willkürlich“ gemeint, sondern nach einem Auswahlverfahren mit einem kontrollierten Zufallsprozess (z. B. verdeckte Adresskarten aus gut durchmischter Einwohnerkartei ziehen), oder das unerwartete Eintreten eines Randphänomens.

Zur Abschätzung des Inferenzfehlers dient also die

- Wahrscheinlichkeitsrechnung als mathematische Theorie zur Beschreibung unsicherer / zufälliger Phänomene (\rightarrow Stochastik)

Unverzichtbare Voraussetzung für induktive Statistik

Teilweise auch für sich genommen interessant (z.B. Modelle sozialer Mobilität)

- Methodologie der Datengewinnung

Stichprobendesigns:

Wie gewinnt man geeignete Stichproben? (hier bestenfalls nur kurz; Veranstaltung für Nebenfachstudierende; Methodenvorlesung)

Konzipierung des Erhebungsinstruments: Wie erhebt man was?

(Operationalisierung komplexer Konstrukte (z.B. Integrationsfähigkeit), Gestaltung des Instruments (z.B. Techniken der Fragebogenerstellung))

siehe Methodenveranstaltungen, teilweise auch „Wirtschafts- und Sozialstatistik“ im Nebenfach

Datenproduktion

- * Amtliche Statistik (durch statistische Ämter (Europa, Bund, Länder, teilweise Städte, auch „ausgelöst“ (Bundesbank, Bundesagentur für Arbeit)), Unterstützung politischer und wirtschaftlicher Entscheidungen, auf gesetzlicher Grundlage (informationelle Selbstbestimmung, dafür aber oft Auskunftspflicht), meist rein deskriptive Analyse)
- * freiwillige Umfragen (wissenschaftlich, kommerziell, teilweise auch von Behörden/ Städten)

- Allgemeiner: Quellen der Unsicherheit beim statistischen Schließen

Kausalität vs. Zusammenhang.

Messfehler und Messungenaugigkeit.

Inferenzfehler bei Stichproben.

1.2.3 Überblick über die Veranstaltung

I. Beschreibende (und explorative) Statistik

- Grundbegriffe
- Beschreibung eindimensionaler Merkmale
- Beschreibung mehrdimensionaler Merkmale, d.h. Beschreibung von Zusammenhängen

II. Wahrscheinlichkeitsrechnung = Mathematische Modellierung und Analyse des Zufälligen / Unsicheren

III. Induktive Statistik = Anwendung der Wahrscheinlichkeitstheorie auf die Analyse von zufälligen (\approx „repräsentativen“) Stichproben

- z.B. Schätze den wahren Anteil der Rot/Grün-Wähler in der Grundgesamtheit möglichst genau (*Punktschätzung*)
- Gib ein Intervall an, das den wahren Wert mit hoher Wahrscheinlichkeit enthält (*Intervallschätzung*)
- Mit welcher Wahrscheinlichkeit begehe ich einen Fehler, wenn ich aus der Stichprobe schließe, dass Frauen systematisch weniger verdienen als Männer? (*Hypothesentests*)
- Ausblick auf komplexere statistische Verfahren

Im Rahmen der Vorlesung Statistik I beschäftigen wir uns ausschließlich mit deskriptiver (beschreibender) und explorativer Statistik. Wahrscheinlichkeitsrechnung und statistische Inferenz (Schließen von Stichproben auf die Grundgesamtheit) werden Thema der Vorlesung Statistik II im Sommersemester sein.

Gliederung von Statistik I

1. Einführung
2. Häufigkeitsverteilungen
3. Lage- und Streuungsmaße
4. Konzentrationsmaße
5. Analyse von Zusammenhängen
6. Regression

Anhang Kleiner Exkurs:

„Linke Einstellung“ ist keine Ausrede, die davor bewahrt, sich mit Statistik und empirischer Sozialforschung zu beschäftigen:

Nach längerer Diskussion „... scheint jetzt festzustehen, dass die Methodenlehre der empirischen Sozialforschung nicht ersetzt, sondern verfeinert und auf ihren eigentlichen Funktionsbereich beschränkt in der kritischen Sozialwissenschaft einen Platz behalten soll.

So unterscheiden sich die Entwürfe einer Sozialwissenschaft, die sich emanzipatorischen Interessen verpflichtet, von anderen durch den Nachdruck auf Ideologiekritik, durch eine veränderte - nämlich kritische - Begriffsbildung, durch die offene Berücksichtigung politischer Interessen bei der Problemdefinition und der Veröffentlichung der Resultate, nicht aber auf der Ebene von Forschungsverfahren und Forschungstechniken. Dass die Interaktion des Forschers mit dem Forschungsgegenstand selbst eine politische Dimension hat, sich als politische Praxis begreifen und planvoll betreiben lässt, ist bislang kaum ins Bewusstsein der kritischen Sozialwissenschaftler getreten.

Dies hängt unter anderem damit zusammen, dass sich nicht nur manche Gruppen der Studentenbewegung an einem stark vereinfachten Positivismusbegriff orientieren, dass manche kritische Sozialwissenschaftler

die Mängel ihrer eigenen Ausbildung - viele haben die Methodenlehre der empirischen Sozialforschung im Laufe ihres Studiums nur unzureichend kennengelernt - in Abwehrhaltungen gegen alles, was mit Forschungstechniken und Statistik zu tun hat, verwandelt haben. Diese weithin vorfindbare Korrelation - konformistische Sozialwissenschaftler verstehen mehr von der Methodenlehre der Sozialforschung und gehen unbefangener mit ihr um - hat die Anhänger der kritischen Theorie in den Nachteil gebracht, zwar die Untersuchungsansätze anderer Sozialwissenschaftler politisch kritisieren zu können, aber nur in seltenen Fällen die methodische Durchführung der kritisierten Untersuchungsansätze. So erscheint es nicht mehr angebracht, Untersuchungen, die sich der Methoden der empirischen Sozialforschung bedienen, in Bausch und Bogen als positivistisch zurückzuweisen. Weder die Korrelation von politischen Einstellungen und Vertrautheit mit der Methodenlehre der Sozialforschung noch die Konsequenz unkritischer Abwehr der Methodenlehre als positivistisch sind der Entwicklung einer kritischen Sozialwissenschaft förderlich.“

Fuchs, W. (1976): Empirische Sozialforschung als politische Aktion. In Ritsert, J. (Hg.): Zur Wissenschaftslogik einer kritischen Soziologie. Suhrkamp, Frankfurt/M., 147-174. (Zitat auf S 148f.)

1.3 Literatur

Vorbemerkung: Es gibt sehr viele gute Statistikbücher, die sich meist in den wesentlichen Aspekten inhaltlich auch stark überlappen. Hier, in alphabetischer Reihenfolge, eine – mündlich kommentierte – Auswahl von Büchern mit unterschiedlichen Schwerpunkten:

Bamberg, G. & Baur, F. & Krapp, M. (2012¹⁷): Statistik. R. Oldenbourg Verlag, München, Wien.

Burkschat, M., Cramer, E. & Kamps, U. (2012²): Beschreibende Statistik: Grundlegende Methoden der Datenanalyse (EMIL@A-stat) Springer Verlag, Berlin, Heidelberg, New York.

Fahrmeir, L. & Künstler, R. & Pigeot, I. & Tutz, G. (2011⁷): Statistik - Der Weg zur Datenanalyse. Springer Verlag, Berlin, Heidelberg, New York.

Caputo, A. & Fahrmeir, L. & Künstler, R. & Lang, S. & Pigeot, I. & Tutz, G. (2009⁵): Arbeitsbuch Statistik. Springer Verlag, Berlin, Heidelberg, New York.

Cramer, E. & Neslehova, J. (2015⁶): Vorkurs Mathematik: Arbeitsbuch zum Studienbeginn in Bachelorstudiengängen. Springer Verlag, Berlin.

Galata, R. & Scheid, S. (2012): Deskriptive und Induktive Statistik für Studierende der BWL: Methoden - Beispiele - Anwendungen. Carl Hanser, München.

Genschel, U. & Becker, C. (2005): Schließende Statistik: Grundlegende Methoden (EMIL@A-stat) Springer Verlag, Berlin, Heidelberg, New York.

Hellbrück, R. (2016³): Angewandte Statistik mit R. Eine Einführung für Ökonomen und Sozialwissenschaftler. Gabler, Wiesbaden.

Hatzinger, R. & Hornik K. & Nagel H. & Maier M.J. (2014²): *R: Einführung durch angewandte Statistik.* Pearson Studium, München.

Jann, B. (2005¹²): **Einführung in die Statistik.** R. Oldenburg Verlag, München, Wien.

Litz (2003³): Statistische Methoden in der Wirtschafts- und Sozialwissenschaften. R. Oldenburg Verlag, München, Wien.

**Quatember (2011³): Statistik ohne Angst vor Formeln: Das Studienbuch fÃ¼r
Wirtschafts- und Sozialwissenschaftler.** Pearson, München.

Rohwer, G. & Pötter, U. (2001): Grundzüge der sozialwissenschaftlichen Statistik. Juventa (Grundlagentexte Soziologie). Weinheim, München.

Rohwer, G. & Pötter, U. (2002): Wahrscheinlichkeit. Begriff und Rhetorik in der Sozialforschung. Juventa (Grundlagentexte Soziologie). Weinheim, München.

Schira, J. (2016⁵): Statistische Methoden der VWL und BWL: Theorie und Praxis. Pearson Studium, München.

Toutenburg, H. & Heumann, C. (2009⁷): Deskriptive Statistik. Springer Verlag, Berlin, Heidelberg, New York.

Toutenburg, H. (2008): Induktive Statistik. Springer Verlag, Berlin, Heidelberg, New York.

Toutenburg, H. & Schomaker, M. & Wißmann, M. & Heumann, C. (2009): Arbeitsbuch zur deskriptiven und induktiven Statistik. Springer Verlag, Berlin, Heidelberg, New York.

Wagschal, U. (1999): Statistik für Politikwissenschaftler. R. Oldenburg Verlag, München, Wien.

1.4 Grundbegriffe

1.4.1 Statistische Einheiten und Gesamtheiten

Statistische Einheiten: Objekte, an denen interessierende Größen erhoben werden.

Grundgesamtheit: Die Menge aller für eine bestimmte Fragestellung relevanten statistischen Einheiten heißt *Grundgesamtheit* (Universum, Population).

Stichprobe: Häufig ist es nicht sinnvoll, nicht möglich, oder zu teuer, alle Elemente der Grundgesamtheit zu untersuchen („*Vollerhebung*“). Stattdessen wird nur eine bestimmte Auswahl untersucht, also eine *Stichprobe* der Grundgesamtheit.

Abgrenzungsproblematik: Die Grundgesamtheit muss durch sachliche, räumliche und zeitliche Kriterien exakt festgelegt sein. Die Kriterien richten sich nach dem Untersuchungsziel.

Gesamtheit: Im Rahmen der deskriptiven Statistik wird keine Verallgemeinerung der aus einer Stichprobe gewonnenen Ergebnisse auf die ihr zugrundeliegende Grundgesamtheit angestrebt. Es ist also in den folgenden Kapiteln nicht nötig zu unterscheiden, ob die zu analysierenden Daten aus einer Stichprobe stammen oder bereits die Grundgesamtheit darstellen. Wir sprechen dann einfach von einer *Gesamtheit* von statistischen Einheiten, die analysiert werden sollen.

Notation:

- In einer Gesamtheit mit n (*Stichprobenumfang, bzw. Umfang der Gesamtheit*) Elementen werden die statistischen Einheiten mit $\omega_1, \omega_2, \dots, \omega_n$ bezeichnet.
- Gesamtheit $\Omega = \{\omega_1, \dots, \omega_n\}$.
- Bezieht man sich auf ein festes, aber beliebiges Element der Grundgesamtheit, so schreibt man meist ω (ohne Index).

1.4.2 Merkmale und Merkmalsausprägungen

Merkmale: Inhaltlich interessant sind nicht die Einheiten an sich, sondern bestimmte Eigenschaften oder *Merkmale* der Einheiten (Variablen).

Merkmalsausprägungen: Ausprägung eines Merkmals für eine konkret vorliegende statistische Einheit.

Wertebereich: Alle prinzipiell möglichen Ausprägungen eines Merkmals.

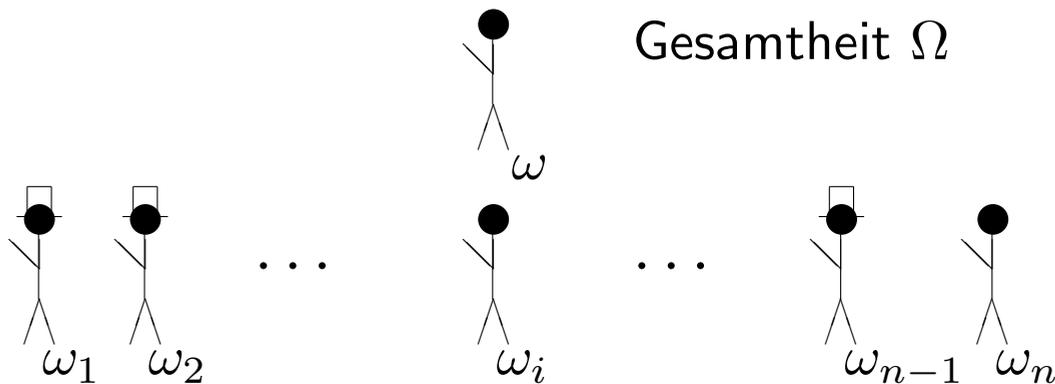
Notation: Merkmale werden typischerweise mit Großbuchstaben bezeichnet (X, Y, Z , etc.), Ausprägungen mit dem zugehörigen Kleinbuchstaben (x, y, z). Der Wertebereich wird mit \mathbb{W} bezeichnet. Manchmal werden auch bei den Merkmalen Indizes verwendet, z.B. X_1 . Die Ausprägungen werden dann konsequenterweise mit x_1 bezeichnet

Formal ist jedes Merkmal eine Funktion.

Schreibweisen:

- $X(\omega)$ Merkmalsausprägung der Einheit $\omega \in \Omega$, $X_1(\omega)$
- $X(\omega) = x$, $X_1(\omega) = x_1$
- $X(\omega_i) = x_i$, $X_1(\omega_i) = x_{1i}$ (als $(x_1)_i$ denken)

Die Elemente von \mathbb{W} werden wir später mit a_1, \dots, a_k bezeichnen.



Bsp. 1.1.

Verknüpfungen von Merkmalen/ „Rechnen mit Merkmalen“

Man kann mit Merkmalen rechnen. Insbesondere gilt: Seien X und Y Merkmale mit Wertebereich $\mathbb{W} \subseteq \mathbb{R}$ und c und d reelle Zahlen. Dann ist $Z = c \cdot X + d \cdot Y$ wieder ein Merkmal mit $Z(\omega) = c \cdot X(\omega) + d \cdot Y(\omega)$

Teilmengen der Gesamtheit: Oft wird auch die Menge aller Einheiten benötigt, bei denen das Merkmal X einen bestimmten Wert, üblicherweise mit x bezeichnet, annimmt.

$$\{ \omega \in \Omega \mid X(\omega) = x \}$$

Bsp. 1.2. *Mietspiegel*

- Grundgesamtheit
- Statistische Einheiten
- Merkmale und Wertebereich
- Merkmalsausprägungen für die 713-te Wohnung im Datensatz
- Defacto wird nur eine Stichprobe ausgewertet (ca. 3000 Wohnungen), in Statistik I wird aber, wie gesagt, nicht zwischen Stichprobe und Grundgesamtheit unterschieden
- Teilmengen (mit Merkmalen X als „Nettomiete pro qm“, Y als „Wohnfläche“):

1.4.3 Merkmalstypen

Eine adäquate statistische Analyse hängt entscheidend davon ab, welche inhaltliche Struktur \mathbb{W} (also die Menge der möglichen Merkmalsausprägungen) hat, sie spielt eine gröse Rolle für die Formulierung von „Klassen von Problemlösungen“ (vgl. ??)

Drei mögliche Unterscheidungen von Merkmalen:

a) Stetige, quasi-stetige und diskrete Merkmale

- Diskret: Das Merkmal kann nur endlich viele (oder abzählbar viele) Ausprägungen annehmen.
- Stetig: Das Merkmal kann (im Prinzip) alle Werte in einem Intervall annehmen (überabzählbar viele verschiedene Ausprägungen).
- Quasi-stetig: Zwischenform. Jede Messung hat nur endliche Genauigkeit, ist also eigentlich diskret, kann aber als stetig behandelt werden.
- Kategorisierung: Oft werden stetige Daten auch absichtlich diskretisiert, nämlich bei Gruppenbildung (gruppieren, klassieren, kategorisieren).

b) Skalenniveau

Diese formalen Überlegungen entsprechen einer großen praktischen Schwierigkeit: In statistischer Software werden Merkmalsausprägungen in der Regel mit Zahlen codiert, die Software erkennt aber oft nicht notwendig, welche Bedeutung diese Zahlen haben. Ob die Variable „Augenfarbe“ mit den Ausprägungen 0=blau, 1=grün, 2=braun oder die Variable „Fehlversuche“ der an der Klausur teilnehmenden Hörer(innen), wiederum mit den Ausprägungen 0, 1, 2 betrachtet wird, macht für das Software-Paket unter Umständen keinen Unterschied. Es gibt stets einen Durchschnittswert aus, der aber im ersten Fall absolut sinnlos ist.

Man teilt die Merkmale nach Eigenschaften der Struktur von \mathbb{W} ein. Man spricht dann von verschiedenen Skalenniveaus. Das Skalenniveau eines Merkmals bestimmt, welche statistischen Verfahren inhaltlich sinnvoll angewendet werden können.

- *Nominalskala*: Ein Merkmal heißt *nominalskaliert*, wenn die möglichen Ausprägungen Namen oder Kategorien sind, die keine natürliche Ordnung haben.

Häufig werden auch hier (etwa zur Datenanalyse am PC) den Ausprägungen Zahlen zugeordnet. Diese Zahlen sind aber nur Stellvertreter ohne inhaltliche Bedeutung; ihre Zuordnung kann völlig willkürlich erfolgen (solange eindeutig).

- *Ordinalskala*: Ein Merkmal heißt *ordinalskaliert*, wenn sich die möglichen Ausprägungen ordnen lassen.

Man kann beliebige Zahlen zuordnen, solange diese die Ordnung erhalten.

Die Abstände der Merkmalsausprägungen lassen sich nicht sinnvoll interpretieren, insbesondere kann strenggenommen kein Durchschnittswert als Mittelwert gebildet werden. (Durchschnittsnoten sind hier allerdings ein Graubereich)

Beachte, entscheidend sind inhaltliche Aspekte. Geht es z. B. um persönliche Vorlieben, dann kann es bei der Augenfarbe durchaus eine Ordnung geben.

- *Intervallskala*: Ein Merkmal heißt *intervallskaliert*, wenn die Abstände der möglichen Merkmalsausprägungen sinnvoll interpretiert werden können.
- *Verhältnisskala / Ratioskala*: Ein Merkmal heißt *verhältnisskaliert*, wenn es intervallskaliert ist und zusätzlich ein sinnvoll interpretierbarer Nullpunkt existiert.
- Verhältnisskala und Intervallskala werden oft zur *Kardinalskala* zusammengefasst. Ein kardinalskaliertes Merkmal wird auch als *metrisch* bezeichnet.
- Beachte: Metrische Merkmale sind oft stetig oder quasi-stetig (z.B. Größe, Einkommen), können aber auch diskret sein (z.B. Anzahlen).

Zusammenfassende Darstellung:

Hat man den Ausprägungen Zahlen zugeordnet, so sind je nach Skalenniveau folgende Berechnungen sinnvoll:

Skala	Häufigkeiten auszählen	Größen vergleichen	Differenz bilden	Quotienten bilden
Nominalskala				
Ordinalskala				
Intervallskala				
Verhältnisskala				

Man sieht eine deutliche Hierarchie: Alles was auf einer Nominalskala erlaubt ist, ist auch auf der Ordinalskala erlaubt usw., aber nicht umgekehrt! Das bedeutet: Man darf Verfahren der niedrigeren Stufe auch auf der höheren Stufe anwenden (etwa ein Verfahren der Ordinalskala auch auf Daten einer Intervallskala), aber nicht umgekehrt.

Insbesondere ist zu beachten: Mittelwertbildung setzt Interpretierbarkeit von Summen (Differenzen) voraus, also eine Kardinalskala.

Zulässige Transformationen: Mathematisch etwas exakter charakterisiert man Skalen über die Transformationen, die man durchführen darf, ohne die inhaltliche Struktur zu zerstören, d.h. vor und nach der Transformation sollen die für die jeweilige Skala grundlegenden Operationen jeweils dieselben inhaltliche Ergebnisse liefern.

Transformation („Umrechnung“):

	Transformation
Nominalskala	eindeutige
Ordinalskala	streng monotone
Intervallskala	lineare ($a + bX; b > 0$)
Verhältnisskala	linear affine ($bX; b > 0$)

Damit bleiben auf Intervallskala Verhältnisse von Differenzen gleich, und auf der Verhältnisskala Verhältnisse.

$$Y = a + bX$$

$$\begin{aligned}\frac{Y_1 - Y_2}{Y_3 - Y_4} &= \frac{a + bX_1 - (a + bX_2)}{a + bX_3 - (a + bX_4)} = \\ &= \frac{b \cdot (X_1 - X_2)}{b \cdot (X_3 - X_4)} = \\ &= \frac{X_1 - X_2}{X_3 - X_4}\end{aligned}$$

Abschließende weiterführende Bemerkungen:

- Grauzone: Manchmal werden bei Ordinalskala durchaus Differenzen und Mittelwerte gebildet (z.B. Durchschnittsnoten, Skalenindizes). Aus statistischer Sicht oft problematisch \Rightarrow Vorsicht bei der Interpretation.
- Hat man bei einer Nominalskala nur zwei Ausprägungen und kodiert diese mit '0' und '1' und berechnet formal den Mittelwert, so ist dieser inhaltlich nicht absolut sinnlos, sondern gibt den Prozentsatz der Ausprägung '1' an.
- Nochmals: Die Einteilung, welches Skalenniveau eine Variable besitzt, ist letztlich oft auch von der inhaltlichen Fragestellung abhängig. Sie bestimmt (natürlich!), welche statistischen Verfahren sinnvoll sind. So können je nach inhaltlicher (!!) Fragestellung auch meist als nominal eingestufte Merkmale manchmal als ordinal betrachtet werden (z.B. Rechts-Links-Skala bei Parteienpräferenz)

c) Qualitative und quantitative Merkmale

- **Qualitativ:** Das Merkmal beschreibt eine Eigenschaft / eine Qualität und kein Ausmaß. Das Merkmal besitzt nur endlich viele Ausprägungen und ist nominal- oder ordinalskaliert.
- **Quantitativ:** Das Merkmal gibt messbar ein Ausmaß wieder. Das Merkmal ist sinnvoll in Zahlen messbar und intervall- oder verhältnisskaliert.

Vorsicht: Die Unterscheidung in qualitative und quantitative Merkmale ist zu unterscheiden vom Begriff qualitative / quantitative Sozialforschung. Auch bei qualitativen Merkmalen wird im soziologischen Sinn quantitativ gearbeitet (Häufigkeitsverteilung etc. statt Einzelfallbeschreibung).

1.4.4 Erhebungsformen

Experiment vs. Beobachtungsdaten:

- Experiment: Die Daten werden gezielt erzeugt. Insbesondere können die interessierenden Größen direkt beeinflusst werden.
- Beobachtungsdaten: Die Daten sind prinzipiell bereits vorhanden und müssen nur noch „erhoben“ werden.

Vollerhebung vs. Stichprobe vgl. oben:

- Vollerhebung: Alle statistischen Einheiten der Grundgesamtheit werden untersucht.
- Stichprobe: Nur ein Teil der Gesamtheit wird untersucht. Dieser soll möglichst repräsentativ für die Grundgesamtheit sein.
- Gründe für Stichproben:
 - + Geringerer Aufwand.
 - + Vollerhebung nicht möglich (z.B. in der Qualitätskontrolle, zerstörende Prüfung).

Auswahltechniken:

- Einfache Zufallsstichprobe,
- Klumpenstichprobe,
- Geschichtete Stichprobe.

Studientypen:

- Querschnittsstudie: An einer Menge von Einheiten werden zu einem Zeitpunkt mehrere Merkmale erhoben.
- Zeitreihe: Ein Merkmal wird wiederholt zu verschiedenen Zeitpunkten erhoben.
- Longitudinal- / Paneldaten: An einer festen Menge von statistischen Einheiten werden wiederholt (die gleichen) Variablen erhoben.

Analysearten:

- Primärerhebung / -analyse:
Daten werden im Rahmen des Forschungsprojekts erhoben und analysiert.
- Sekundäranalyse:
- Tertiäranalyse:
Analyse von aggregierten (zusammengefassten) Daten.
- Metaanalyse:
Sekundäranalyse oder Tertiäranalyse (= Metaanalyse im engeren Sinn) von mehreren Studien.