Leuphana Universität Lüneburg

Statistik für alle

Fakultät W
Wirtschaftswissenschaften
Professur 'Statistik und Freie Berufe'
Univ.-Prof. Dr. Joachim Merz

Skriptum zur Vorlesung und Übung

Fünfte Auflage 2014
Impressum: Statistik für alle – Skriptum zur Vorlesung und Übung,
herausgegeben von der Leuphana Universität Lüneburg,
Fakultät W - Wirtschaftswissenschaften.
Univ.-Prof. Dr. Joachim Merz, Forschungsinstitut Freie Berufe,
Professur 'Statistik und Freie Berufe'.
Campus, Scharnhorststraße1, Gebäude 4, 21335 Lüneburg
E-Mail: ffb@leuphana.de
www.leuphana.de/ffb

Gedruckt auf 100 % Altpapier, chlorfrei gebleicht.

„A basic literacy in statistics will one day be as necessary for efficient citizenship as the ability to read or write“
H.G. Wells.

Vorwort

Statistik - Ziele und Hintergrund

Grundlegende Statistikkenntnisse sind für eine Zivilgesellschaft Voraussetzung und ganz unabhängig von einzelnen Studiengängen notwendiger Bestandteil eines Studiums überhaupt.

Vor diesem Hintergrund und dem allgemeinen Ziel, in einer Welt rapide zunehmender Informationsmengen Konzepte und Werkzeuge zur Informationskomprimierung für die Gewinnung zentraler Aussagen und Tendenzen bereit zu stellen, werden Grundlagen der Statistik mit Schwerpunkt auf die zusammenfassende Beschreibung in der Leuphana-Veranstaltung *Statistik* gelegt.

Statistische Informationen (Graphen, Kennzahlen etc.) sind wesentliche Bausteine zur Untermauerung von Argumenten, sei es im beruflichen, politischen, aber auch privaten Bereich. Zeitungen und andere gesellschaftlich wichtige Medien bleiben ohne das Verständnis für Zahlen und Graphen der deskriptiven Statistik unverständlich. Vorkenntnisse der deskriptiven Statistik (Mittelwerte, Streuung etc.) sind zudem eine notwendige Voraussetzung für darauf aufbauende Themenbereiche in allen Studiengängen.

Statistik - Didaktisches Konzept

Statistik wird mit ineinander verzahnten Ansätzen angeboten: Mit der **Vorlesung** werden die Studierenden mit statistischen Konzepten und praktischen Werkzeugen auf anschauliche und verständliche Weise in die Grundlagen der Statistik eingeführt. Die zu Beginn der Vorlesung **jeweils erhobene Umfrage zu den Lebens- und Wohnbedingungen der Studierenden** ist die empirische nicht an ein Fach gebundene Basis zur praktischen Umsetzung der Konzepte an eigenen Daten. Eingebunden in eine **umfassende Motivation für alle Studiengänge**, dienen zudem **Skriptum, themenspezifische Folien (ppts)** und **Übungsaufgaben** als fachlicher Hintergrund. Die **Seminare in kleinen Gruppen** dienen der Diskussion, Vertiefung und Anwendung der gewonnenen Erkenntnisse. Dort wird mit den eigenen Umfragedaten und Open Office (plattformunabhängige freeware) die praktische Umsetzung geübt.

Das vorliegende Skriptum soll vorlesungsbegleitend helfen, den Blick auf das Wesentliche, auf das Verständnis der Methoden und ihrer Anwendungen zu erleichtern. Ich empfehle, den Stoff mit der angegebenen Literatur zu vertiefen: Manchmal hilft ein anderer Blickwinkel, die Dinge besser zu begreifen. Das Verstehen, das verständige Umgehen mit Statistik als ein wesentlicher Baustein, Theorie mit Empirie zu verbinden, ist mir ein wichtiges Anliegen.

Statistik im Leuphana Semester - Überblick

1 2	Deskriptive Statistik	Allgemeine Grundlagen	Motivation, Einführende Beispiele und Bedeutung der Statistik für alle Lebensbereiche Statistische Einheiten, Merkmale und Umfragedesign
3 4		Konzeption und Ansätze zur Informationskomprimierung - Statistische Analyse eines einzelnen Merkmals	Grafische und tabellarische Zusammenfassung Komprimierung mithilfe von zentralen Indikatoren - Lageparameter/Mittelwerte - Streuung um zentrale Werte - Konzentrationsanalyse
5 6		Entdeckung von Mustern - Statistische Analyse mehrerer Merkmale	Kreuztabellen/Zweidimensionale Häufigkeitsdarstellung Analyse des Zusammenhangs zwischen Merkmalen/Korrelationsanalyse
7	Ausblick	Ausblick Schließende Statistik	Von der Stichprobe zur Grundgesamtheit, Wahrscheinlichkeit, Hypothesentests Mustererkennung, Data Mining und Multivariate Verfahren Statistik und Computing (SPSS, SAS, Stata , R…)

Für den problemorientierten Einstieg und den Umgang mit dem Computer als Hilfsmittel werden Tabellenkalkulatoren (wie z.B. Excel, Open Office), SPSS (Statistical Package for the Social Sciences) und andere Programmpakete verwendet.

Nicht zu vergessen: Studium und späterer Beruf sollen auch Spaß machen. Die Cartoons im Skriptum sind entsprechende Lockerungsübungen.

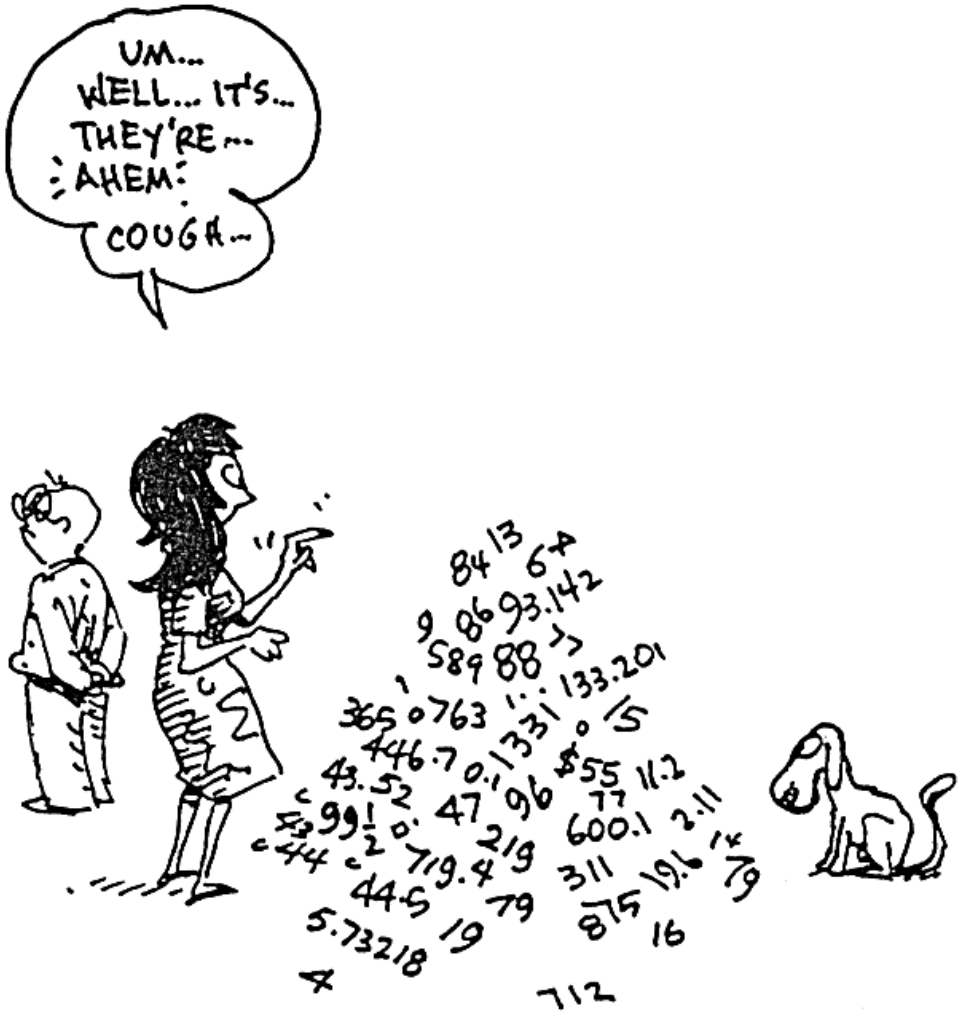
Viel Spaß und Erfolg!

Lüneburg, im August 2014 Univ.-Prof. Dr. Joachim Merz

Univ.-Prof. Dr. Joachim Merz

STATISTIK FÜR ALLE

THEMENBEREICHE



Statistics is …

… to compress information

… to quantify uncertainty

Gonick, Smith (1993)

Univ.-Prof. Dr. Joachim Merz

STATISTIK FÜR ALLE

GLIEDERUNG

Gonick, Smith (1993)

Statistik – Warum ist sie so wichtig?! Deskriptive Statistik, Wahrscheinlichkeitsrechnung und induktive Statistik

Moderne Statistik ist **Informationskomprimierung**. Dazu zählen in erster Linie Ansätze, mit denen eine Vielzahl von Informationen auf zentrale **Indikatoren** und **Kennzahlen** verdichtet werden können (**deskriptive Statistik**). Sind aus Kosten- und anderen Vereinfachungsgründen Erkenntnisse **aus Stichproben** für eine **übergeordnete Grundgesamtheit** zu gewinnen, dann ist es notwendig, etwas über die Signifikanz der Stichprobenergebnisse auszusagen (Wahrscheinlichkeitsrechnung, **schließende Statistik**).

Statistische Informationen (Graphen, Kennzahlen etc.) sind wesentliche Bausteine zur **Untermauerung von Argumenten**, sei es im **beruflichen, politischen** aber auch **privaten Bereich**. Zeitungen und andere gesellschaftlich wichtige Medien bleiben ohne das Verständnis für Zahlen und Graphen der deskriptiven Statistik unverständlich. Vorkenntnisse der deskriptiven Statistik (Mittelwerte, Streuung etc.) sind zudem eine notwendige Voraussetzung für eine darauf aufbauende schließende Statistik (Hypothesentest etc.)

Die ehemalige Bundesministerin für Bildung und Forschung, Edelgard Bulmahn hat in ihrem Vorwort des Gutachtens der Kommission zur Verbesserung der informationellen Infrastruktur zwischen Wissenschaft und Statistik (2001) die Wichtigkeit von statistischen Informationen betont:

„Gute politische Entscheidungen brauchen als Grundlage aussagekräftige statistische Informationen zur Situation und Entwicklung von Wirtschaft und Gesellschaft. Nur dann können Sozial- und Wirtschaftswissenschaften treffsichere Analysen erstellen und damit die Handlungsgrundlage für die Politik verbessern.“

Mit entsprechenden Erhebungen mit einer Fülle von Einzeldaten (Mikrodaten) werden die Informationen für zielkonforme Analysen gewonnen. Das Ziel ist es, aus der Vielzahl der Daten dann wesentliche Informationen wie Trend und durchschlagende Phänomene zu gewinnen. Vor einer multivariaten Analyse, einer Analyse mit konkurrierenden Erklärungsfaktoren steht die zusammenfassende und komprimierende Beschreibung der Situation aus der Gesamtheit der Daten: die Deskription, oder beschreibende Statistik, die im Vordergrund dieses Skriptums steht.

Statistik I - Deskription:

Beschreibende Statistik mit Verfahren zur Aufbereitung statistischer Daten bezogen auf die beobachteten Werte (Informationsaufbereitung und -verdichtung). Umfasst die Darstellung eines Datenmaterials in Form von Kennzahlen, Tabellen und Grafen.

Statistik II - Wahrscheinlichkeitsrechnung und induktive Statistik:

Dient der Überprüfung allgemeingültiger Theorien. Informationsbewertung durch Inferenzstatistik (schließende Statistik): Wahrscheinlichkeitsaussagen über die Vereinbarkeit der in den Daten erfassten Realität (Empirie) mit den aus einer Theorie abgeleiteten Hypothesen.

Die Wahrscheinlichkeitsrechnung ist notwendig, um von Teilerhebungen (Stichproben, 'sample') auf eine Grundgesamtheit zu schließen (induktive Statistik).

Zum Aufbau von Statistik für alle

Einführung und allgemeine Grundlagen
- Beispiele, Begriff und Aufgaben
- Träger der Wirtschaftsstatistik und statistische Quellen
- Adäquationsproblem und sachgerechte Interpretation
- Statistische Einheiten, Massen, Merkmale und Meßskalen

Statistische Analyse eines einzelnen Merkmals
- Eindimensionale Häufigkeitsverteilungen
- Lageparameter
- Streuungsmaße
- Konzentration und Verteilung

Statistische Analyse mehrerer Merkmale
- Zweidimensionale Häufigkeitsverteilungen
- Korrelationsrechnung

I Einführung und allgemeine Grundlagen

Beispiele, Begriffe, Aufgaben und Quellen sowie statistische Einheiten, Massen und Skalen als Grundlage für die deskriptive Statistik

1 Einführende Beispiele

1.1 Anwendungsorientierte Statistik: Mikroanalyse der individuellen Wirkungen der Steuerreform 1990 - Mikroökonomische Theorie, Mikrodatenbasis, Mikroökonometrie und Mikrosimulation

Mikroökonomisches Modell Multipler Markt- und Nichtmarktmäßiger Aktivitäten Privater Haushalte

- Steuern und Transfers
- Sozioökonomische Charakteristika
- Mikroökonomisches Modell optimaler Zeitallokation

Mikrodaten und Merge

Berechnung individueller Steuervariablen

- Steuerschuld
- Grenzsteuersätze

- Sozio-ökonomisches Panel 1. Welle 1984
- ESt-/LSt-Statistik 1983
- Steuerrecht 1983/1990

Erweiterte Mikrodatenbasis

- Sfb 3 - Nebenerwerbstätigkeits-umfrage 1984 (BfLR)

Merge

- Steuervariablen
- Regionale Wirtschafts- und Arbeitsmarktdaten

Mikroökonometrisches Modell und Schätzung

- 3 stufiges selektionskorrigiertes Modell multiplen Arbeits(Aktivitäts)angebots
- Eigenarbeit, Nebenerwerb/Schwarzarbeit und Haupterwerb
  - Partizipation
  - Löhne/Einkommen
  - Zeitallokation

Mikrosimulation der Steuerreform 1990 für die Jahre 1990 und 2000

Dynamische Mikrosimulation demografischer Entwicklungen

Hochrechnung der Mikrodaten nach dem Prinzip des minimalen Informationsverlustes (MIL)

- Simultane Hochrechnung mit der demografischen Situation 1990
- Simultane Hochrechnung mit der demografischen Situation 2000

Mikrosimulation der Steuerreform 1990 für 1990 und 2000

- Mikrosimulation mit dem Statischen Sfb3-Mikrosimulationsmodell MICSIM
- Zeitallokationseffekte auf individuelle multiple markt- und nichtmarktmäßige Aktivitäten

Abb. I.1: Mikrosimulation der Steuerreform 1990: Struktur des Analysesystems
Quelle: Merz, J. (1991a)

1.2 Statistik und EDV: Deskription und Inferenz am Beispiel des Programmpakets SPSS (Statistical Package for the Social Sciences)

SPSS – Funktionen:

- Datenhandling, Datenbearbeitung
- Analysemodule
- Grafiken
- Utilities

www.spss.com

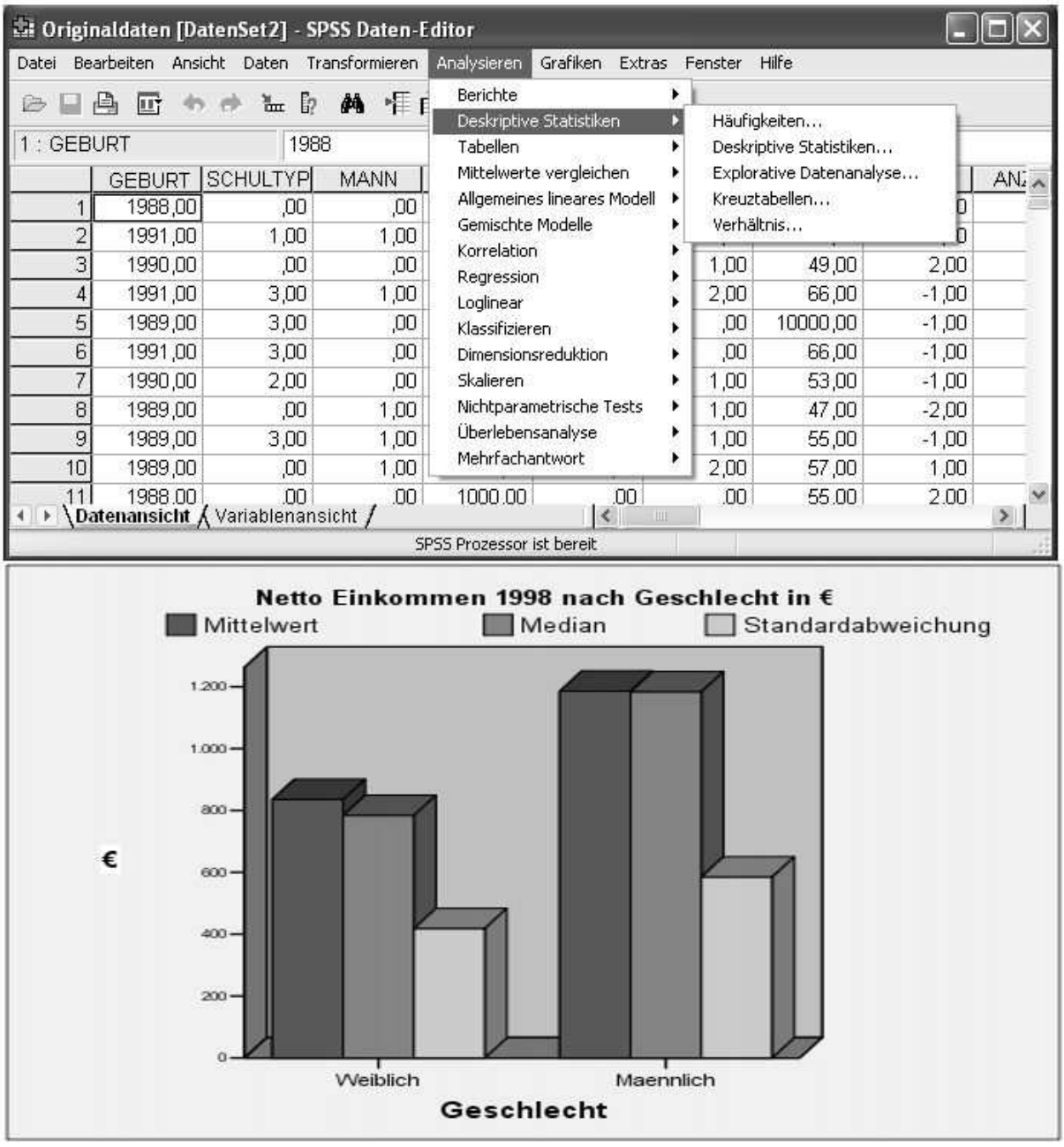


Abb. I.2 SPSS, Screenshot Datenansicht und Beispiel Grafikoutput

1.3 Problemorientierte Statistik: Studien zur aktuellen und zukünftigen Situation der Erde – Club of Rome und Intergovernmental Panel on Climate Change (IPCC)

Club of Rome: Grenzen des Wachstums mit ersten Studien aus den 70er Jahren
http://www.clubofrome.de/

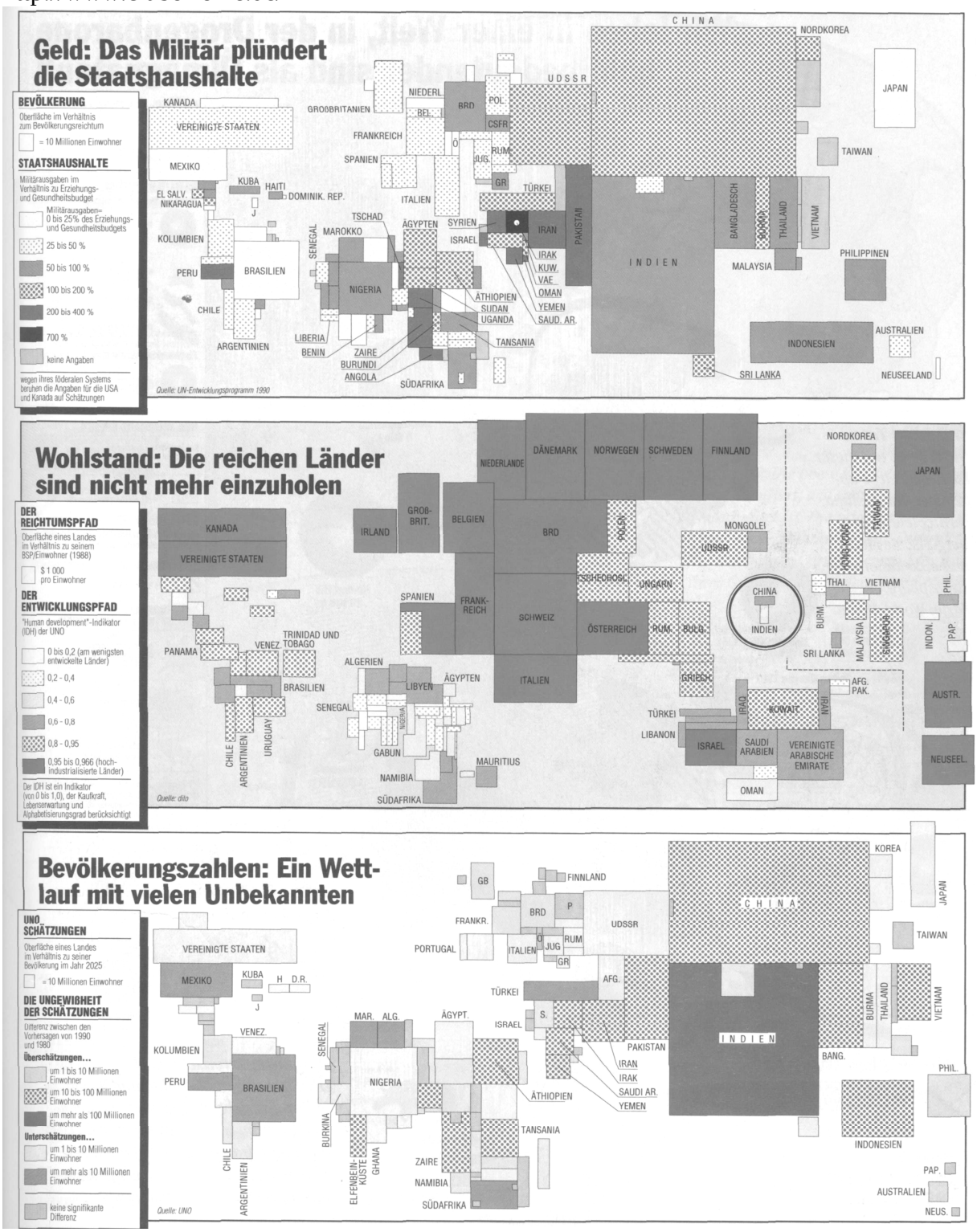


Abb. I.3: Blick in die Zukunft: Militär, Wohlstand, Bevölkerung
Quelle: Club of Rome (1991), S. 31

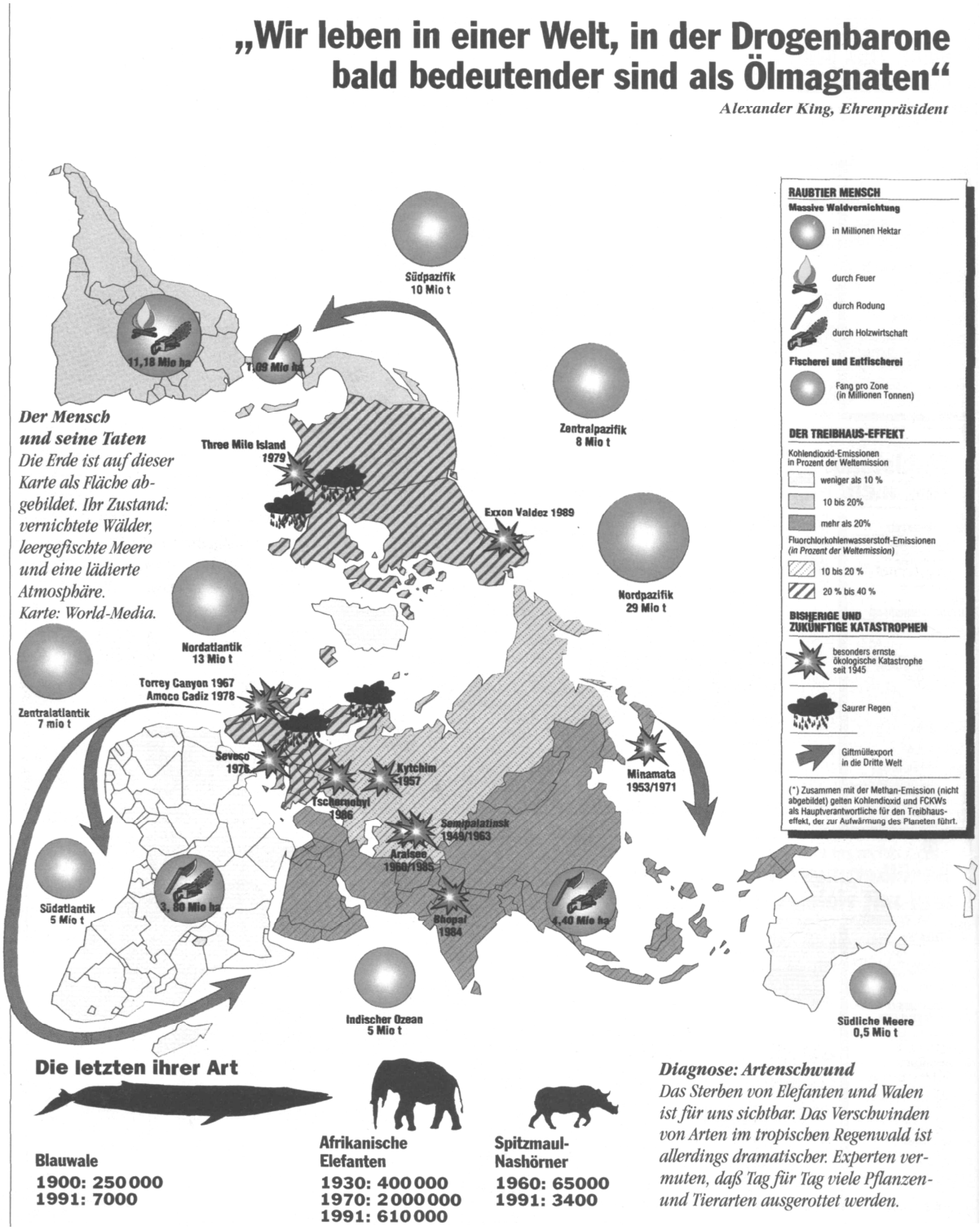


Abb. I.4: Blick in die Zukunft: Der Mensch und seine Taten
Quelle: Club of Rome (1991), S. 32

Tab. I.1: Weltweites Wachstum in ausgewählten Sektoren

	1970			1990	
Weltbevölkerung	3,6	Mrd		5,3	Mrd
Kraftfahrzeuge	250,0	Mio		560,0	Mio
gefahrene Kilometer/Jahr (nur OECD-Länder)					
PKW	2.584,0	Mrd		4.489,0	Mrd
LKW	666,0	Mrd		1.536,0	Mrd
Ölverbrauch/Jahr	17,0	Mrd	Barrel	24,0	Mrd
Kohleverbrauch/Jahr	2,3	Mrd	Tonnen	5,2	Mrd
Kapazität E-Werke	1,1	Mrd	Kilowatt	2,6	Mrd
Strom aus Kernkraft/Jahr	79,0	Mrd	Terawatt-Std.	1.884,0	Mrd
Getränkeverbrauch/Jahr					
nicht alkoholisch/Jahr	23,0	Mrd	Liter	58,0	Mrd
Bierverbrauch/Jahr	19,0	Mrd	Liter	29,0	Mrd
Aluminium für Getränkebehälter	72.700,0	Mrd	Tonnen	1.251.900,0	Mrd
Müll aus Gemeinden/Jahr (nur OECD-Länder)	302,0	Mio	Tonnen	420,0	Mio

Quelle: Meadows et al. (1992), S. 27

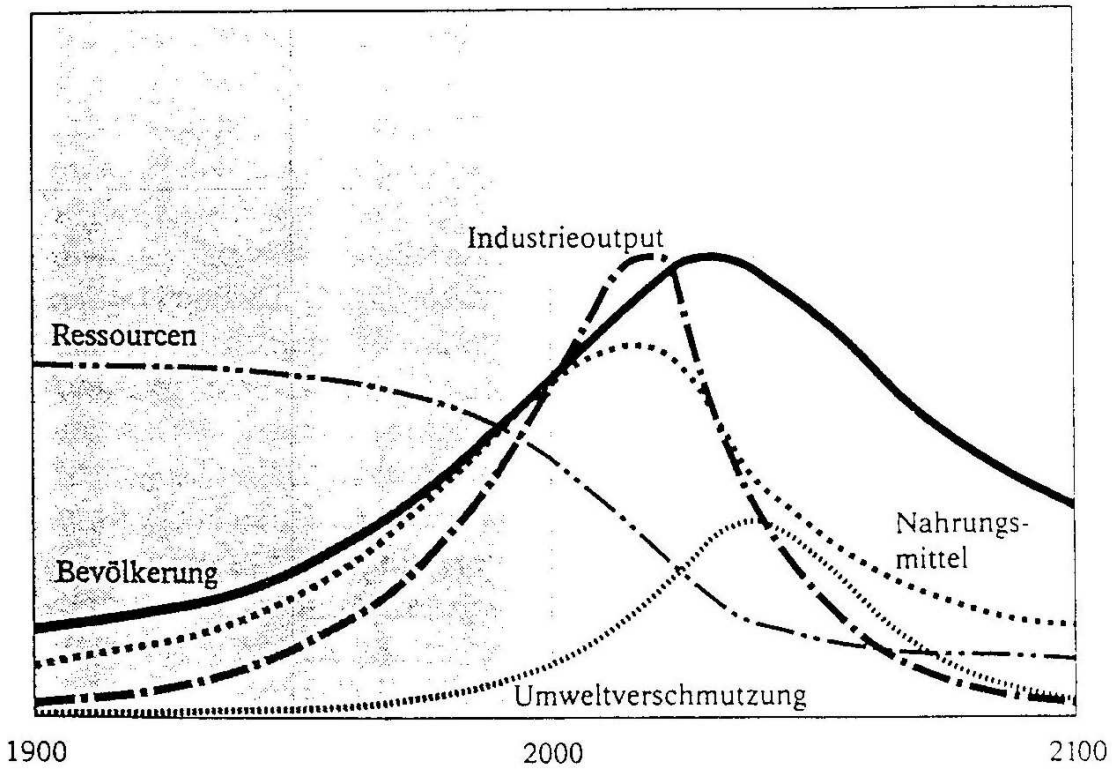


Abb. I.5: Szenario 1: 'Standardlauf' von Grenzen des Wachstums
Quelle: Meadows et al. (1992), S. 166

Intergovernmental Panel on Climate Change (IPCC) http://www.ipcc.ch/
Weltklimabericht der UNEP und WMO

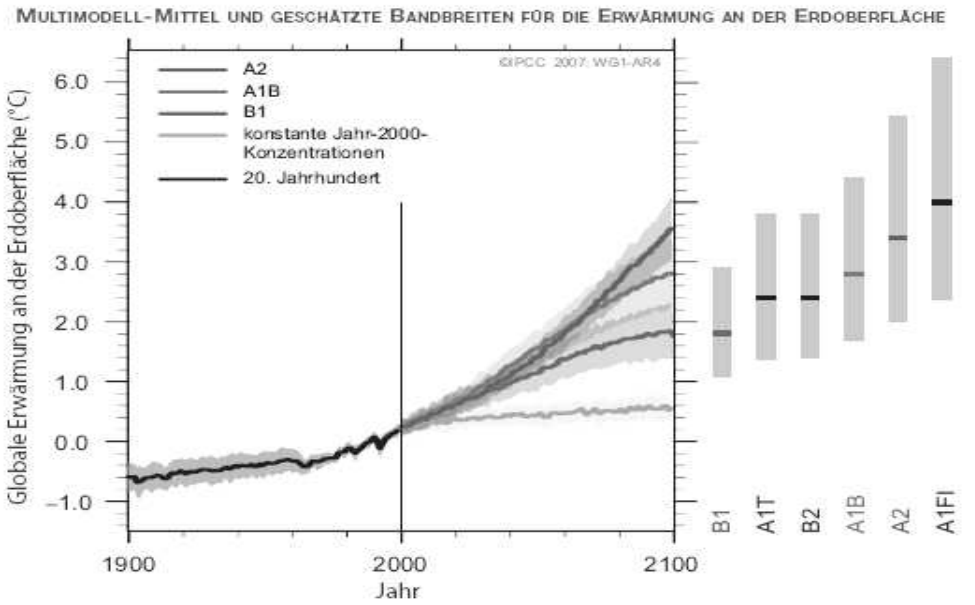


Abbildung SPM.5. *Die durchgezogenen Linien sind globale Multimodell-Mittel der Erwärmung an der Erdoberfläche (relativ zu 1980–99) für die Szenarien A2, A1B und B1, dargestellt als Verlängerungen der Simulationen für das 20. Jahrhundert. Die Schattierung kennzeichnet die Bandbreite von plus/minus einer Standardabweichung der einzelnen Modell-Jahresmittel. Die orange Linie stellt das Resultat des Experiments dar, bei dem die Konzentrationen auf Jahr-2000-Werten konstant gehalten wurden. Die grauen Balken auf der rechten Seite zeigen die beste Schätzung (durchgezogene Linie innerhalb des Balkens) und die abgeschätzte wahrscheinliche Bandbreite für die sechs SRES-Musterszenarien. Die Herleitung der besten Schätzungen und wahrscheinlichen Bandbreiten in den grauen Balken beinhaltet sowohl die AOGCMs im linken Teil der Abbildung als auch die Resultate einer Hierarchie von unabhängigen Modellen sowie beobachtungsgestützte Randbedingungen. (Abbildungen 10.4 und 10.29)*

Abb. I.6: Erwärmung der Erdoberfläche
Quelle: http://www.bmu.de/klimaschutz/downloads/doc/39255.php Stand August 2007

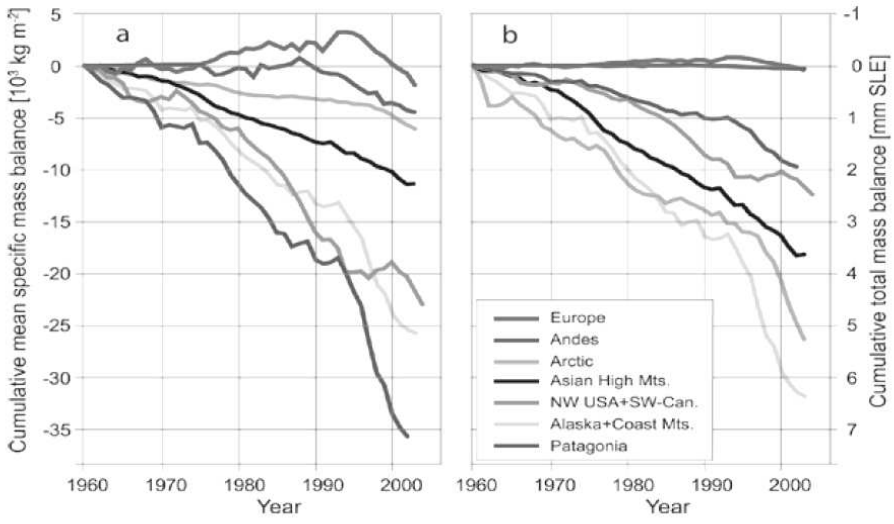


Figure 2.6: *Cumulative mean specific mass balances (a) and cumulative total mass balances (b) of glaciers and ice caps, calculated for large regions (Dyurgerov and Meier, 2005). The mass balance of a glacier is the sum of all mass gains and losses during a hydrological year. Mean specific mass balance is the total mass balance divided by the total surface area of all glaciers and ice caps of a region, and it shows the strength of change in the respective region. Total mass balance is presented as the contribution from each region to sea-level rise. [WGI 4.5.2, Figure 4.15]*

Abb. I.7: Veränderung der Gletschermassen
Quelle: Climate Change and Water, Technical Paper of the Intergovernmental Panel on Climate Change, S. 20

1.4 Anwendungsorientierte Statistik: Fragen zur Wohnsituation aus dem Sozio-Ökonomischen-Panel (SOEP), Leben in Deutschland, Befragung 2007 zur sozialen Lage der Haushalte

Auszug zu Wohnungsfragen aus dem Haushaltsfragebogen

5. Wie würden Sie die Wohngegend hier beschreiben?

Ein reines Wohngebiet mit überwiegend Altbauten ☐
Ein reines Wohngebiet mit überwiegend Neubauten ☐
Ein Mischgebiet mit Wohnungen und Geschäften bzw. Gewerbebetrieben ☐
Ein Geschäftszentrum (Läden, Banken, Verwaltungen) mit wenigen Wohnungen ☐
Ein Gewerbe- bzw. Industriegebiet mit wenigen Wohnungen ☐

10. Wie groß ist die Wohnfläche dieser Wohnung insgesamt? ☐☐☐ qm

11. Und wie viele Räume hat Ihre Wohnung?

☞ *Gemeint sind Räume ab 6 qm, **ohne** Küche und **ohne** Bad.* ☐☐ Räume

12. Wie beurteilen Sie insgesamt die Größe Ihrer Wohnung? Ist sie für Ihren Haushalt ...

– viel zu klein ☐
– etwas zu klein ☐
– gerade richtig ☐
– etwas zu groß ☐
– viel zu groß? ☐

14. Wie ist Ihre Wohnung ausgestattet? Gehört zu Ihrer Wohnung ...

	Ja	Nein
– Küche	☐	☐
– Bad / Dusche innerhalb der Wohnung	☐	☐
– Fließend Warmwasser / Boiler	☐	☐
– WC innerhalb der Wohnung	☐	☐
– Zentralheizung oder Etagenheizung	☐	☐
– Balkon / Terrasse	☐	☐
– Keller / Abstellräume	☐	☐
– Eigener Garten / Gartenbenutzung	☐	☐
– Alarmanlage	☐	☐
– Klimaanlage	☐	☐
– Sonnenkollektor, Solarenergieanlage	☐	☐

15. Haben Sie oder Ihr Vermieter seit Anfang 2006 an dieser Wohnung eine oder mehrere der folgenden Modernisierungen vorgenommen?

Eine Küche eingebaut ☐
Bad, Dusche oder WC innerhalb der Wohnung eingebaut ☐
Zentralheizung oder Etagenheizung eingebaut ☐
Neue Fenster eingebaut ☐
Sonstige größere Maßnahmen ☐

Nein, nichts davon ☐ → *Sie springen auf Frage **18!***

16. Erfolgte diese Modernisierung auf Kosten des Vermieters oder auf Ihre eigenen Kosten?

Auf Kosten des Vermieters ☐ → *Sie springen auf Frage **18!***
Auf eigene Kosten ☐
Teils / teils ☐

22. Wie hoch ist derzeit die monatliche Miete?

☐☐☐☐ EURO Zahle keine Miete ☐ → *Sie springen auf Frage **37!***

2 Begriff, Aufgaben und Entwicklung der Statistik

2.1 Begriff und Aufgaben der Statistik

Statistik:

- quantitative Informationen über bestimmte Tatbestände (Bevölkerungsstatistiken, Umsatzstatistik etc.)
- formale Wissenschaft, die sich mit Methoden der Erhebung, Aufbereitung und Analyse von Information (numerische Daten) beschäftigt

Wirtschafts- und Sozialwissenschaften:
- Entscheidungsgrundlage für private Haushalte, Unternehmen, Staat
- Informationssammlung
- Informationsreduktion (Komprimierung)
- Herausarbeiten von Gesetzmäßigkeiten

Basis der empirischen Wirtschafts- und Sozialforschung

Deskription: beschreibende Statistik

Inferenz: schließende Statistik (Wahrscheinlichkeitsrechnung, Stichproben)

Heinz Grohmann (1986a):

"Statistik ist die methodisch geregelte, zielgerichtete Gewinnung zusammenfassender zahlenmäßiger Informationen über reale Massenerscheinungen." (S. 9)

Statistik im traditionellen Sinn:	beschreibend, ohne Wahrscheinlichkeit (erste Art statistischer Information)
Schließende Statistik:	mit Wahrscheinlichkeit (zweite Art statistischer Information)

Gerd Hansen (1974):

"Die Statistik hat zunächst die Aufgabe, Informationen über die Struktur bestimmter Erscheinungen des Wirtschafts- und Soziallebens zu sammeln, aufzubereiten und zu charakterisieren (**beschreibende Statistik**)." (S. 1)

"Die weitere Aufgabe der Statistik ist es, das Ergebnis einer solchen Beschreibung zu verwenden, um auf **allgemeine Regelmäßigkeiten in wirtschaftlichen und sozialen Beziehungen zu schließen**. Dies geschieht dadurch, daß man den **hypothetischen Befund**, der sich aus einer wissenschaftlichen Theorie über solche Regelmäßigkeiten ableiten läßt, mit dem **empirischen Befund** der statistischen Informationen vergleicht.

Die Statistik liefert auf diese Weise Entscheidungskriterien für die Frage, ob eine wissenschaftliche Theorie mit dem empirischen Befund vereinbar ist oder nicht (Falsifizierung von Theorien im Sinne Poppers). Man spricht hierbei von schließender Statistik (**induktiver Statistik oder statistischer Inferenz**)." (S. 2)

Beispiele:

Problem: Erfassung der Arbeitslosensituation in Niedersachsen (Arbeitsamtsbezirke) und in den fünf neuen Bundesländern

Fragen: Durchschnittliche Dauer der Arbeitslosigkeit?
Anteil der Altersgruppen?
Regionale Differenzierung?
Einfluß der Berufsqualifikation?
Vergleichbarkeit der Informationen?

Fazit: Beschreibend → Deskriptive Statistik

Problem: Qualitätskontrolle im Produktionsbereich eines Unternehmens

Fragen: Annahme oder Ablehnung des 'Loses'?
Eingriff in den Produktionsprozeß?

Fazit: Operationale Funktion (Entscheidungshilfe) → Schließende, Induktive Statistik

Problem: Wirtschaftspolitische Behauptung: Transfer des Staates (z.B. Arbeitslosengeld) verlängert signifikant die Arbeitslosigkeitsdauer

Fragen: Wie kann dies operationalisiert werden?
Gibt es zwei 'identische Gruppen' mit und ohne Transfers (Soziale Experimente: USA z.B. New Jersey Income Maintenance Experiment)?
Wie ist das Verhalten zu quantifizieren?
Daten: Querschnitt (Umfrage), Panel?

Fazit: Deskription und Inferenz im Rahmen einer empirischen Wirtschaftsforschung; Raum-sachliche, Raum-zeitliche Begrenzung der Aussage

2.2 Zur geschichtlichen Entwicklung

Geschichtlich: **Erhebungsstatistiken**

Erhebung über staatskundliche Phänomene:
- Bevölkerung, Ackerfläche, Goldbestand (Ägypten 2500 v. Chr. etc.)

Neuere Zeit: **Wahrscheinlichkeitsrechnung**

Blaise PASCAL (1623-1662), Pierre de FERMAT (1601-1665)
de Moivre, Laplace: Frankreich
Bernoulli, Euler: Schweiz
Gauss: Deutschland
Kolmogorov, Tschebyscheff, Markoff: Rußland
19.-20. Jahrhundert Induktive Statistik

Statistik ist für viele Bereiche von Bedeutung:

Wirtschaft:	Unternehmen (Marketing), Private Haushalte (Einkommen und Konsum), Staat (Wirtschafts- und Sozialpolitik) etc.
Soziologie:	Gruppenverhalten (Sozioökonomie)
Medizin:	Rauchgewohnheiten → Lungenkrebs
Psychologie:	Lernerfolg
Physik/Mathematik:	Atome, Unschärfebereiche, Zufall
Biologie:	Mendelsche Gesetze
Umwelt:	Verschmutzungsgrade
etc.	

Benutzergruppen statistischer Information im Überblick finden sich in Abb. I.6.

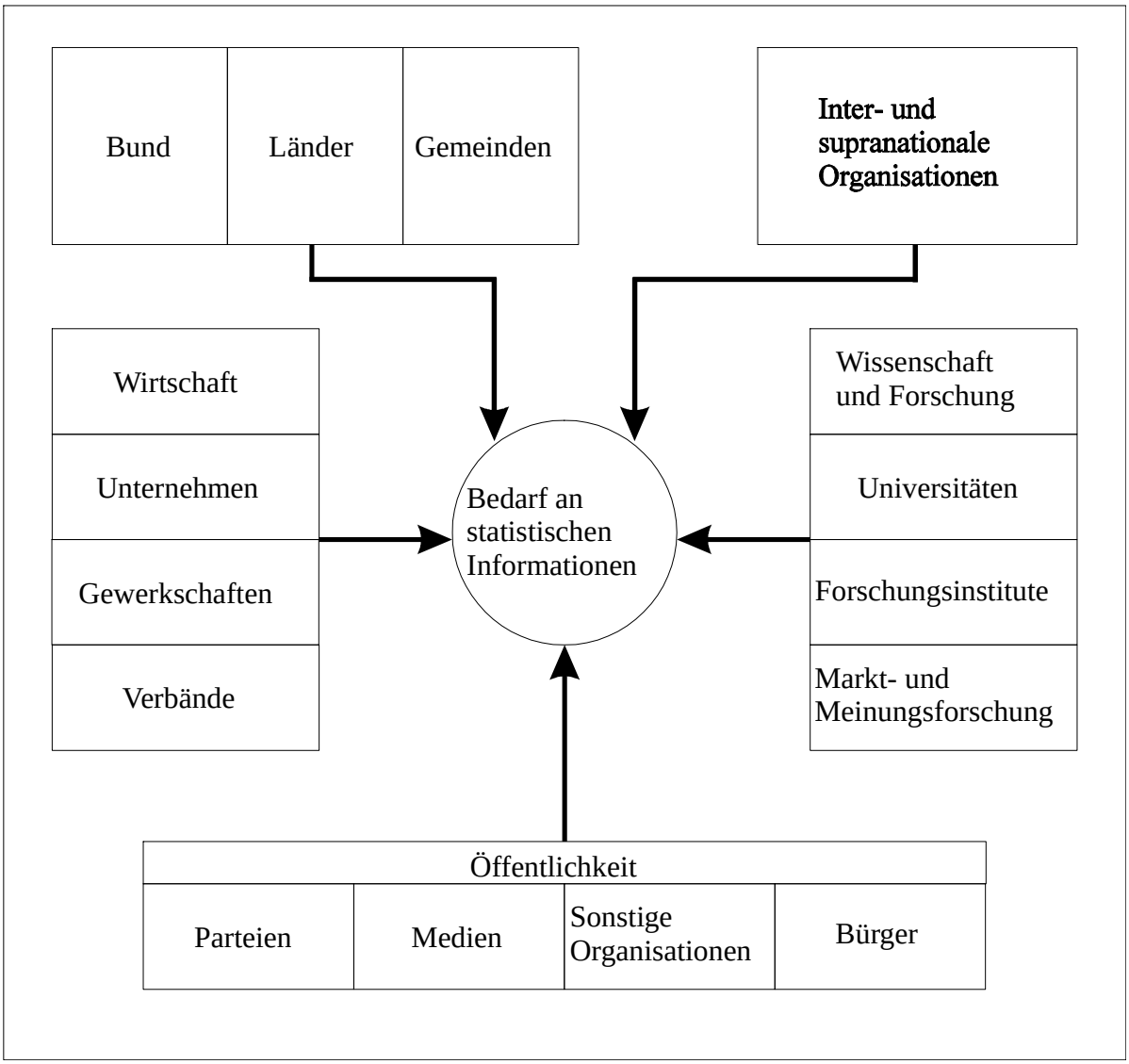


Abb. I.6: Benutzergruppen statistischer Informationen
Quelle: Statistisches Bundesamt (1989)

3 Träger der Wirtschaftsstatistiken und statistische Quellen

3.1 Amtliche Statistik

Legale Basis: Gesetze, Rechtsverordnungen (Bundesstatistikgesetz: BStatG 1987)

Statistische Ämter

Statistisches Bundesamt, Statistische Landesämter, Statistische Ämter der Städte, Gemeinden und Kommunen, Forschungsdatenzentren der Statistischen Ämter des Bundes und der Länder

Das Statistische Bundesamt gibt Informationsbroschüren über seine Aufgabe, Aufbau und Arbeitsweise heraus, so z.B.:
- Statistisch gesehen

Diese Broschüe und weiteres Material sind kostenlos erhältlich bei:

Statistisches Bundesamt
Gustav-Stresemann-Ring 11
65189 Wiesbaden
Tel.: 0611/75-2405
Fax: 0611/75-3330
www.destatis.de
info@destatis.de

Allgemeine Aufgabenbeschreibung: Statistisches Bundesamt (Hrsg.), Das Arbeitsgebiet der Bundesstatistik, Kohlhammer Verlag, Mainz 1988

Zum Ablauf von Bundesstatistiken vgl. Abb. I.7.

Veröffentlichungen des Statistischen Bundesamtes

- Statistisches Jahrbuch für die Bundesrepublik Deutschland
- Wirtschaft und Statistik (monatlich)
- Fachserien 1-19

Zur Übersicht des Veröffentlichungssystems des Statistischen Bundesamtes vgl. Abb. I.8.

GENESIS-Online: Datenbankzugriff auf das statistische Informationssystem des Bundes

Beispiele:

Statistiken aus Befragungen:
Volkszählung (1970/71, 1987), Mikrozensus (jährlich), Einkommens- und Verbrauchsstichprobe (EVS), Zeitbudgeterhebung (1991/92 und 2001/02)...

Ressortstatistik

- Deutsche Bundesbank www.bundesbank.de
  Veröffentlichungen:
  - Monatsberichte
  - Statistische Beihefte zu den Monatsberichten
    Reihe Bankenstatistik
    Reihe Kapitalmarktstatistik
    Reihe Zahlungsbilanzstatistik
    Reihe Saisonbereinigte Wirtschaftszahlen
    Reihe Devisenkursstatistik

- Bundesagentur für Arbeit www.arbeitsagentur.de
  Veröffentlichungen:
  - Amtliche Nachrichten der BA (ANBA)
  - Monatlicher Arbeitsmarktbericht
  - Jahresbericht

- IAB: Institut für Arbeitsmarkt- und Berufsforschung www.iab.de
  Veröffentlichungen:
  - Mitteilungen aus der Arbeitsmarkt- und Berufsforschung

Weitere Hinweise: Rinne (1994), Kunz (1987), Grohmann (1986a), v.d. Lippe (1996)

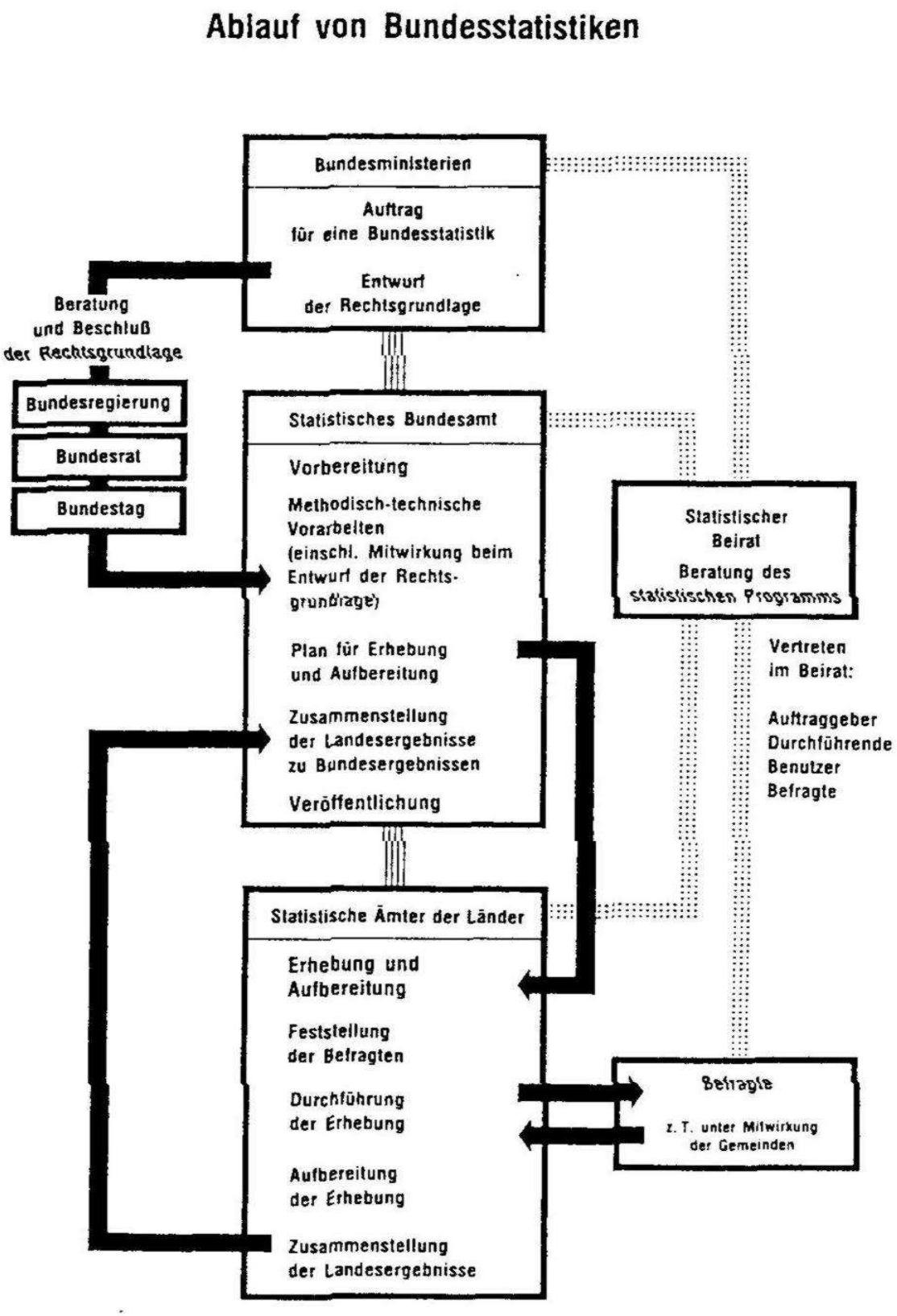


Abb. I.7: Ablauf von Bundesstatistiken
Quelle: Statistisches Bundesamt (1988), S. 47

Zusammenfassende Veröffentlichungen			
Allgemeine Querschnittsver-öffentlichungen	Thematische Querschnittsver-öffentlichungen	Veröffentlichungen zu Organisations und Methodenfragen	Kurzbroschüren

Fachserien
1 Bevölkerung und Erwerbstätigkeit
2 Unternehmen und Arbeitsstätten
3 Land- und Forstwirtschaft, Fischerei
4 Produzierendes Gewerbe
5 Bautätigkeit und Wohnungen
6 Handel, Gastgewerbe, Reiseverkehr
7 Außenhandel
8 Verkehr
9 Geld und Kredit
10 Rechtspflege
11 Bildung und Kultur
12 Gesundheitswesen
13 Sozialleistungen
14 Finanzen und Steuern
15 Wirtschaftsrechnungen
16 Löhne und Gehälter
17 Preise
18 Volkswirtschaftliche Gesamtrechnungen
19 Umweltschutz

Systematische Verzeichnisse				
Unternehmens und Betriebs-systematiken	Güter-systematiken	Personen-systematiken	Regional-systematiken	Sonstige Systematiken

Karten

Statistik des Auslandes

Fremdsprachige Veröffentlichungen

Abb. I.8: Veröffentlichungssystem des Statistischen Bundesamtes
Quelle: Statistisches Bundesamt (1989)

3.2 Nichtamtliche Statistik

- Wirtschaftsverbände, Berufsorganisationen
- Industrie- und Handelkammer (IHK), Kammern
- Markt- und Meinungsforschungsinstitute (Infratest, Marplan, Emnid, Allensbach…)
- Arbeitnehmer- und Arbeitgeberorganisationen
- Wirtschaftsforschungsinstitute

  von Interessenverbänden:

  - IW - Institut der Deutschen Wirtschaft (Köln), www.iwkoeln.de
  - WSI - Wirtschafts- und Sozialwissenschaftliches Institut (Düsseldorf), www.wsi.de

  gemeinnützig und unabhängig:

  - DIW - Deutsches Institut für Wirtschaftsforschung (Berlin), www.diw.de, z.B. Vierteljährliche Volkswirtschaftliche Gesamtrechnung
  - ZBW - Deutsche Zentralbibliothek für Wirtschaftswissenschaften (Kiel), www.zbw.eu, Leibniz-Informationszentrum Wirtschaft
  - Ifo - 'Information und Forschung' - Ifo Institut für Wirtschaftsforschung (München), www.ifo.de, CESifo Economic Studies (vierteljährlich), Ifo-Geschäftsklimaindizes
  - IfW - Institut für Weltwirtschaft an der Universität Kiel (Kiel), www.uni-kiel.de/ifw
  - RWI - Rheinisch-Westfälisches Institut für Wirtschaftsforschung (Essen), www.rwi-essen.de
  - IWH - Institut für Wirtschaftsforschung Halle (Halle a.d. Saale), www.iwh-halle.de

  Jahresgutachten von DIW, Ifo, IfW, RWI und IWH

- GESIS, Leibniz-Institut für Sozialwissenschaften, Mitglied der Gesellschaft Sozialwissenschaftlicher Infrastruktureinrichtungen e.V. (GESIS), www.gesis.org/das-institut/
  Wissenschaftliche Abteilungen:
  - Survey Design and Methodology (SDM)
  - Dauerbeobachtung der Gesellschaft (DBG): German Microdata Lab – GML, Zentrum für Sozialindikatorenforschung – Zsi, Survey Programme: Allgemeine Bevölkerungsumfrage ALLBUS, International Social Science Programme ISSP, Comparative Study of Electoral Systems CSES, German Longitudinal Election Study GLES
  - Datenarchiv für die Sozialwissenschaften: Datenservice zu nationalen und international-vergleichenden Umfragen zu soziologischen und politikwissenschaftlichen Fragestellungen, ALLBUS

3.3 Internationale Organisationen

- UN (Statistical Yearbook, Demographic Yearbook, Yearbook of National Accounts Statistics, New York), http://unstats.un.org
- OECD (Paris), http://www.oecd.org
- EUROSTAT, http://epp.eurostat.ec.europa.eu
- ILO (International Labour Organization, Genf), http://www.ilo.org
- WHO (World Health Organisation, Genf), http://www.who.int/GHO/
- IMF (International Monetary Fund, Washington), http://www.imf.org
- Multinationale Konzerne (Fachabteilungen)

3.4 Aufgaben und Quellen der Wirtschaftsstatistik im vereinten Deutschland

Ehemalige DDR

- Ministerrat der DDR, Staatliche Zentralverwaltung für Statistik
- oft Vollerhebung, z.B. Berufstätigenerhebung (BTE) ohne den X-Bereich (Stasi, Armee)

Gravierende Unterschiede in den Statistiksystemen (vor allem: Wirtschaftsstatistik)

Im Rahmen der Wiedervereinigung wurde eine Umstrukturierung bzw. ein Neuaufbau der Einrichtungen vorgenommen (z.B. Arbeitsämter, BA)

Zur Vereinheitlichungsdiskussion vgl. Allgemeines Statistisches Archiv, Bd. 76, 1992

Statistiken für die fünf neuen Bundesländer

- Sozialreport '90
- Statistisches Bundesamt: Neue Publikationen (monatlich)
- Übersicht zum Stand der Einführung wichtiger ausgewählter Bundesstatistiken
- IAB-Werkstattbericht 'Neue Bundesländer'
- Presseinformationen der Bundesagentur für Arbeit
- Sozio-ökonomisches Panel (SOEP-Ost)
  1. Welle 1984 (West), 1. Welle 1990 (Ost)
- Infratest- 'Befragung Ost'

Bei einem Ereignis wie der deutschen Wiedervereinigung ist der sich hieraus ergebende strukturelle Bruch in einer ökonomisch/statistischen Bewertung besonders zu beachten. Konkret entsteht dieser hier durch die Ausweitung der Grundgesamtheit um rund 16 Millionen Menschen mit grundlegend anderen demografischen Vorrausetzungen z.B. im Bereich Lebensumstände, Einkommen, Lebensstandard, Gesundheit usw. Die sich durch die Einbeziehung dieser Faktoren ergebenden Veränderungen müssen vor allem bei der Vergleichbarkeit der Daten berücksichtigt werden (z.B. wird in Gutachten meist gesondert die Situation vor (1989) und nach der Widervereinigung (1990) ausgewiesen → siehe z.B. tabellarischen Anhang im Gutachten der Sachverständigenrates Wirtschaft).

4 Das Adäquationsproblem und einige wissenschaftstheoretische Bemerkungen

4.1 Wissenschaftstheoretische Grundlagen: Zur Struktur und Anwendung wissenschaftlicher Theorien

Erklärung der Welt, 'Kritischer Rationalismus'

Der Kritische Rationalismus versteht unter einer Theorie allgemein ein System wissenschaftlicher Sätze über die Realität.

Albert (1964):

"Die zentralen Bestandteile realwissenschaftlicher Theorien haben den Charakter von nomologischen Hypothesen (Gesetzen), also empirisch gehaltvollen Aussagen über die Struktur der Realität, die infolgedessen anhand der Tatsachen nachgeprüft werden können."

Logische Struktur einer Theorie:

Aus Axiomen (Grundsätzen) werden Theoreme (abgeleitete Sätze) deduziert.

Inhaltliche Struktur einer Theorie:

1. Geltungsmodus: Nur Aussagen mit empirischem Geltungsanspruch (nicht nur denkbare Situationen)
2. Widerspruchsfreiheit
3. Operationalität (eindeutig definierte Begriffe, überprüfbar)
4. Empirischer Gehalt
5. Prüfbarkeit (Falsifizierbarkeit) und Bewährung
6. Allgemeinheit

Für ein **Explanandum**, das den zu erklärenden Tatbestand beschreibt, ist ein **Explanans** zu finden, das ein **allgemeines Gesetz** und **Anwendungsbedingungen** enthält (Albert 1964).

Beispiel:

Explanandum: 'Die Ausgaben für den privaten Konsum sind gestiegen'

Konsumhypothese: 'Wenn sich das verfügbare Einkommen der privaten Haushalte um einen bestimmten Betrag erhöht, dann steigen die Konsumausgaben im Mittel um einen bestimmten (anderen) Betrag'

→ Allgemeines Gesetz: 'Das verfügbare Einkommen ist um einen bestimmten Betrag gestiegen'

Explanans = Allgemeines Gesetz und Anwendungsbedingung

Aus Explanans kann dann das Explanandum logisch abgeleitet werden.: Wenn das verfügbare Einkommen um x steigt, erhöhen sich die Konsumausgaben um f(x).

Karl Popper (1964):

$$\left.\begin{array}{ll} G & (\text{allgemeines Gesetz: Wenn A, dann C}) \\ A & (\text{singuläre Aussage: Nun A}) \end{array}\right\} \text{Explanans (Prämissen)}$$

Also C (Conclusion) Explanandum

Prüfung: Bewährungsgrad einer Theorie durch permanente Falsifikationsversuche feststellen

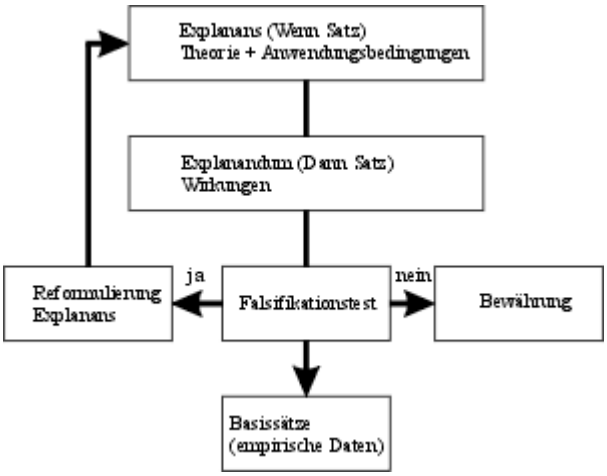


Abb. I.9: Der Falsifikationsprozess
Quelle: Hujer und Cremer (1978), S. 12

Thomas Kuhn betrachtet in seinem Buch 'Structure of Scientific Revolutions', 2nd ed., Chicago 1970, ausführlich den **Paradigmenwechsel**. Kuhn interpretiert ein Paradigma als eine Menge von Wechselbeziehungen, die aber noch unterentwickelt sind. Ein Paradigma ist dann ein Gedankengebäude, das Antworten zu bestimmten Fragen liefern kann, auf denen dann Ausweitungen der Theorie vorgenommen werden können. Diese 'normal science' beschäftigt sich also mit dem 'puzzle solving': Offen gebliebene Fragen einer revolutionären Theorie werden gelöst.

Neue Paradigmen tauchen dann auf, wenn Widersprüche in den bestehenden Paradigmen entdeckt werden, d.h. alte Paradigmen werden dann fallengelassen, wenn sie immer mehr Fragen nicht beantworten können. Der Zeitablauf ist also durch wechselseitige Phasen von normaler und revolutionärer Wissenschaft gekennzeichnet.

4.2 Das Adäquationsproblem: Allgemeine Problemstellung und statistische Operationalisierung

Der theoretische (idealtypische) Begriff ist mit einem empirisch feststellbaren Begriff zu verbinden.

Adäquationsproblem:

Die Diskrepanz zwischen theoretischem und statistischem Begriff sollte so klein wie möglich werden (Grohmann (1986a), S. 18, Blind, **Frankfurter Schule**).

Beispiel:

Frage: Es ist zu klären, ob und wie die Käufe eines Verbrauchsgutes von der Anzahl, Größe und dem Einkommen der Haushalte abhängen.

Theorie: Wirtschaftswissenschaften, Welche Haushaltsdefinition soll verwendet werden? (möglich: Haushalt = Wirtschaftseinheit, d.h. Gruppe von Personen, die einen gemeinsamen Verbrauchsplan aufstellen, die gemeinsam wirtschaftlich handeln) (Mikroökonomie).

Realität: Eine gemeinsame Kaufentscheidung wird eher selten getroffen.

Praxis: Laut Volkszählung umfaßt ein Haushalt alle diejenigen Personen, die in der gleichen Wohnung leben und den Lebensunterhalt überwiegend gemeinsam betreiben.

Welche Einkommensdefinition soll verwendet werden? Das Geldeinkommen (Lohn und Gehalt) gehört grundsätzlich zum Einkommen. Was ist aber mit dem 13. Monatsgehalt, einmaligen Zahlungen etc.? Wie verhält es sich mit einer mietfreien Werkswohnung, einem Firmenwagen oder Naturaleinkommen, 'fringe benefits' oder laufendem Brutto-/Nettoeinkommen?

5 Sachgerechte Interpretation: 'How (not) to lie with statistics'

Illustratives zu 'How to lie with statistics': Huff (1978), Krämer (1991), Schwarze (1990), S. 17-19.

5.1 Some pitfalls

- willkürlicher Bezug

Beispiel:

Von 2500 Studenten nehmen 50 an der Statistik-Klausur teil. Keiner von ihnen besteht die Klausur. Der Dozent behauptet, die Durchfallquote betrage zwei Prozent.

- fehlende Sachlogik

Beispiel:

Es wird beobachtet, dass der Anstieg von Storchbrütungen in einer Region mit einem Anstieg der Geburten einher geht. Basierend auf dieser Beobachtung wird ein statistisch 'bewiesener' Zusammenhang vermutet.

- Herausgreifen bestimmter 'passender' Werte

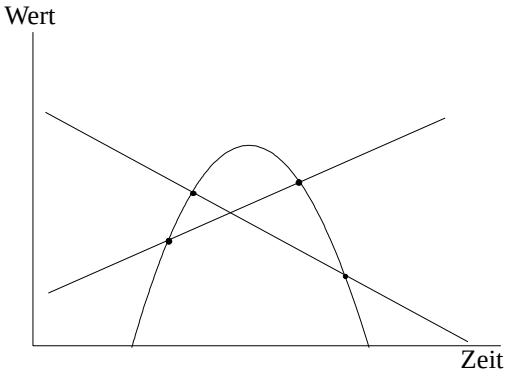


Abb. I.10: Herausgreifen bestimmter 'passender' Werte

- Maßstabsmanipulation:

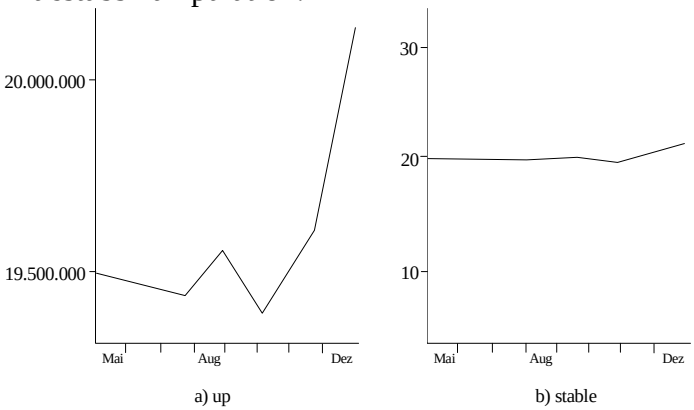


Abb. I.11: Maßstabsmanipulation

5.2 How not to lie with statistics

- Wahl eines problemgerechten operablen statistischen Begriffs (Adäquationsproblem)
- eindeutige Bezugs- und Berechnungsangaben
- sachgerechte, ehrliche Präsentation
- adäquater Vergleich
- Vorsicht mit Extrapolationen, Vorhersagen weit außerhalb des Datenstützbereichs unzulässig
- ... 'be honest'

6 Statistische Einheiten und statistische Massen

6.1 Statistische Einheiten

Statistische Einheiten = Merkmalsträger der Untersuchung, Einzelobjekt, Proband

Statistische Einheiten sind real, klar voneinander abgrenzbar, zählbar:
z.B. Haushalt, Person, Betrieb, Gemeinde, Fläche, Region, PKW, Beobachtungspunkt in einer Stadt (bei Verkehrszählung) etc.

Identifikations- oder Abgrenzungskriterien:
- sachlich
- räumlich
- zeitlich

Beispiel:

Sfb 3 Nebenerwerbstätigkeitsumfrage 1984 (Merz, Helberger und Schneider (1985)):

Statistische Einheit: Personen in Privathaushalten (nicht in Anstalten)

Identifikation:

- sachlich: Person über 14 Jahre mit definierter Nebenerwerbstätigkeit (Mehrfacherwerbstätigkeit oder Mehrfachtätigkeit?)
- räumlich: Gebiet der Bundesrepublik Deutschland (einschließlich West-Berlin)
- zeitlich: vergangene drei Monate aus einem 'Quartal' 1984

6.2 Statistische Massen

Statistische Masse = Gesamtheit aller statistischen Einheiten, die vom Untersuchungsziel her gleichartig sind (übereinstimmende Identifikationskriterien, sachliche, räumliche und zeitliche Abgrenzung)

Beispiele:

- Zahl der Arbeitslosen in der Bundesrepublik im Monat Februar 2010
- Rechnungen des Unternehmens McAlles im Monat Oktober 2010

Arten statistischer Massen:

- **Bestandsmassen**

  Für einen **Zeitpunkt** definiert, z.B. Kassenbestand eines Warenhauses am 31.12.2010, Wohnbevölkerung in der Bundesrepublik am Stichtag (z.B. 25.5.1987, Stichtag der letzten Volkszählung)

- **Ereignis- (Bewegungs-) Massen**

  Für einen **Zeitraum** definiert, z.B. Eheschließungen in der Bundesrepublik im Jahre 2010, Scheckeingänge der Bank X im Monat März 2010

  Die Verknüpfung von Bestands- und Bewegungsmassen erfolgt durch **Fortschreibung**:

$$\underset{\text{(Bestandsmasse)}}{\text{Anfangsbestand}} + \underset{\text{(Bewegungsmasse)}}{\text{Zugang}} \text{ ./. } \text{Abgang} = \underset{\text{(Bestandsmasse)}}{\text{Endbestand}}$$

Beispiel:

	Zugelassene Kraftfahrzeuge in Lüneburg am	1.1.2010
+	Neuzulassungen	1.1. - 31.12.2010
./.	Abmeldungen	1.1. - 31.12.2010
=	zugelassene Kfz in Lüneburg am	31.12.2010

7 Merkmale, Merkmalsausprägungen und Meßskalen

7.1 Merkmale und Merkmalsausprägungen

Merkmal = **Eigenschaft** einer statistischen Einheit (Merkmalsträger)

Beispiele:

- Merkmale einer Person (statistische Einheit):
  Alter, Geschlecht, Einkommen, Berufsausbildung... (sozioökonomische Merkmale)
- Merkmale eines Haushalts (statistische Einheit):
  Haushaltsgröße, Alter des 'Haushaltsvorstandes', Anzahl der Kinder, Anzahl der Erwerbstätigen...

- Merkmal X: ALTER, SEX, AGE OF HEAD...

Merkmalsausprägung = **Mögliche Werte** (Kategorien) eines Merkmals

Beispiele:

Statistische Einheit = Studentin

Merkmal:	ALTER	STUDIENFACH
Merkmalsausprägung:	21 Jahre	BWL
(Merkmalswert, Beobachtungswert)		

7.2 Meßskalen und ihre Eigenschaften

Meßskalen der Merkmalsausprägungen haben unterschiedliches Meßniveau:

Nominalskala

- keine natürliche Reihenfolge, Merkmalsausprägungen sind gleichberechtigt nebeneinander:
  z.B. Geschlecht, Hautfarbe, Religion, Staatsangehörigkeit (Codes)

Ordinalskala

- Rangskala, natürliche Rangordnung, Abstände nicht quantifizierbar:
  z.B. Examensnoten, Güteklassen, Bundesligatabelle, * oder *** [Sterne] Hotel

Metrische Skala

- Kardinalskala, Abstände sind angebbar (Maßsystem):

  - *Intervallskala*:
    - mit Abständen, aber ohne Bezugspunkt:
      z.B. Abstand zwischen Gefrier- und Siedepunkt des Wassers in 100 Teilen, Kalenderzeit

  - *Verhältnisskala*:
    - mit Abständen und mit natürlichem Bezugspunkt:
      z.B. Körpergröße (cm), Alter (Jahre), Einkommen

  (- *Absolutskala*:
    - metrische Skala mit natürlichem Nullpunkt und natürlicher Einheit:
      z.B. Stückzahlen)

Zur schematischen Abgrenzung von Meßskalen siehe Tab. I.2.

Tab. I.2: Schematische Abgrenzung von Meßskalen

Merkmale	Skala	gleich oder verschieden	natürliche Reihenfolge	konstanter Wertabstand	natürlicher Nullpunkt	natürliche Einheit	Rechenoperationen
qualitative	Nominalskala	X					Häufigkeiten
intensitätsmäßige	Ordinalskala	X	X				Median
quantitative	Intervallskala	X	X	X			Addition und Subtraktion
quantitative	Verhältnisskala	X	X	X	X		Division und Multiplikation
quantitative	Absolutskala	X	X	X	X	X	Division und Multiplikation

7.3 Diskrete und stetige Merkmale

Diskretes Merkmal

abzählbare Ausprägungen:
z.B. Erwerbstätigkeit (0/1), Anzahl der Studentinnen und Studenten im Hörsaal

Stetiges Merkmal

überabzählbare Ausprägungen (kontinuierlich):
z.B. Länge, Gewicht
(Approximativ stetig: z.B. Geld)

7.4 Quantitative und qualitative Merkmale

Quantitative Merkmale

Abstände zwischen Merkmalsausprägungen sind durch reelle Zahlen meßbar:
z.B. Länge, Alter

Qualitative Merkmale

Kategoriale Abstufung:
z.B. Farbe, Noten

Mit dieser Unterscheidung ist es schwierig, ordinalskalierte Daten einzuordnen.

8 Statistische Untersuchungen: Erhebung, Aufbereitung und Analyse

8.1 Vorgehensweise bei statistischen Untersuchungen

1. Schritt: Abbildung materieller Fragestellung in statistisches Konstrukt

Ausgangspunkt: Problemstellung aus Theorie (Empirische Überprüfung von Hypothesen) oder Praxis (Wert- oder Zielvorstellungen): z.B. Überprüfung der Theorie Dualer Arbeitsmärkte, Erreichung des Ziels 'Hoher Beschäftigungsstand'

Theorie, Praxis → materielle Fragestellung

- Entwicklung eines Begriffssystems aus der Fachwissenschaft (idealtypischer Begriff):

- z.B. Entwicklung eines Modells über den Arbeitsmarkt,
  Definiton von Arbeitsmarktindikatoren für den Zielkomplex 'Hoher Beschäftigungsstand'

- Übersetzung von materieller Fragestellung in statistische Konstrukte, Messung (Adäquationsproblem)

- z.B. Messung der Arbeitslosenquote, Ermittlung der Zahl der offenen Stellen

2. Schritt: Erhebung

Nach Festlegung der zu untersuchenden Objekte und Merkmale erfolgt die Beobachtung oder Befragung (Erhebung) (siehe VI.2).

3. Schritt: Aufbereitung und Darstellung des Beobachtungsmaterials

Ziel: Verdichtung, Straffung und Strukturierung des Urmaterials (bei Befragungen, Umfragen: 'Editing' der Daten)

Gruppierung der Daten nach Merkmalsklassen, grafische/tabellarische Darstellung der klassifizierten Daten

Berechnung von beschreibenden Maßzahlen wie Mittelwerte, Streuung, Zusammenhangsmaße

4. Schritt: Analyse durch Schluß von der Stichprobe auf die Grundgesamtheit

Ziel: Aussagen über unbekannte Gesamtmasse (Grundgesamtheit), z.B. Studenten der Universität Lüneburg, Bevölkerung Niedersachsens

Die Ergebnisse der bekannten Teilmasse (Stichprobe) werden auf die Grundgesamtheit mit einer bestimmten Wahrscheinlichkeit übertragen, z.B. durchschnittliche Körpergröße einer Stichprobe der Studenten der Statistik I Vorlesung → alle Studenten (Lüneburg)

Methoden:

- Schätzen der unbekannten Größen (Parameter)
- Testen von Hypothesen über diese Größen

Beispiele:

a) Anteil der Kommilitoninnen und Kommilitonen an den Studenten

b) Individuelle Wirkungen der Steuerreform '90: Ist der geschätzte Koeffizient b zu der Anzahl der Kinder signifikant von Null verschieden, um das Arbeitsangebot im Nebenerwerb zu erklären?

$$= f\left(..., \underset{*}{\underset{\sim}{b}} \cdot \text{Anzahl der Kinder}, ... \right)$$

*=signifikant von Null verschieden?

---

5. Schritt: Sachgerechte Interpretation der Ergebnisse

Interpretation der Ergebnisse im Sinne der untersuchten materiellen Frage

Beachtung der Einschränkungen aus:

- Definition
- verwendeten Methoden
- zeitliche, räumliche und sachliche Begrenzung
- Güte der Daten

Falsifikation: wissenschaftliche Theorie oder politische Zielvorstellung widerlegt oder nicht?

Aus den fünf Schritten seien zwei vertieft: Erhebung sowie Aufbereitung und Analyse.

8.2 Erhebung: Erhebungsarten und Erhebungstechnik

Erhebung

Bei einer vorbestimmten Menge von **Merkmalsträgern** (Untersuchungseinheiten, Objekte, Probanden) werden eine Anzahl von **Merkmalen** erhoben und deren **Ausprägungen** erfaßt.

Erhebungsarten

- Primärstatistische Erhebungen
  - ausschließlich zu statistischen Zwecken

- Sekundärstatistische Erhebungen
  - bereits vorhandene, zunächst für andere Zwecke gesammelte Daten (z.B. Lohnsteuerstatistik)

- Vollerhebung
  - alle Einheiten werden erfaßt (z.B. Volkszählung)

- Teilerhebung
  - ausgewählte Einheiten (z.B. Mikrozensus)

Erhebungstechnik

- Schriftliche Befragung, Fragebogen (Questionnaire)
  - offene Fragen (ohne Antwortvorgabe), geschlossene Fragen (mit vollständigen Antwortvorgaben)

- Mündliche Befragung, Interviewer, CATI (Computer Aided Telephone Interview), CAPI
  - z.B. ISR, Ann Arbor Michigan, Panel Study of Income Dynamic (PSID), Infratest Sozialforschung, München

- Online-Erhebung bzw. -Umfrage
  - z.B. FFB-Online Erhebung zu Freien Berufen 2005/2006

- Beobachtung

8.3 Aufbereitung und Analyse

Aufbereitung und Analyse umfaßt alles vom Urmaterial (individueller Fragebogen, Zählblätter etc.) bis zum Ergebnis (Grafik, Tabelle)

1. Prüfen des Urmaterials auf Vollständigkeit, Widerspruchsfreiheit und Glaubwürdigkeit (Editing)

Widersprüche:

sozial:	80jähriger Schüler?
ökonomisch:	Großbetrieb mit Jahresumsatz von EUR 100,-
gesetzlich:	dreijährige Witwe
instiutionell:	katholischer Pfarrer, verheiratet

2. Verschlüsseln (Kodierung) von qualitativen Merkmalsausprägungen:

Zuordnung von Variablenwerten wie z.B.:

SEX = 1: männlich, 2 : weiblich (dummy variables)

FAMSTD = 1: ledig, 2: verheiratet, 3: geschieden, 4: verwitwet

Mehrstellige Codes: Arbeitsstätte, Beruf, Waren, Krankheit etc.
z.B. Beruf: ISCO-Code (Systematik der Wirtschaftszweige)

3. Übertragen von Informationen auf Datenträger (z.B. CD-ROM) und maschinell unterstütztes Editing

4. Auswerten des Urmaterials nach Gruppen, Ausprägungen

- Deskription: Grafiken, Tabellen
- Inferenz: Regressionsanalyse, Multivariate Analyse

- Erwerb von fertigen Ergebnissen (Tabellen aus dem Statistischen Bundesamt)
- Eigene Software, Anwendungssoftware zur Analyse der Individualinformationen
- Datenbanken (dbase, ORACLE mit SQL,...)

9 Tabellarische und grafische Darstellung

9.1 Zur Präsentation von Informationen

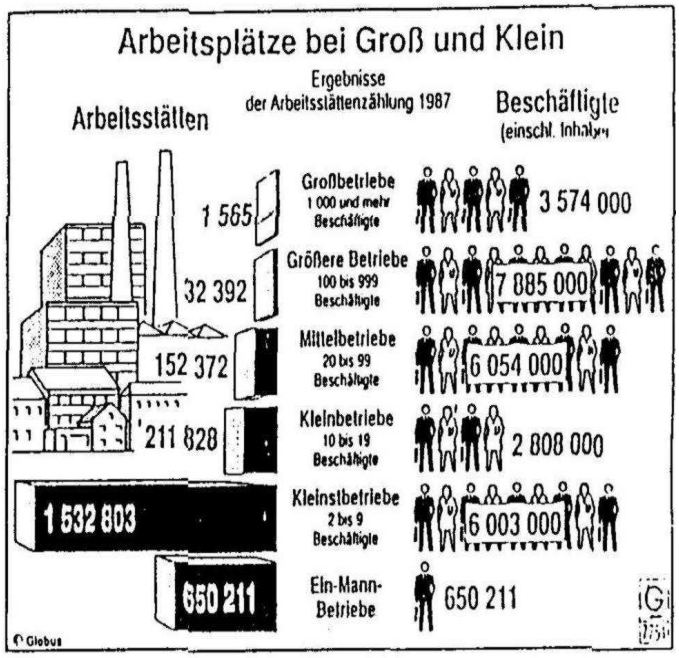
KLAR . . . **GROß** . . . *motivierend!*

'Weniger ist **MEHR**!!!', vollständige Beschriftung!

Beispiele:

- VENN-Diagramm aus einführendem Beispiel zur Mikroanalyse der Steuerreform
- Abbildungen mit Bildinformation (vgl. Abb. I.12a,b)
- charts (Software)
- Wirkung auch durch Text alleine:

> *Die Erde ist etwa 4.500.000.000 Jahre alt.*
>
> *Vergleicht man diese Zeitspanne mit dem Leben eines 45jährigen Menschen, so traten die ersten Säugetiere vor acht Monaten in Erscheinung und Menschen gibt es erst seit wenigen Tagen.*
>
> *Vor etwa einer Stunde erlernte der Mensch den Ackerbau und vor einer Minute begann die industrielle Revolution.*
>
> *In diesen 60 Sekunden hat der Mensch die Rohstoffreserven unseres Planeten geplündert. Boden, Wasser und Luft verseucht und unzählige Pflanzen und Tiere ausgerottet.*



Wirtschaftsdaten der EU-Länder

EU der 27 - Bevölkerung und Wirtschaftskraft

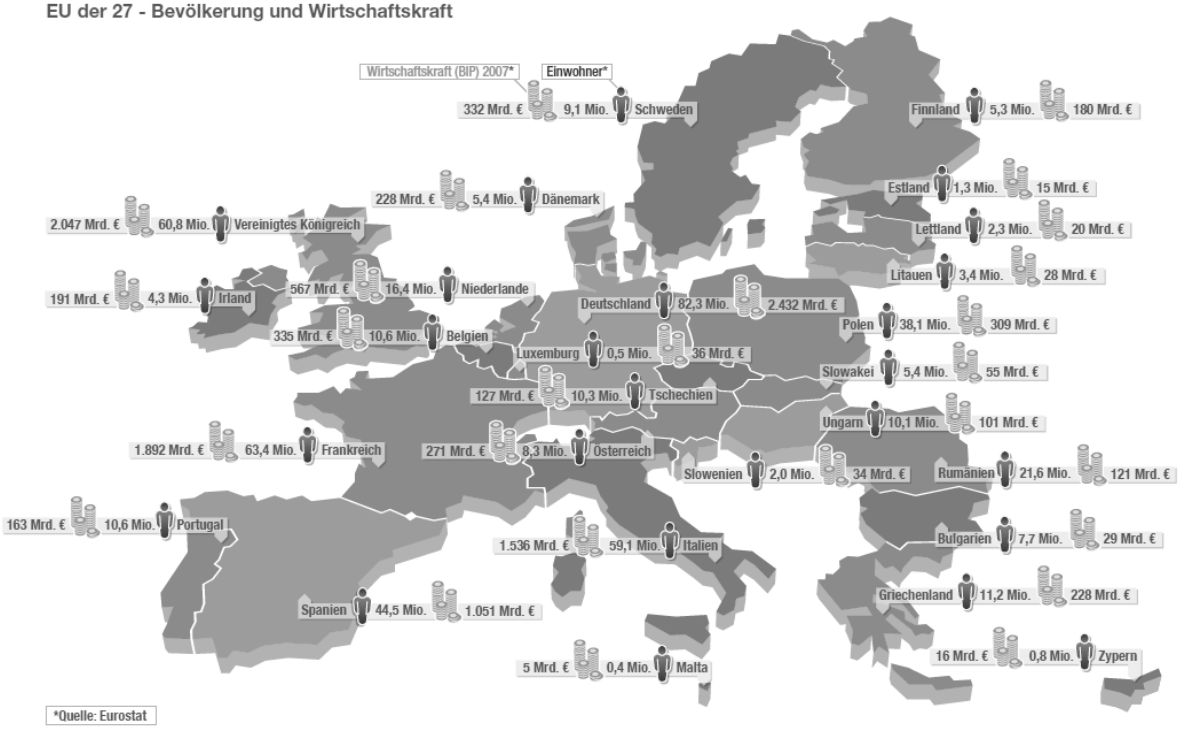


Abb. I.12a: Grafische Darstellung mit Bildsymbolen

Quelle: transcontact Verlagsgesellschaft, Bonn, für die KKB Bank AG, 1989, S. 31, Bundeszentrale für politische Bildung (2009)

Abb. I.12b: Grafische Darstellung mit Bildsymbolen
Quelle: transcontact Verlagsgesellschaft, Bonn, für die KKB Bank AG, 1989, S. 42-43

9.2 Tabellenaufbau und grafische Darstellung

Systematische und übersichtliche Zusammenstellung von Daten mit
- ausreichend informierender Überschrift
- möglichst Zwischensummen
- kein leeres Tabellenfeld (Information einfügen, wie z.B. nicht vorhanden, not available oder keine ausreichende Besetzungszahl)

Normblatt DIN 55301:

Überschrift (Titel und wichtige Angaben)

Vorspalte zum Kopf / Kopf zur Vorspalte	Tabellenkopf				
Vorspalte					
		Fach			

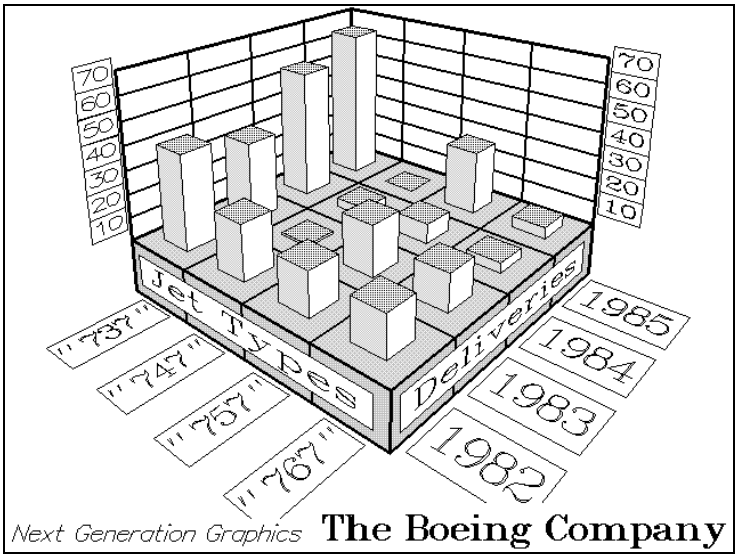
(Erläuterungen)

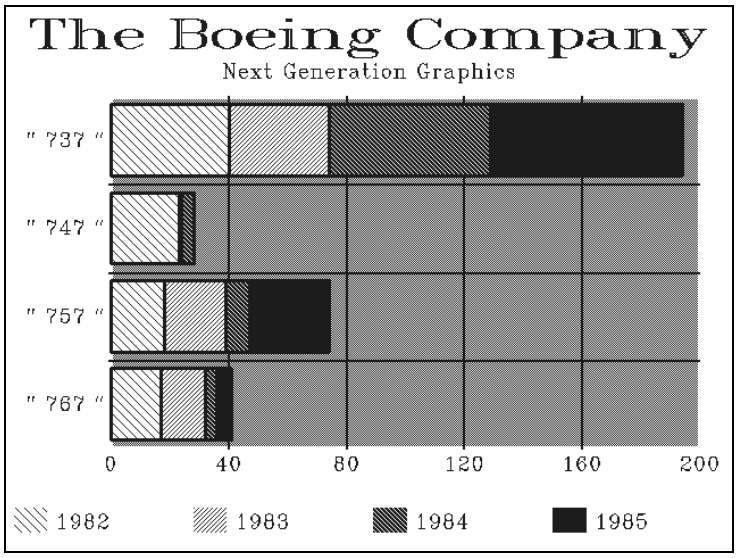
Grafische Darstellung:

- Stab- oder Säulendiagramm, Balkendiagramm (für nominal und ordinal skalierte Daten)
- Flächendiagramm (flächenproportionale Darstellung)
- Kreisdiagramm ('Pie charts', Tortendiagramm)
- Piktogramm (mit Bildsymbolen)
- Kartogramm (innerhalb einer Landkarte)
- Kurvendiagramm
- Histogramm (Häufigkeiten eines klassifizierten Merkmals)
- Polygonzug (Verbindungslinie der Mittelpunkte der Oberkanten des Histogramms)

Beispiele:

- Schwarze (1990), S. 50ff.
- Grafiken aus der Vorlesung
- 2 D- und 3 D-Grafiken aus Computerprogrammen (Chart, Boeing-Graph, Harvard Graphics, EXCEL etc., vgl. Abb. I.13)





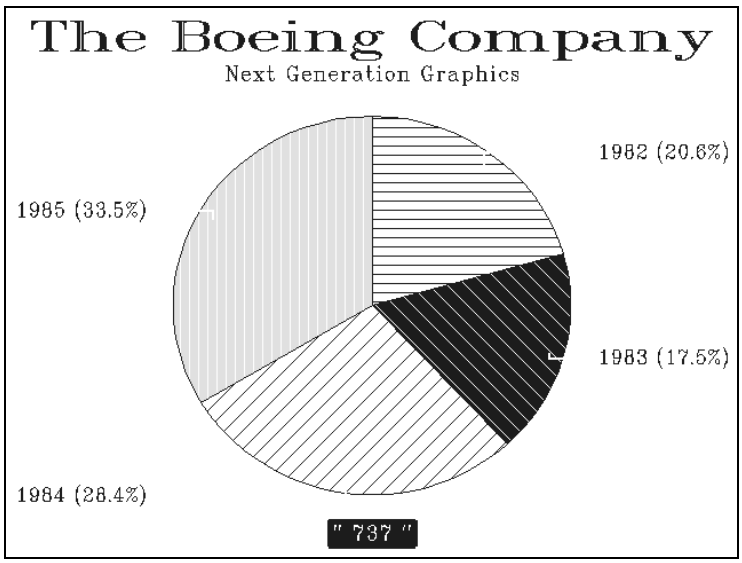


Abb. I.13: Boeing-Graph: 3 D/2 D-Darstellung

10 Datenschutz und Datensicherheit

Datenschutz ist die Aufgabe, Daten vor Mißbrauch u.a. bei der Datenverarbeitung zu bewahren und damit der Beeinträchtigung schutzwürdiger Belange der Betroffenen entgegenzuwirken.

Diskussion besonders im Umfeld des Volkszählungsgesetzes ('gläserner' Mensch, Orwell`s 1984, 'big brother')

Berufskodex für Statistiker (Internationales Statistisches Institut ISI, 1986, S.238, zitiert nach Rinne (1994), S. 32):

> *Statistical data are unconcerned with individual identies. They are collected to answer questions such as "how many?" or "what proportional?", not "who?". The identities and records of co-operating (or non-cooperating) subjects should therefore be kept confidential, whether or not confidentiality has been explicitly pledged.*
>
> *Statisticians should take appropriate measures to prevent their data from being published or otherwise released in a form that would allow any subject´s identity to be disclosed or inferred.*

Grundgesetz:

- Persönlichkeitsrecht nach Grundgesetz (GG) Art. 2, Abs. 1
  (Individuelles Persönlichkeitsrecht auf Achtung seiner Würde und Eigenwertes abgeleitet aus GG Art. 1, Abs. 2)
- Recht auf Informationsfreiheit nach GG Art. 5, Abs. 1, S. 1 ('jeder hat das Recht, ...sich aus allgemein zugänglichen Quellen ungehindert zu unterrichten.')

Grundlegende gesetzliche Regelungen:

- Bundesstatistikgesetz (BStatG) 1987 (Schutz von Daten aus statistischen Erhebungen)
- Bundesdatenschutzgesetz (BDSG) vom 20.12.1990
- Landesdatenschutzgesetze

Datenschutzbeauftragte des Bundes, der Länder, von Institutionen

Datensicherheit

Organisatorische und technische Aufgabe zur Gewährleistung der Sicherheit von Datenbeständen und Datenverarbeitungsabläufen (Datenzugriff nur für Berechtigte, unverfälschte Verarbeitung der Daten etc.)

Weitere Informationen zu Datenschutz und Datensicherheit: Rinne (1994), Kap. 2.3
und im Internet unter:

www.gesetze-im-internet.de/bdsg_1990/index.html

Keyconcepts

Begriff und Aufgaben der Statistik
Träger der Wirtschaftsstatistiken
Adäquationsproblem
Sachgerechte Interpretation
Statistische Massen und Einheiten
Merkmale, Merkmalsausprägungen
Meßskalen
Vorgehensweise bei statistischen Untersuchungen
Tabellarische, grafische Darstellung

II Statistische Analyse eines einzelnen Merkmals

Informationskomprimierung mit Hilfe von Häufigkeitsverteilungen und Kennzahlen zur Lage, Streuung und Konzentration eines einzelnen Merkmals.

Die statistische Analyse *eines* Merkmals konzentriert sich auf die Analyse/Beschreibung *einer* Dimension von Merkmalsträgern (z.B. Einkommen von Haushalten, Umsätze mehrerer Geschäftsvorgänge/Filialen/Jahre)

Komprimierende Beschreibung über:
- Häufigkeitsverteilung
- Lageparameter
- Streuungsmaße bzw. -parameter
- Konzentration der Verteilung

1 Eindimensionale Häufigkeitsverteilungen und ihre Darstellung

Von Interesse:
- ein Merkmal einer statistischen Masse
- beobachtete Häufigkeiten der Merkmalsausprägungen

Darstellungsformen: Tabellen, Grafiken

Es ist zu unterscheiden zwischen qualitativen und quantitativen Größen.

1.1 Häufigkeitsverteilung nominalskalierter (qualitativer) Merkmale

Statistische Masse mit

n	statistische Einheiten
A	qualitatives Merkmal
A_i	i-te Merkmalsausprägung (i = 1,...,k) mit
$n(A_i) = n_i$	absolute Häufigkeit des Merkmals in der Klasse i

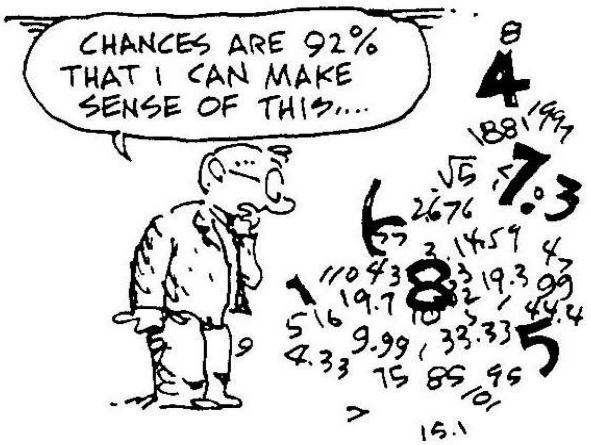
Absolute Häufigkeit

Anzahl der statistischen Einheiten der Klasse i: $\mathrm{n}(\mathrm{A}_\mathrm{i}) = n_i$

Es gilt: $\sum_{\mathrm{i}=1}^{\mathrm{k}} \mathrm{n}(\mathrm{A}_\mathrm{i}) = \sum_{i=1}^{k} n_i = n_1 + n_2 + ... + n_k = n$

Für Vergleichszwecke geeignet:

DATA ARE THE STATISTICIAN'S RAW MATERIAL, THE NUMBERS WE USE TO INTERPRET REALITY. ALL STATISTICAL PROBLEMS INVOLVE EITHER THE COLLECTION, DESCRIPTION, AND ANALYSIS OF DATA, OR *THINKING* ABOUT THE COLLECTION, DESCRIPTION, AND ANALYSIS OF DATA.



THIS CHAPTER CONCENTRATES ON DATA *DESCRIPTION*. HOW CAN WE REPRESENT DATA IN USEFUL WAYS? HOW CAN WE SEE UNDERLYING PATTERNS IN A HEAP OF NAKED NUMBERS? HOW CAN WE SUMMARIZE THE DATA'S BASIC SHAPE?

WELL, TO DESCRIBE DATA, THE FIRST THING YOU NEED IS SOME ACTUAL DATA TO DESCRIBE... SO LET'S COLLECT SOME DATA!

Gonick, Smith (1993)

Relative Häufigkeit

$$h(A_i) = h_i = \frac{n_i}{n} = \frac{n(A_i)}{n} \qquad h_i = \frac{\text{Absolute Häufigkeit der Klasse i}}{\text{Umfang der statistischen Masse}}$$

Es gilt: $\sum_{i=1}^{k} h(A_i) = \sum_{i=1}^{k} h_i = 1$

Häufigkeitstabelle: Tabellarische Darstellung

Tab. III.1: Häufigkeitstabelle: Allgemeiner Aufbau

Merkmalsausprägung	Absolute Häufigkeit	relative Häufigkeit
A_1	n_1	h_1
.	.	.
.	.	.
A_i	n_i	h_i
.	.	.
.	.	.
A_k	N_k	h_k
Summe	N	1

Beispiel:

Erwerbstätige untergliedert nach der Stellung im Beruf

Tab. III.2: Erwerbstätige nach der Stellung im Beruf in der BRD 1987 (in 100.000)

Ausprägung A_i des Merkmals 'Stellung im Beruf'		Absolute Häufigkeit $n(A_i) = n_i$	Relative Häufigkeit $h(A_i) = \frac{n_i}{n}$	bzw. [%]
Selbständige	(A_1)	23	0,085	8,5
Mithelfende Familien-angehörige	(A_2)	5	0,018	1,8
Beamte	(A_3)	24	0,091	9,1
Angestellte/Auszubildende (kfm./techn.)	(A_4)	110	0,410	41,0
Arbeiter/Auszubildende (gewerblich)	(A_5)	107	0,396	39,6
Insgesamt		269 $\left(\sum_{i=1}^{5} n_i\right)$	1,000	100,0

Quelle: Statistisches Bundesamt 1989, Volkszählung 1987

In der BRD waren 1987 somit von den insgesamt 269 · (100.000) = 26,9 Mio. Erwerbstätigen 8,5 % Selbständige und 39,6 % (10,7 Mio.) Arbeiter.

Häufigkeitsverteilung

Die **Häufigkeitsverteilung qualitativer Merkmale** heißt die Funktion $h(A_i)$, die jeder Merkmalsausprägung A_i den Anteil der statistischen Einheiten mit dieser Merkmalsausprägung (relative Häufigkeit)

$$h(A_i) = \frac{n(A_i)}{n} = \frac{n_i}{n} \qquad (i = 1, ..., k)$$

zuordnet. Die Häufigkeitsverteilung ist also die Gesamtheit der relativen Häufigkeiten.

Grafische Darstellung

- Kreisdiagramm ('pie-chart'): Die Kreisfläche wird in entsprechende Anteile aufgeteilt.
  Hinweis zur Berechnung der Anteile an der Kreisfläche: $\sum_{i=1}^{k} h(A_i) = 100\ \% = 360^\circ$.
- Balkendiagramme

Beispiele:

- **Kreisdiagramm**

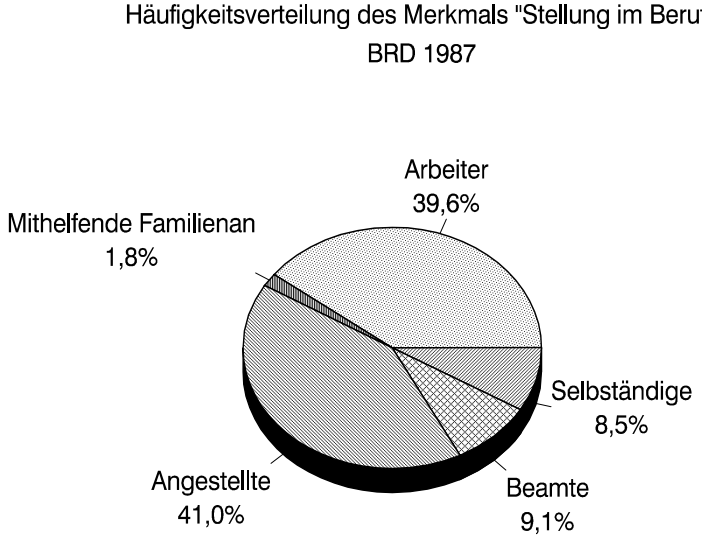


Abb. III.1: Häufigkeitsverteilung des Merkmals 'Stellung im Beruf' in der BRD 1987
Quelle: Statistisches Bundesamt 1989, Volkszählung 1987

- **Balkendiagramm** ('bar-chart'): Balken proportional zu den entsprechenden Einheiten

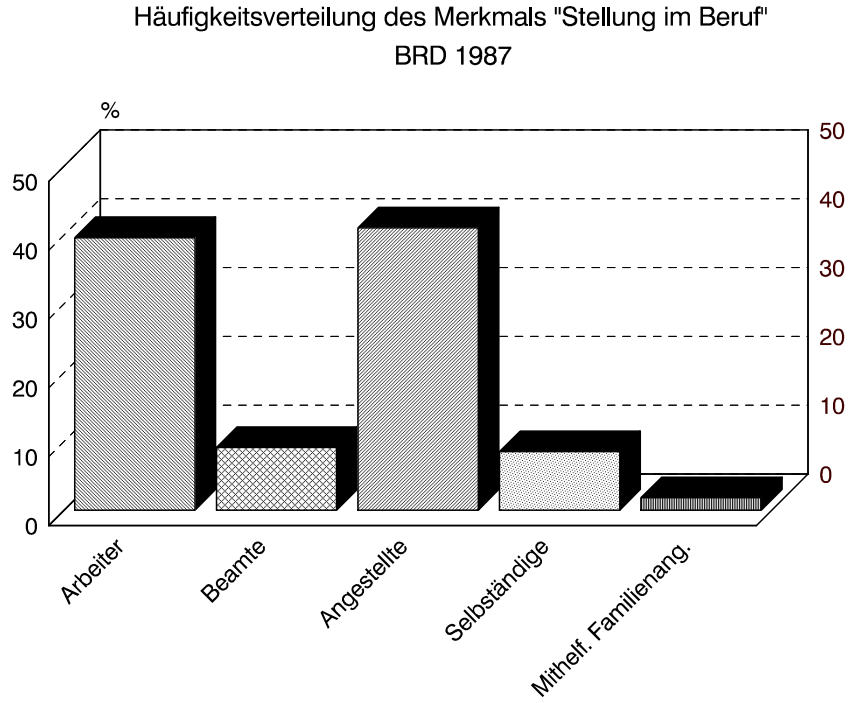
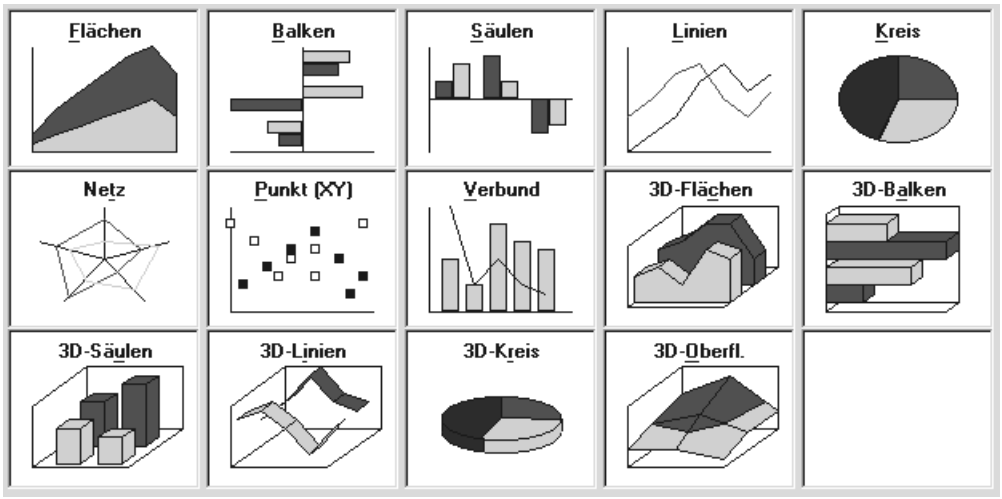


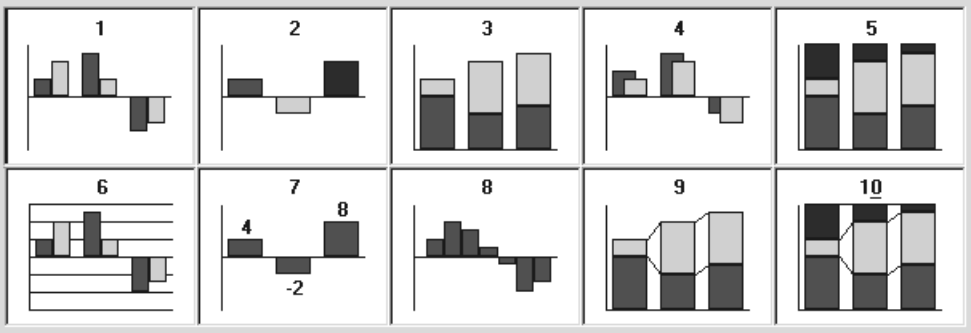
Abb. III.2: Häufigkeitsverteilung des Merkmals 'Stellung im Beruf' in der BRD 1987 Quelle: Statistisches Bundesamt 1989, Volkszählung 1987

- **Grafische Darstellungsmöglichkeiten in einem Tabellenkalkulationsprogramm z.B. Microsoft Excel**

Hauptmenü:



Untermenü Säulen:



1 Einfaches Säulendiagramm
2 Säulendiagramm für eine Datenreihe mit unterschiedlichen Mustern
3 Gestapelt
4 Überlappend
5 100 % gestapelt
6 Mit horizontalen Gitternetzlinien
7 Mit Wertebeschriftungen
8 Stufendiagramm (Rubriken ohne Zwischenraum)
9 Gestapelt mit Linien, die die Daten in derselben Datenreihe verbinden
10 100 % gestapelt mit Linien, die die Daten in derselben Datenreihe verbinden

Abb. III.3: Grafische Darstellungsmöglichkeiten in dem Tabellenkalkulationsprogramm Microsoft Excel

- **ET, Econometrics Toolkit**
  z.B.: 20 Studentinnen/Studenten werden nach der Augenfarbe befragt.
  Jede Beobachtung (= Proband), jede statistische Einheit erhält für das Merkmal Augenfarbe eine Merkmalsausprägung (Code):

  Codes: 1 = blaue Augen
  2 = grüne Augen
  3 = rote Augen
  4 = gelbe Augen

  ET: 1 Data entry and manipulation (Main Menu)
  - Data (Read or edit data)
    variable name (eyes) ↵
    input of data

  5 Histograms, plots, descriptive statistics
  - Histogram for individual or frequencies
    variable name (eyes) ↵

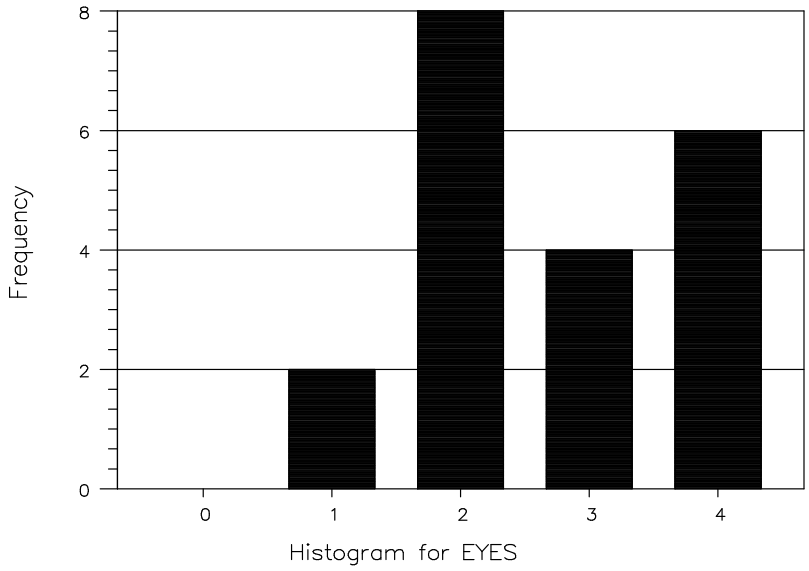
```
                Main Menu
 1 Data entry and mani
 2 Current sample and
 3 Management and disp
 4 File system and out
 5 Histograms, plots,
 6 Regression model es
 7 Probability,Matrice
 8 Tests: t, F, fits,
 9 Editor for text and
 System: Give DOS comm
 Graphics: set screen
 QUIT: leave ET (Save
Data: Rows= 200 (   1
Cols=100,Observations
```

```
Histograms, plots, desc. stats

Command  Keys Function                  Page 5
Describe D  1 Descriptive statistics
Histogram H  2 Histogram for ind. or freq
Scatter  P  3 Plot variables in scatter diag.
Identify B  4 Box-Jenkins time series ID
ARIMA    A  5 ARIMA and ARMAX time series
Stepwise S  6 Stepwise linear regression
Crosstab C  7 Cross tabulation for 2 vars.
XCorrel  X  8 Cross correlation, time series
Give command by letter or #, or ↑/↓ and ↵.
Use PgUp and PgDn for other command groups.
Press ESC to clear and return to main menu.
```

```
F1=HELP   F2=Variables  F3=Namelists     F4=Matrices    F5=Scalars   F6=List all
F7=Output(Full/Basic)B  F8=Option MenusN F9=Color/Mono  F10=Mode(Menu/Command) M
```

```
DATA LISTING (Current sample)     ↵/ESC            Press ESC to interrupt list.
Observation      EYES
      1        4.0000
      2        2.0000
      3        3.0000
      4        1.0000
      5        3.0000
      6        3.0000
      7        2.0000
      8        2.0000
      9        2.0000
     10        4.0000
     11        4.0000
     12        2.0000
     13        4.0000
     14        1.0000
     15        2.0000
     16        2.0000
     17        4.0000
     18        2.0000
     19        3.0000
     20        4.0000
```

```
     Histogram for EYES     computed using   20 observations
     Obs. out of range: too low=   0, too high=   0
     Individual data      Mean=    2.700, std.dev.=    1.031
                                    Frequency          Cumulative
    Lower Limit    Upper Limit    Total   Relative   Total   Relative
0        -.500           .500       0      .0000       0      .0000
1         .500          1.500       2      .1000       2      .1000
2        1.500          2.500       8      .4000      10      .5000
3        2.500          3.500       4      .2000      14      .7000
4        3.500          4.500       6      .3000      20     1.0000
```



- **Verbale Häufigkeitsdarstellung:**

  "Wenn wir für eine Minute schweigen sollten für jeden Menschen, der 1982 an Hunger starb, wären wir nicht in der Lage, den Beginn des 21. Jahrhunderts zu feiern, weil wir dann immer noch still sein müßten."

  Kubas Staatspräsident Fidel Castro, 1983

1.2 Häufigkeitsverteilung metrisch skalierter, diskreter Merkmale

Ein metrisch skaliertes (quantitatives), diskretes Merkmal nimmt nur bestimmte Zahlenwerte (aus den reellen Zahlen) an: meist nichtnegative ganze Zahlen: 0,1,2,...; z.B.:

- Anzahl der Personen in einem Haushalt
- Anzahl der Verkäufe eines bestimmten Produktes
- Anzahl der Räume in privaten Wohnungen

x_i Merkmalswert für die i-te Merkmalsausprägung des Merkmals x

Wie für nominalskalierte Größen erhält man durch Auszählen der jeweiligen statistischen Einheiten die:

Absolute Häufigkeit

Anzahl der statistischen Einheiten mit dem Merkmalswert $x_i : n(x_i) = n_i$

Relative Häufigkeit

$$h(x_i) = \frac{n_i}{n} = \frac{n(x_i)}{n}$$

$h(x_i)$ kann als diskrete Funktion aufgefaßt werden:

Häufigkeitsfunktion $h(x) = h(x_i)$ *mit* $i = 1,...,k$ *Merkmalsausprägungen*

Bei metrisch skalierten Merkmalen mißt die Differenz zweier Merkmalswerte ihren Abstand. Daher: Berechnung von Anteilen (relative Häufigkeit) für mehrere Merkmalswerte.

Häufig: Berechnung des Anteilwertes für Merkmalswert kleiner oder gleich x_i: kumulierte Häufigkeiten

Kumulierte absolute Häufigkeit

$$n(x \le x_i) = n_1 + n_2 + ... + n_i = \sum_{j=1}^{i} n_j$$

Kumulierte relative Häufigkeit

$$h(x \le x_i) = h(x_1) + h(x_2) + ... + h(x_i) = \sum_{j=1}^{i} h(x_j)$$

Verteilungsfunktion

Die Verteilungsfunktion ist die Funktion $F(x_i)$, die jedem Merkmalswert x_i den Anteilswert aller statistischen Einheiten zuordnet, die einen Merkmalswert kleiner oder gleich x_i $(x \le x_i)$ haben.

$$F(x_i) = h(x \le x_i) = \sum_{j=1}^{i} h(x_j) = \sum_{j=1}^{i} \frac{n_j}{n}$$

Beispiel:

Größe der Privathaushalte in der BRD 1987

Tabellarische Darstellung: Häufigkeitstabelle

Tab. III.3: Größe der Privathaushalte in der BRD 1987 (in 100.000)

Anzahl der Personen x_i	Absolute Häufigkeit $n(x_i) = n_i$	Relative Häufigkeit $h(x_i) = \frac{n_i}{n}$	Verteilungs-funktion $F(x_i)$
$x_1 = 1$	88	0,34	0,34
$x_2 = 2$	74	0,28	0,62
$x_3 = 3$	46	0,18	0,80
$x_4 = 4$ und mehr	54	0,20	1,00
Insgesamt	262	1,00	

Quelle: Statistisches Bundesamt 1989, Volkszählung 1987

Aus der Häufigkeitsverteilung ergibt sich z.B., dass 18 % aller Privathaushalte aus drei Personen bestehen.

Aus der Verteilungsfunktion ergibt sich z.B., dass 80 % aller Privathaushalte drei oder weniger Personen hatten.

Grafische Darstellung

Häufigkeitsfunktion: als Stabdiagramm (es gibt keine Zwischenwerte); Stablänge = relative Häufigkeit

Verteilungsfunktion: als Treppenfunktion (Summe der bisherigen Stablängen)

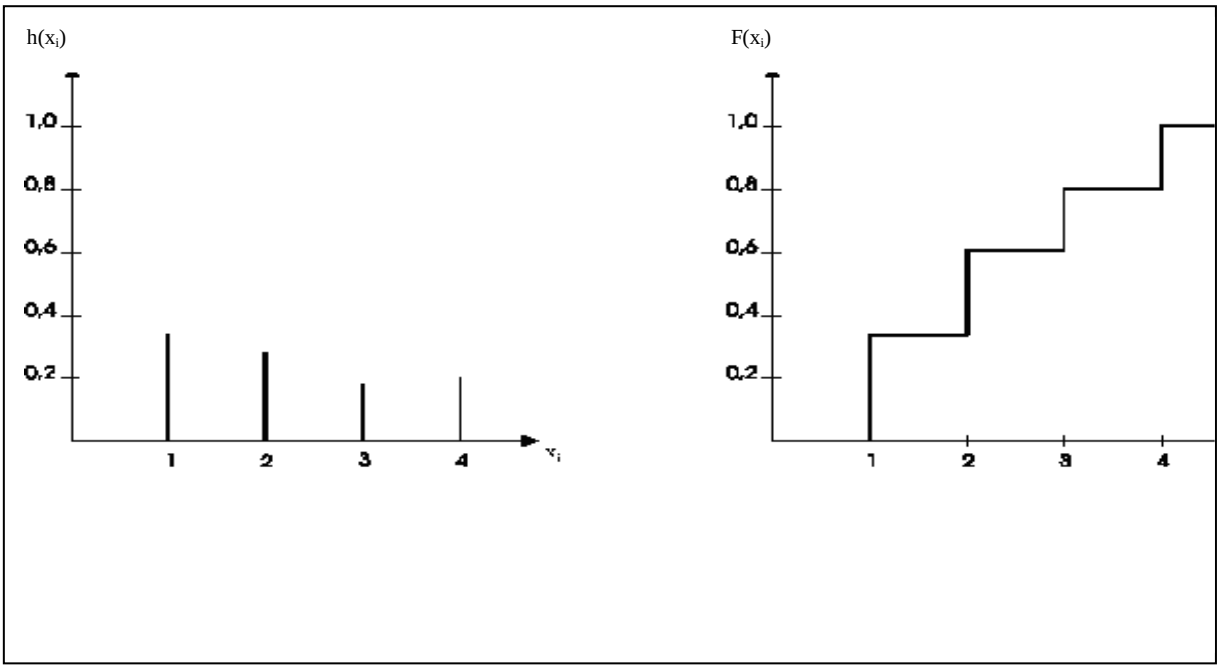


Abb. III.4: Häufigkeitsfunktion und Verteilungsfunktion: Grafische Darstellung der Größe der Privathaushalte in der BRD 1987 (vgl. Tab. III.3)

Wie groß ist der Anteil der Privathaushalte mit mehr als 1 aber weniger als 4 Personen?

$$\begin{aligned} h(1<x<4) &= h(x\le 3)-h(x=1) \\ &= F(3)-F(1) \\ &= 0,80-0,34=0,46 \end{aligned}$$

46 % aller Haushalte sind 2- oder 3-Personenhaushalte.

1.3 Häufigkeitsverteilung metrisch skalierter (quantitativer) stetiger Merkmale

Merkmalswerte können sich in allen reellen Zahlen ausprägen. Da jeder Merkmalswert in der Regel nur einmal beobachtet wird (bei beliebig genauer Messung) ist zur Darstellung erst eine **Klassenbildung** notwendig.

Klasseneinteilung
- Möglichst gleiche Klassenbreiten, bei großem Variationsbereich der Daten auch unterschiedliche Klassenbreiten verwenden;
- Anzahl der Klassen (k) nicht zu groß $\left(\sqrt[3]{n},...,\sqrt[2]{n}\,(n=100,5-10\,Klassen)\right)$;
- Der häufigste Wert der Urliste sollte die Klassenmitte der Klasse mit der größten Häufigkeit bilden;
- Für einen Vergleich mit anderen Verteilungen: gleiche Klassen bilden

Generelles Ziel: Struktur des Ausgangsmaterials klar und unverfälscht herausarbeiten!

Nach Klasseneinteilung: Auszählen der Merkmalswerte je Klasse ergibt Häufigkeiten für die einzelnen Klassen

Häufigkeitsverteilung metrisch skalierter, stetiger Merkmale
ist die Funktion

$$h(x_i)=h_i=h\left(x_i^u\le x<x_i^o\right),$$

die jeder Klasse i $(i=1,...,k)$ eine relative Häufigkeit $\frac{n_i}{n}$ zuordnet, wobei

x_i^u Untergrenze der Klasse i, x_i^o Obergrenze der Klasse i

n_i Zahl der beobachteten Merkmalswerte im Intervall $\left[x_i^u, x_i^o\right]$

Die klassierte Häufigkeitsverteilung wird mit Rechtecken als **Histogramm** grafisch dargestellt. Damit die Rechteckflächen auch bei unterschiedlichen Klassenbreiten proportional den (relativen) Häufigkeiten sind, werden die relativen Häufigkeiten über die Dichtefunktion (= Höhe der Rechtecke) normiert.

Dichtefunktion
normierte relative Häufigkeiten (vor allem für unterschiedliche Klassenbreiten)

$f(x_i) = \frac{n_i}{n \cdot \Delta x_i}$, wobei Δx_i = Breite der i-ten Klasse

Aus Rechteckfläche: Höhe $f(x_i) \cdot$ Breite Δx_i =Anteilswert

also $f(x_i) \cdot \Delta x_i = \frac{n_i}{n} \Rightarrow f(x_i) = \frac{n_i}{n \cdot \Delta x_i}$ Häufigkeitsdichte

Beispiele:

Monatliches Haushaltsnettoeinkommen der Haushalte in der BRD im Jahre 2009

Tabellarische Darstellung

Tab. III.4: Monatl. Haushaltsnettoeinkommen der Haushalte in Deutschland im Jahre 2009

Einkommens-klasse $x_i^u \le x < x_i^o$ (in EUR)	Klassen-breite	absolute Häufigkeit n_i (=Anz. HH in 1000)	relative Häufigkeit $h_i = h(x_i^u \le x < x_i^o)$ $h_i = \frac{n_i}{n}$	Verteilungs-funktion. in den Klassen-obergrenzen $F(x_i^o)$	Dichte-funktion $f(x_i) = \frac{n_i}{n \cdot \Delta x_i}$
unter 500	500	1.400	0,037	0,037	7,46468E-05
500 - unter 1000	500	6.200	0,165	0,203	3,30579E-04
1000 - unter 1500	500	7.500	0,200	0,403	3,99893E-04
1500 - unter 2000	500	7.200	0,192	0,595	3,83898E-04
2000 - unter 2500	500	4.800	0,128	0,722	2,55932E-04
2500 -unter 3000	500	3.800	0,101	0,824	2,02613E-04
3000 - unter 3500	500	2.400	0,064	0,888	1,27966E-04
3500 - unter 4000	500	1.700	0,045	0,933	9,06425E-05
4000 - unter 4500	500	860	0,023	0,956	4,58544E-05
4500 - unter 5000	500	740	0,020	0,976	3,94561E-05
5000 - unter 5500	500	270	0,007	0,983	1,43962E-05
5500 - unter 7000	1500	640	0,017	1,000	1,13747E-05
		37.510	1,000		

Haushalte mit einem Einkommen über EUR 7.000 nicht in der Tabelle erfasst, ihr Anteil beträgt 1,8 % der Stichprobe (n = 9708).

Quelle: Sozio-ökonomisches Panel (Welle Z (26)), 2009), eigene Berechnungen

Grafische Darstellung als Histogramm mit unterschiedlichen Klassenbreiten:

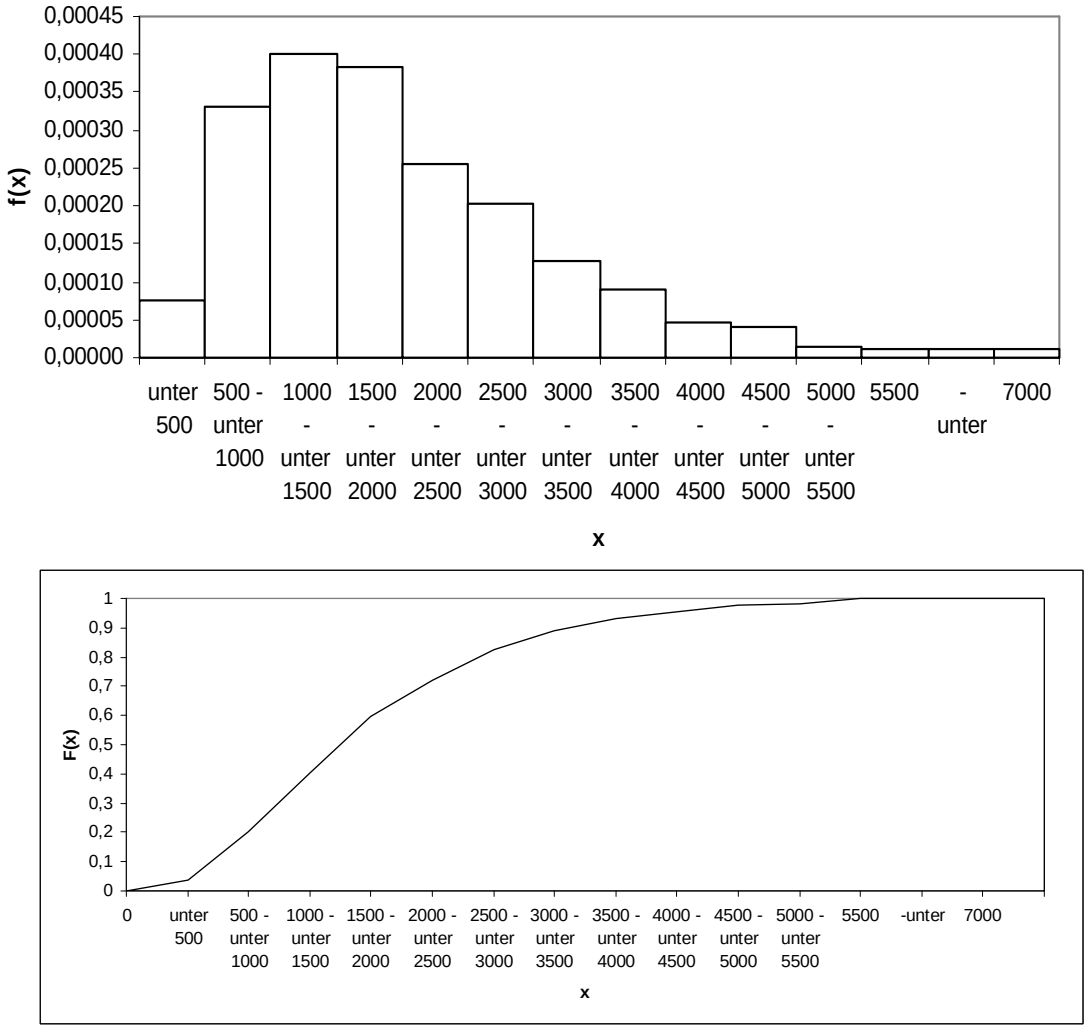


Abb. III.6: Histogramm und Verteilungsfunktion des monatlichen Haushaltsnettoeinkommens der Haushalte in der BRD 2009

Quelle: Sozio-ökonomisches Panel (Welle Z (26), 2009), eigene Berechnungen

Zur Berechnung der Anteilswerte innerhalb einer Klasse

Wieviel Prozent der statistischen Einheiten besitzen einen Merkmalswert kleiner oder gleich x?

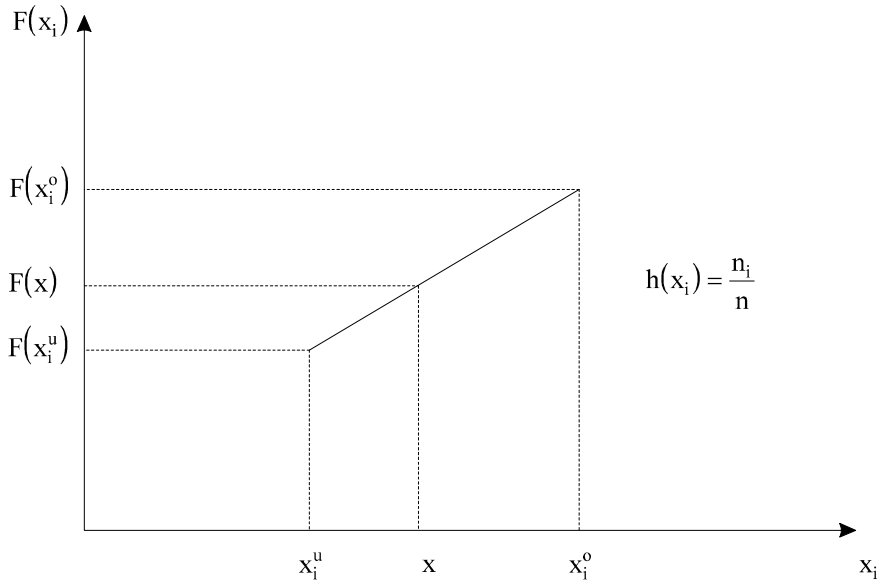
Durch lineare Interpolation (Annahme: Gleichverteilung in der Klasse):

Es gilt:

$$\frac{F(x)-F\left(x_i^u\right)}{h(x_i)}=\frac{x-x_i^u}{\Delta x_i} \qquad \textit{bzw.} \qquad \frac{F(x)-F\left(x_i^u\right)}{x-x_i^u}=\frac{h(x_i)}{\Delta x_i}$$

Daraus folgt:

$$F(x)=F\left(x_i^u\right)+\frac{x-x_i^u}{\Delta x_i}\cdot h(x_i)$$



Beispiel:

Monatliches Haushaltsnettoeinkommen (Tabelle III.4)

$$F\left(x=1800\right)=0,403+\frac{1800-1500}{500}\cdot 0,192=0,518$$

Etwa 52 % der Haushalte haben weniger als 1.800 EUR Nettoeinkommen pro Monat.

1.4 Computergestützte grafische Darstellung

PC-Programme wie
- WORD
- EXCEL
- CHART
- 3D-Boeing Graph, Sunrise
- Harvard Graphics
- in Verbindung mit Analyseprogrammen:
  SPSS-PC, SAS-PC, ET, LIMDEP, GAUSS, MICRO-TSP, SYSTAT, STATA...

Beispiele: siehe PC-Vorführung in der Vorlesung

Gonick, Smith (1993)

2 Lageparameter

Häufigkeitsverteilungen waren eine erste Stufe der Informationsverdichtung. Weitergehende und stärkere Verdichtung ist durch Maßzahlen möglich wie:

Lageparameter: verschiedene Mittelwerte
Streuungsmaße: Varianzen etc.
Konzentration: Gini, Lorenzkurve

2.1 Häufigster Wert (Modus)

Modus oder Modalwert = Wert, der am häufigsten vorkommt.

Der **Modalwert D** eines metrisch skalierten diskreten Merkmals ist derjenige Merkmalswert x, für den die relative Häufigkeit $h(x)$ ihr Maximum annimmt.

Die Klasse mit der größten Häufigkeitsdichte $f(x_i) = \frac{n_i}{n \cdot \Delta x_i}$ heißt **modale Klasse,** ihre Klassenmitte definiert man als Modalwert D (metrisch skaliert, stetig).

Der Modalwert ist nur dann aussagekräftig, wenn die Verteilung eingipflig (unimodal) ist. Bei einer mehrgipfligen (u-förmig, etc.) Verteilung ist er wenig sinnvoll.

Beispiele:

- Privathaushalte in der BRD 2010
  häufigster Haushaltstyp: 1-Personenhaushalt, Modus D = 1 (diskretes Merkmal)
- Einkommensverteilung in der BRD 1992
  Modus D = 2500,- (größte Dichte $f(x)$, stetiges Merkmal)

2.2 Median (Zentralwert)

Der **Median oder Zentralwert** halbiert das Datenmaterial, d.h. 50 % aller Einheiten liegen oberhalb und 50 % aller Einheiten liegen unterhalb dieses Wertes (Median = '50-Prozentpunkt')

Dazu ist es notwendig, die Einheiten nach der Größe ihrer Merkmalswerte zu ordnen.

Ungruppiertes Material

Gegeben sind: n beliebige Merkmalswerte $x_1, x_2, ..., x_n$;
geordnet nach Größe: $x_{(1)} \le x_{(2)} \le ... \le x_{(n)}$;

$x_{(i)}$: i-ter Merkmalswert der geordneten Reihe

Als **Median Z** wird definiert

$$Z = x_{\left(\frac{n+1}{2}\right)} \quad , \text{ falls n ungerade}$$

$$Z = \frac{1}{2}\left[x_{\left(\frac{n}{2}\right)} + x_{\left(\frac{n}{2}+1\right)} \right] \quad , \text{ falls n gerade}$$

Beispiel:

Monatsgehälter in der Fa. DALLES & CO.

Männer: 1650, 2030, 1840, 1520, 1670; n = 5
Frauen: 1710, 1960, 2570, 1490 ; n = 4

Geordnete, sortierte Werte:
Männer: 1520, 1650, 1670, 1840, 2030
Frauen: 1490, 1710, 1960, 2570
M+F: 1490, 1520, 1650, 1670, 1710, 1840, 1960, 2030, 2570

$$Z_M = x_{\left(\frac{n+1}{2}\right)} = x_{(3)} = 1670 \qquad (\text{n ungerade})$$

$$Z_F = \frac{1}{2}\left[x_{\left(\frac{n}{2}\right)} + x_{\left(\frac{n}{2}+1\right)} \right] = \frac{1}{2}\left(x_{(2)} + x_{(3)} \right) = \frac{1}{2}(1710+1960) = 1835 \qquad (n \text{ gerade})$$

$$Z_{M+F} = x_{\left(\frac{n+1}{2}\right)} = 1710$$

Im allgemeinen kann man den Median einer zusammengefaßten Grundgesamtheit **nicht** aus den Medianen der Teilgesamtheiten berechnen!

Grafische Darstellung als 'Box and Whisker'-Plots

'Box and Whisker'-Plots beschreiben eine Verteilung mit dem Median und dem Bereich um den Median, in dem 50 % aller (geordneten) Werte liegen. Damit liegen 25 % unterhalb und 25 % aller Daten oberhalb der Box.

Beispiel:

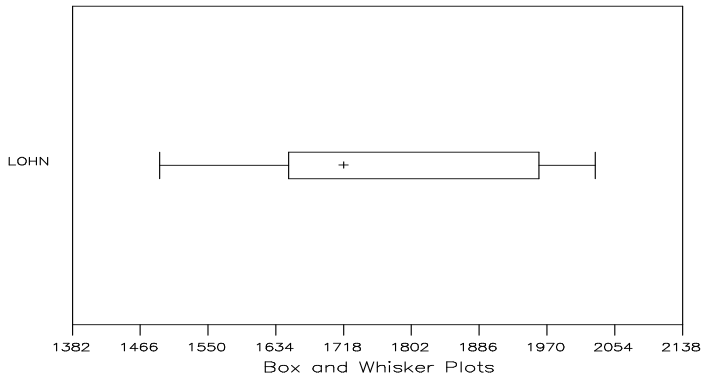


Abb. III.7: ET-'Box and Whisker'-Plot der Monatsgehälter der Fa. BRUCH, DALLES & Co.

Interpretation:
50 % der Beobachtungen liegen zwischen 1650 und 1960; der Median liegt nicht im Zentrum, damit sind die Daten nicht symmetrisch verteilt.
Vertikale Striche: kleinster ($\pm 1{,}5 \cdot$Boxbreite) bzw. größter Wert

---

Gruppiertes Material

Falls das Urmaterial nur gruppiert vorliegt, erhält man den Median **nur approximativ mit Hilfe der Verteilungsfunktion** $F(x)$.

Da 50 % der Merkmalswerte einen kleineren Merkmalswert als den Median Z haben, gilt:

$$h(x \le Z) = F(Z) = 0{,}5$$

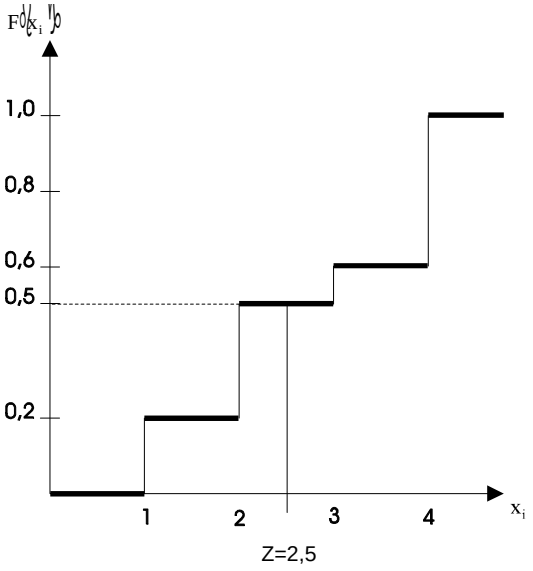
Lineare Interpolation

$$Z = x_i^u + \frac{F(z) - F(x_i^u)}{f(x_i)} = x_i^u + \frac{F(z) - F(x_i^u)}{n_i / n} \cdot \Delta x_i$$

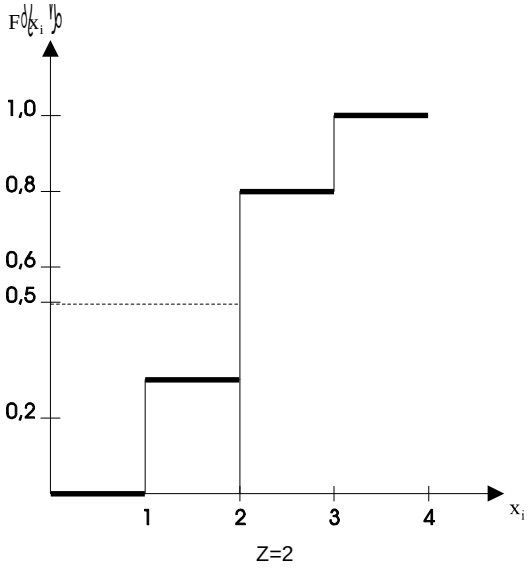
mit x_i^u als Klassenuntergrenze der Klasse mit F(Z) = 0,50
Bei **metrisch skalierten, diskreten Merkmalen erhält man den Median Z** durch Ablesen der Stelle x aus der Treppenfunktion an der Stelle $F(x) = 0{,}5 \quad \Rightarrow \quad Z = x$.

Bestimmung des Medians aus der Verteilungsfunktion bei diskreten Merkmalen
F(Z) = 0,50

a) Falls F(z) auf einer Treppenstufe den Wert 0,5 annimmt, dann ist der Median gleich den Abzissen des Mittelwertes dieser Treppenstufe.
b) Falls F(z) den Wert 0,5 nicht annimmt, ist der Median gleich dem kleinsten Merkmalswert, an dem die Verteilungsfunktion größer als 0,5 ist.



a)



b)

Beispiel:

Privathaushalte $F(2) = 50\ \% \ \Rightarrow \qquad Z = 2$

d.h. ca. 50 % aller Privathaushalte sind 1 oder 2-Personenhaushalte (approximativ).

Z ist hier die Stelle, an der die Verteilungsfunktion F(x) den Wert 0,5 erstmals überschreitet.

Bei **metrisch skalierten, stetigen Merkmalen** halbiert der Median die Fläche des Histogramms. Bestimmung gegebenenfalls durch lineare Interpolation.

Eine Verallgemeinerung des Median (Z = 50 % Quantil) ist das Konzept der p-Quantile, vgl. 3.2.

Beispiel:

Monatliches Haushaltsnettoeinkommen BRD 2009

Der Median liegt innerhalb der Einkommensklasse 1500-2000 EUR.
Lineare Interpolation

$$Z = x_i^u + \frac{F(z) - F\left(x_i^u\right)}{\frac{n_i}{n}} \cdot \Delta x_i$$

hier: $$Z = 1500 + \frac{0,50 - 0,403}{0,192} \cdot 500 = 1.753,82$$

d.h. 50 % aller Haushalte haben in der BRD 2009 weniger als 1.753,82 EUR (approximativ) monatliches Nettoeinkommen verdient.

2.3 Arithmetisches Mittel

Das **arithmetische Mittel** $\bar{x}$ gibt an, welchen Merkmalswert **jede** statistische Einheit haben würde, wenn die gesamte Merkmalssumme gleichmäßig auf alle statistischen Einheiten verteilt wäre (Ersatzwert).

Ungruppiertes Material

Das arithmetische Mittel $(\bar{x})$ ist der Durchschnitt $(\varnothing)$ aus den Merkmalswerten aller statistischen Einheiten.

Gegeben: n beliebige Merkmalswerte $x_1, x_2, \ldots, x_n$

Arithmetisches Mittel

$$\bar{x} = \frac{1}{n}\left(x_1 + x_2 + \ldots + x_n\right) = \frac{1}{n}\sum_{i=1}^{n} x_i$$

Über Umformung ergibt sich die Merkmalssumme: $n \cdot \bar{x} = \sum_{i=1}^{n} x_i$

Beispiel:

a) Durchschnittliche Körpergröße von Studentinnen und Studenten in der Vorlesung Statistik I an der Uni Lüneburg:

$$\bar{x} = \frac{1}{5}\sum_{i=1}^{5} x_i = \frac{1}{5}\left(x_1 + x_2 + \ldots + x_5\right) = \frac{1}{5}\left(172 + 178 + 164 + 167 + 171\right) = \frac{1}{5} \cdot 852 = 170,4\,cm$$

b)

i	1	2	3
x_i	3	4	5

$$\bar{x} = \frac{1}{3}(3+4+5) = \frac{1}{3} \cdot 12 = 4$$

$\rightarrow \bar{x}$ kann, muß aber nicht einen der x_i-Werte annehmen!

Gonick, Smith (1993)

Gruppiertes Material

Nach Zusammenfassung des Datenmaterials in k Größenklassen ergibt sich für jede Klasse ein arithmetisches Mittel $\bar{x}_1, \bar{x}_2, ..., \bar{x}_k$.

Die Berechnung des arithmetischen Mittels des gesamten Datenmaterials (über alle Klassen) ergibt sich als

gewichtetes (gewogenes) arithmetisches Mittel (Additionssatz für Mittelwerte):

$$\bar{x} = \frac{\bar{x}_1 n_1 + \bar{x}_2 n_2 + ... + \bar{x}_k n_k}{n_1 + n_2 + ... + n_k} = \frac{1}{n}\sum_{i=1}^{k} \bar{x}_i n_i = \sum_{i=1}^{k} \bar{x}_i \cdot \frac{n_i}{n} = \sum_{i=1}^{k} \bar{x}_i \cdot h(x_i)$$

Gewichte: relative Häufigkeiten

Bei **unbekanntem Gruppenmittel** werden die Klassenmitten x_i^* anstelle von $\bar{x}_i$ verwendet:

$$\bar{x} \approx \frac{1}{n}\sum_{i=1}^{k} x_i^* \cdot n_i = \sum_{i=1}^{k} x_i^* \cdot h(x_i) \qquad \text{['je gleichverteilter desto besser']}$$

Beispiel:

Klausurnoten Statistik I

Note	n_i	n_i/n	$\bar{x}_i$
1	4	0,4	1
2	2	0,2	2
3	3	0,3	3
4	1	0,1	4
5	-	0,0	
	10	1,0	

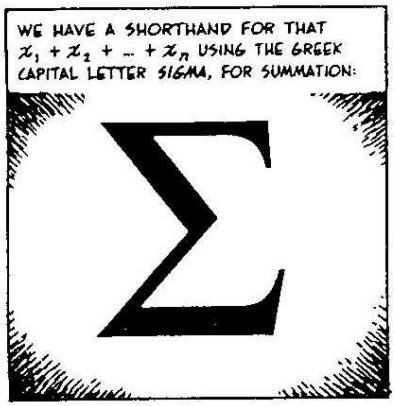
$\bar{x}_i = \text{Note}_i$, z. B.: $\bar{x}_1 = \frac{1}{4}(1+1+1+1) = 1$

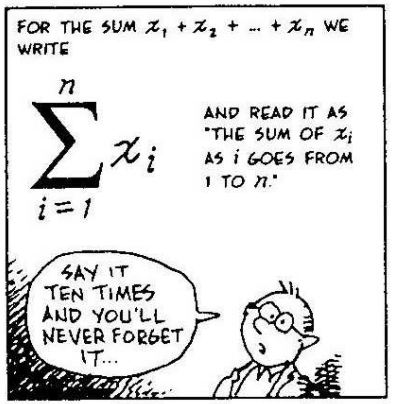
gewogenes arithmetisches Mittel:

$$\bar{x} = \sum_{i=1}^{k} \bar{x}_i \cdot h(x_i) = 1 \cdot 0,4 + 2 \cdot 0,2 + 3 \cdot 0,3 + 4 \cdot 0,1 = 2,1$$

ungruppiertes arithmetisches Mittel:

$$\bar{x} = \frac{1}{10}\sum_{i=1}^{10} x_i = \frac{1}{10}(1+1+1+1+2+2+3+3+3+4) = \frac{1}{10} \cdot 21 = 2,1$$





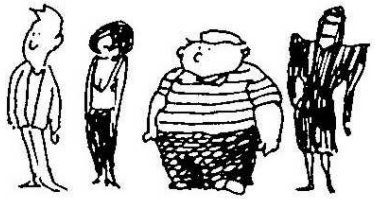
SO... TO REPEAT, THE AVERAGE, OR *MEAN*, OF A SET OF DATA x_i IS

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} \quad \text{OR} \quad \sum_{i=1}^{n} \frac{x_i}{n}$$

IN THE CASE OF OUR 92 PENN STATE STUDENTS, THE MEAN WEIGHT IS

$$\sum_{i=1}^{92} \frac{x_i}{92} = \frac{13,354}{92}$$

$$= 145.15 \text{ POUNDS}$$



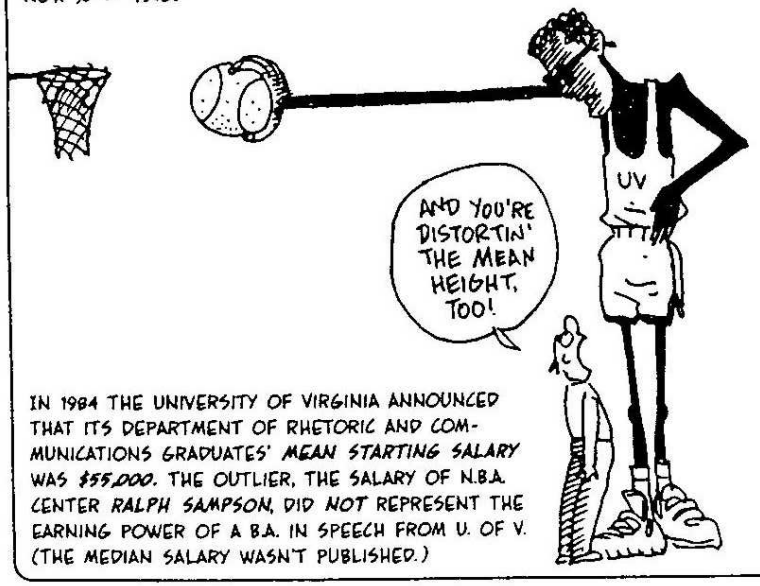
FOR THE $n = 92$ STUDENT WEIGHTS, WE CAN FIND THE MEDIAN FROM THE ORDERED STEM-AND-LEAF DIAGRAM: JUST COUNT TO THE 46TH OBSERVATION. THE MEDIAN IS

$$\frac{x_{46} + x_{47}}{2} = \frac{145 + 145}{2}$$

$$= 145 \text{ POUNDS}$$

```
 9 : 5
10 : 288
11 : 002556688
12 : 00012355555
13 : 0000013555688
14 : 000025555588
15 : 00000000003555555555557
16 : 000045
17 : 000055
18 : 0005
19 : 00005
20:
21 : 5
```

WHY MORE THAN ONE MEASURE OF THE CENTER? EACH HAS ADVANTAGES. FOR EXAMPLE, THE *MEDIAN* IS NOT SENSITIVE TO *OUTLIERS*, OR EXTREME VALUES NOT TYPICAL OF THE REST OF THE DATA. SUPPOSE IN OUR SMALL TV-WATCHING GROUP, ONE PERSON WATCHES *200* HOURS PER WEEK. THEN OUR DATA ARE 3, 5, 7, 7, 200. THE MEDIAN, 7, IS UNCHANGED, BUT THE MEAN IS NOW $\bar{x} = 45.8$!



IN 1984 THE UNIVERSITY OF VIRGINIA ANNOUNCED THAT ITS DEPARTMENT OF RHETORIC AND COMMUNICATIONS GRADUATES' *MEAN STARTING SALARY* WAS *$55,000*. THE OUTLIER, THE SALARY OF N.B.A. CENTER *RALPH SAMPSON*, DID *NOT* REPRESENT THE EARNING POWER OF A B.A. IN SPEECH FROM U. OF V. (THE MEDIAN SALARY WASN'T PUBLISHED.)

Gonick, Smith (1993)

Formale Eigenschaften des arithmetischen Mittels $\bar{x}$

- Die Summe der Abweichungen der Merkmalswerte von $\bar{x}$ ist Null.

$$\sum_{i=1}^{n}\left(x_i-\bar{x}\right)=\sum_{i=1}^{n}x_i-n\cdot\bar{x}=\sum_{i=1}^{n}x_i-\sum_{i=1}^{n}x_i=0 \qquad q.e.d.$$

- Die Summe der quadrierten Abweichungen der Merkmalswerte von $\bar{x}$ ist ein Minimum.

$$\sum_{i=1}^{n}\left(x_i-\bar{x}\right)^2=\min$$

Beweis:

$$d(a)=\sum_{i=1}^{n}\left(x_i-a\right)^2 \qquad \text{a beliebig}$$

Extremaproblem: 1.Ableitung wird Null gesetzt → Extrema (min, max) möglich

$$\frac{d(a)}{da}=-\sum_i 2\left(x_i-a\right)\cdot 1 \qquad \textit{Kettenregel}$$
$$=-2\sum_i\left(x_i-a\right)=0$$

Also:

$$\sum_i\left(x_i-a\right)=0$$
$$\sum_i x_i-n\cdot a=0$$
$$\Rightarrow\ a=\frac{1}{n}\sum_i x_i=\bar{x} \quad q.e.d$$

Fechnersche Lageregel zum Vergleich von arithmetischem Mitteln $\bar{x}$, Zentralwert (Median) Z und Modus D

a) symmetrische Verteilung: $\bar{x} = Z = D$
b) asymmetrische Verteilung: $\bar{x} \neq Z \neq D$
Links- und rechtssteile Verteilungen (Abb. III.8)

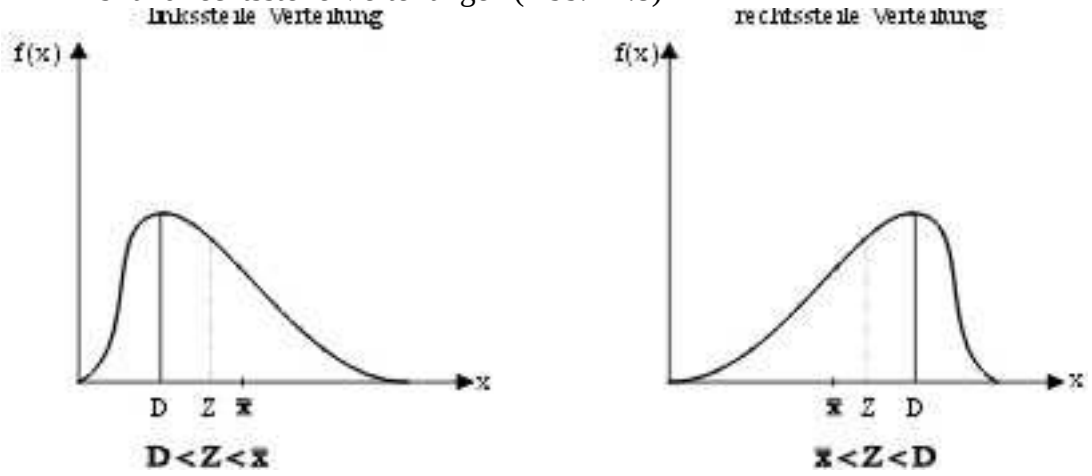


Abb. III.8: Fechnersche Lageregel und links- und rechtssteile Verteilung

2.4 Geometrisches Mittel

Das geometrische Mittel ist sinnvoll bei der Mittlung von Wachstumsraten oder anderen multiplikativ verknüpften Merkmalswerten.

Gegeben: positive Merkmalswerte $x_1, x_2, ..., x_n$

Geometrisches Mittel

$$GM = \sqrt[n]{x_1 \cdot x_2 \cdot ... \cdot x_n} = \left(\prod_{i=1}^{n} x_i \right)^{\frac{1}{n}} \qquad (x_i > 0)$$

oder

$$\log GM = \frac{1}{n}(\log x_1 + \log x_2 + ... + \log x_n) = \frac{1}{n} \sum_{i=1}^{n} \log x_i$$

Beispiel:

Umsätze der Firma F.I.R.M.A. von 2006 - 2010 in Mio. EUR

Jahr	Umsatz	Zuwachsrate in %	Wachstums-faktor
2006	2,0		
2007	2,4	+20,00	1,2000
2008	2,9	+20,83	1,2083
2009	2,7	- 6,89	0,9310
2010	3,1	+14,81	1,1481

Wie groß ist der durchschnittliche relative Umsatzzuwachs (Zuwachsrate) pro Jahr?

$$GM = \sqrt[4]{1,2 \cdot 1,2083 \cdot 0,9310 \cdot 1,1481} = 1,11579$$

Durchschnittliche Zuwachsrate pro Jahr: $(1,11579 - 1) \cdot 100\% = 11,5791\%$

Jahr	Umsatz	$\cdot 1,11579$
2006	2,0	2,2316
2007	2,2316	2,4900
2008	2,4900	2,7783
2009	2,7783	3,1000
2010	3,1000	---

2.5 Harmonisches Mittel

Das harmonische Mittel wird bei der Mittelung von Brüchen mit konstantem Zähler angewandt (z.B.: Geschwindigkeit dividiert durch die Zeit, Preise, Verhältniszahlen).

Gegeben: positive Merkmalswerte $x_1, x_2, ..., x_n$

Harmonisches Mittel

$$HM = \frac{n}{\frac{1}{x_1} + \frac{1}{x_2} + ... + \frac{1}{x_n}} = \frac{n}{\sum_{i=1}^{n} \frac{1}{x_i}}$$

Beispiel:

Fertigungszeiten: Vier Arbeiter sind acht Stunden (= 480 min) lang mit der Herstellung eines Einzelteils beschäftigt:

Arbeiter	Fertigungszeit je Stück in Min.
A	2,3
B	3,0
C	3,4
D	3,7

Wie hoch ist die durchschnittliche Fertigungszeit?

Das arithmetische Mittel ergäbe

$$\overline{x} = \frac{1}{4}(2,3 + 3,0 + 3,4 + 3,7) = 3,1 \text{ min}$$

Besser harmonisches Mittel: Denn Fertigungszeiten je Stück sind arithmetische Mittel, die aus konstanter Arbeitszeit (480 Min.) und der Stückzahl n_i als $\overline{x}_i = 480/n_i$ berechnet sind

$$HM = \frac{n}{\sum_{i=1}^{n} \frac{1}{x_i}} = \frac{4}{\frac{1}{2,3} + \frac{1}{3,0} + \frac{1}{3,4} + \frac{1}{3,7}} = \frac{4}{1,332} = 3,0$$

Die durchschnittliche Fertigungszeit je Stück beträgt also drei Minuten.

3 Streuungsmaße

Mittelwerte bzw. Lageparameter sind in der Praxis wichtige Verteilungsparameter. Sie liefern aber noch unvollständige Beschreibungen einer Häufigkeitsverteilung, denn es läßt sich keine Aussage über die Größe der Abweichungen der einzelnen Merkmalswerte vom Mittelwert machen. Maßzahlen dafür sind die **Streuungsmaße** (vgl. Abb. III.9).

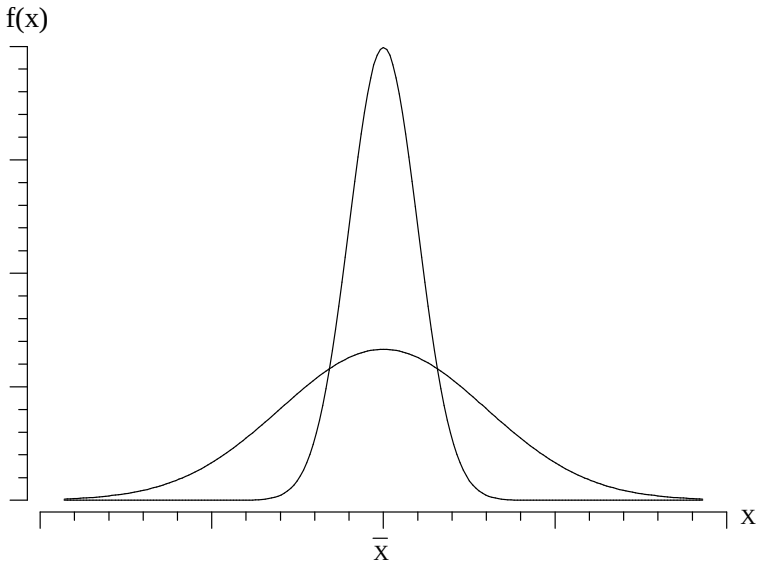


Abb. III.9: Häufigkeitsverteilungen mit gleichem Mittelwert $\bar{x}$ aber verschiedenen Streuungen

3.1 Spannweite

Die Spannweite entspricht dem Konzept des häufigsten Wertes (Modus D).

Spannweite (Range) R

Differenz zwischen dem größten (x_{max}) und dem kleinsten Merkmalswert (x_{min})

$$R = x_{max} - x_{min} = x_{(n)} - x_{(1)}$$

Bei gruppierten Daten werden die Klassengrenzen (Klassenmitten) der Randklassen verwendet.

Der Nachteil der Spannweite R besteht in der Verwendung der extremen Werte (sog. Ausreißer).

Beispiele:

a) Körpergrößen: 172, 178, 164, 167, 171 [cm]

$$x_{max} = x_{(5)} = x_2 = 178$$
$$x_{min} = x_{(1)} = x_3 = 164$$

$$R = x_{max} - x_{min} = 178 - 164 = 14 \text{ cm}$$

b) Temperaturen: 7, 13, -6, 25 [°C]

$x_{max} = 25$

$x_{min} = -6$

$R = 25 - (-6) = 31\ °C$

3.2 Quartilsabweichung und p-Quantile

Ein Streuungsmaß auf der Basis des Mediankonzeptes ist der Quartilsabstand, dessen allgemeine Grundlage die p-Quantile sind.

Für den Median Z ($Z = x_p$ mit $p = 0{,}50 - \text{Quantil}$) gilt, dass 50 % aller Merkmalsträger einen kleineren oder gleich großen (≤) Merkmalswert haben. Z halbiert somit die Fläche der Häufigkeitsdichte/bzw. Häufigkeitsverteilung).

Verallgemeinert gilt für ein p-Quantil, dass p % aller Merkmalsträger einen kleineren oder gleich großen Merkmalswert haben. Damit wird mit einem p-Quantil auch p % der Fläche unter der Häufigkeitsdichte abgetrennt (vgl. Abb. III.10).

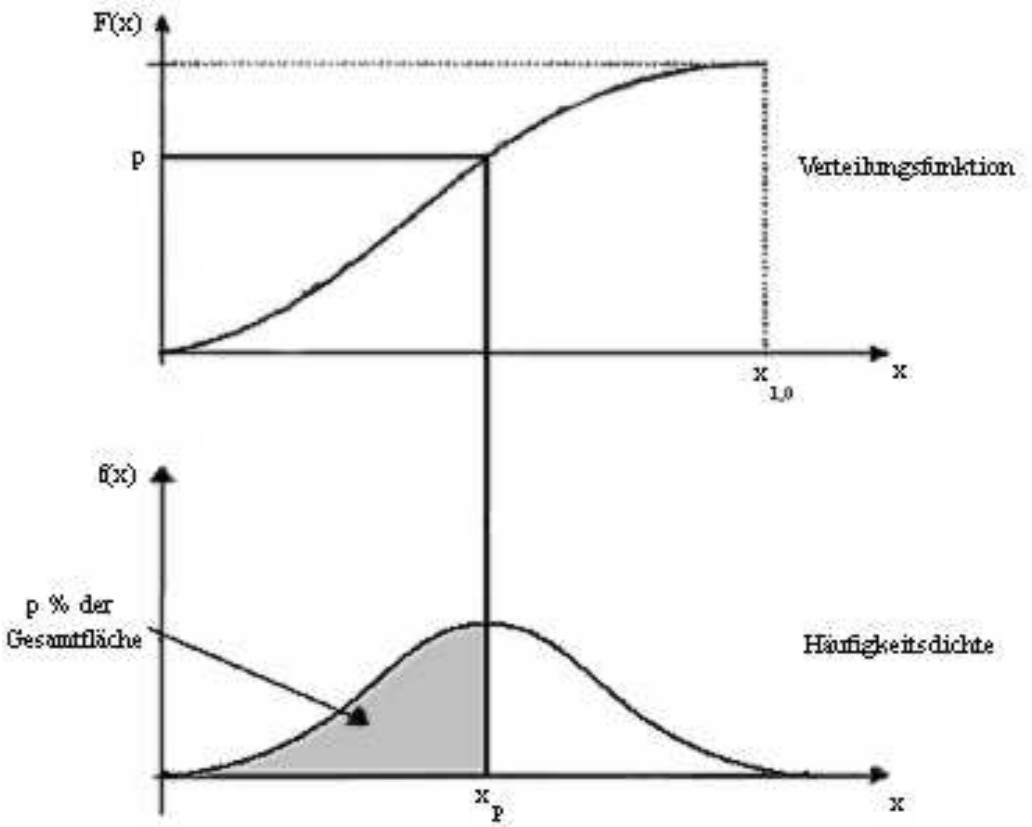


Abb. III.10: p-Quantile, Häufigkeitsdichte und Verteilungsfunktion

Ist also $x_{(1)} \leq .... \leq x_{(n)}$ die geordnete Merkmalsreihe, dann ist allgemein das p-Quantil x_p $(0 < p < 1)$

$$x_p = \begin{cases} x_{(K)} & \text{,falls } n \cdot p \text{ keine ganze Zahl ist} \\ & \text{(K ist dann die auf } n \cdot p \text{ folgende ganze Zahl)} \\ \frac{1}{2}\left(x_{(K)} + x_{(K+1)}\right) & \text{,falls } n \cdot p \text{ eine ganze Zahl ist} \\ & \text{(dann ist K} = n \cdot p) \end{cases}$$

Der Median ist somit ein spezielles p-Quantil, nämlich das mit p = 50 %. Weitere spezielle p-Quantile sind das 0,25-Quantil (unteres Quartil) und das 0,75-Quantil (oberes Quartil) oder die Dezile mit p = 0,10, p = 0,20 etc.

---

p-Quantil der Verteilungsfunktion F(x)

$$F\left(x_p\right) = p \qquad \textit{bzw.} \qquad x_p = F^{-1}\left(p\right)$$

Bei **gruppiertem Datenmaterial** erfolgt die Berechnung von x_p nach der Interpolationsformel:

$$x_p = x_i^u + \frac{F\left(x_p\right) - F\left(x_i^u\right)}{f\left(x_i\right)} = x_i^u + \frac{F\left(x_p\right) - F\left(x_i^u\right)}{n_i / n} \cdot \Delta x_i$$

Quartilsabweichung

p = 0,25 und 0,75

$$QA = \frac{1}{2}\left(x_{0,75} - x_{0,25}\right) \quad F\left(x_{0,25}\right) = 0,25; F\left(x_{0,75}\right) = 0,75; F\left(x_{0,50}\right) = F\left(Z\right) = 0,50$$

In dem Bereich der QA liegen die mittleren 50 % aller Merkmalswerte. Die QA ist nicht von Extremwerten abhängig, sie ist als durchschnittliche Streuung zu interpretieren. Interquartile Spannweiten und 'Box and Whisker'-Plots beschreiben die Bereiche mit 25 % bzw. 50 % der Daten.

Gonick, Smith (1993)

Beispiele:

a) Monatsgehälter in der Fa. DALLES & Co. (siehe Median)
geordnete Werte:
Männer: 1520, 1650, 1670, 1840, 2030 (n = 5)
Frauen: 1490, 1710, 1960, 2570 (n = 4)

Quantile Männer

$$x_{0,25}: n \cdot p = 5 \cdot 0,25 = 1,25 \;\; \textit{keine ganze Zahl}, K = 2, \; x_{0,25} = x_{(2)} = 1650\, EUR$$

25 % aller Monatsgehälter liegen unter 1650 EUR.

$$x_{0,75}: n \cdot p = 5 \cdot 0,75 = 3,75 \; \textit{keine ganze Zahl}, K = 4, x_{0,75} = x_{(4)} = 1840\; EUR$$

Quartilsabweichung: $QA_M = \frac{1}{2}\left(x_{0,75} - x_{0,25}\right) = \frac{1}{2}(1840 - 1650) = 95$

Die 50 % Merkmalswerte um den Median streuen um $\pm 95 \, um \, Z_M = 1670\, EUR$.

Quantile Frauen

$$x_{0,25}: \quad n \cdot p = 4 \cdot 0,25 = 1 \;\; \textit{ganze Zahl}, K = 1$$

$$x_{0,25} = \frac{1}{2}\left(x_{(1)} + x_{(2)}\right) = \frac{1}{2}(1490 + 1710) = 1600\, EUR$$

$$x_{0,75}: \quad n \cdot p = 4 \cdot 0,75 = 3 \;\; \textit{ganze Zahl}, K = 3$$

$$x_{0,75} = \frac{1}{2}\left(x_{(3)} + x_{(4)}\right) = \frac{1}{2}(1960 + 2570) = 2265\, EUR$$

Quartilsabweichung: $QA_F = \frac{1}{2}\left(x_{0,75} - x_{0,25}\right) = \frac{1}{2}(2265 - 1600) = 332,5$

b) Monatliches Haushaltsnettoeinkommen BRD 2009, Quartilsabweichung

Aus $F(0,75)$ folgt:

$$x_{0,75} = x_i^u + \frac{0,75 - F\left(x_i^u\right)}{n_i / n} \cdot \Delta x_i = 2500 + \frac{0,75 - 0,722}{0,101} \cdot 500 = 2.635,86\, EUR$$

Aus $F(0,25)$ folgt:

$$x_{0,25} = 1000 + \frac{0,25 - 0,203}{0,200} \cdot 500 = 1.118,50 \; EUR$$

$$QA = \frac{1}{2}\left(x_{0,75} - x_{0,25}\right) = \frac{1}{2}(2.635,86 - 1.118,50) = 758,68$$

Im Mittel weichen die Haushaltseinkommen um QA =758,68 EUR vom Median Z = 1.753,82 EUR ab.

c) Monatliches Haushaltsnettoeinkommen BRD 2009, Percentile und Spannweite, berechnet auf Grundlage der nicht klassierten Werte der Stichprobe

Tab. III.5: Monatliches Haushaltsnettoeinkommen BRD 2009, Percentile, Quartilsabweichung und Spannweite in EUR

Min	150	Spannweite:	
10 %	850	$R = x_{max} - x_{min}$	29.850,0
20 %	1.140	1. Quartil (25 %)	1.275,0
30 %	1.400	2. Quartil (50 %)	1.924,0
40 %	1.650	3. Quartil (75 %)	2.800,0
Median	1.924	Quartilsabweichung QA	762,5
60 %	2.200		
70 %	2.600		
80 %	3.000		
90 %	4.000		
Max	30.000		

Quelle: Sozio-ökonomisches Panel (Welle Z (26), 2009), eigene Berechnungen

3.3 Mittlere absolute Abweichung

Die mittlere absolute Abweichung ist ein Streuungskonzept hinsichtlich des arithmetischen Mittels.

Da $\sum_{i=1}^{n}(x_i - \bar{x}) = 0$, wird die durchschnittliche absolute Abweichung d gewählt. Positive und negative Abweichungen $(x_i - \bar{x})$ heben sich somit nicht (!) auf.

Mittlere absolute Abweichung d (Mean absolute deviation = MAD)

für ungruppiertes Material:

$$d = \frac{1}{n}\sum_{i=1}^{n}\left|x_i - \bar{x}\right|$$

für gruppiertes Material:

$$d = \frac{1}{n}\sum_{i=1}^{k}\left|x_i^* - \bar{x}\right| \cdot n_i = \sum_{i=1}^{k}\left|x_i^* - \bar{x}\right| \cdot \frac{n_i}{n} = \sum_{i=1}^{k}\left|x_i^* - \bar{x}\right| \cdot h_i,$$

wobei x_i^* die Klassenmitte der Klasse i ist.

Beispiel:

Temperaturen: $-6°, 18°, 12°, 3°$ [°C]

$$\bar{x} = \frac{1}{4}(-6+18+12+3) = \frac{27}{4} = 6,75°C$$

$$d = \frac{1}{n}\sum_{i=1}^{n}|x_i - \bar{x}| = \frac{1}{4}\{|-6-6,75|+|18-6,75|+|12-6,75|+|3-6,75|\}$$

$$= \frac{1}{4}(12,75+11,25+5,25+3,75) = \frac{1}{4}(33) = 8,25$$

Die mittlere absolute Abweichung beträgt 8,25 °C.

3.4 Mittlere quadratische Abweichung: Varianz und Standardabweichung

Konzept des arithmetischen Mittels: Streuungsmaß mit Abweichungen von $\bar{x}$

Gegeben: Merkmal x mit den Ausprägungen $x_1, x_2, ..., x_n$

Varianz

$$s^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

Standardabweichung (standard deviation)

$$s = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2} = \sqrt{s^2} = \sqrt{\mathrm{var}(x)}$$

Die Standardabweichung ist gebräuchlicher, da sie die gleiche Dimension wie die Merkmalswerte aufweist.

Gegenüber der mittleren absoluten Abweichung gewichtet die mittlere quadratische Abweichung (Varianz s^2) größere Abweichungen durch die Quadrierung stärker als kleinere Abweichungen (gebräuchlicher als mittlere absolute Abweichung)

Die Varianz bzw. Standardabweichung wird auch als **empirische Varianz bzw. empirische Standardabweichung** bezeichnet.

Vereinfachte Berechnung von s^2 bzw. s:

$$
\begin{aligned}
s^2 &= \frac{1}{n}\sum_{i=1}^{n}\left(x_i-\bar{x}\right)^2 = \frac{1}{n}\sum_{i=1}^{n}\left(x_i^2-2x_i\bar{x}+\bar{x}^2\right)\\
&= \frac{1}{n}\sum_{i=1}^{n}x_i^2-\frac{1}{n}2\bar{x}\sum_{i=1}^{n}x_i+\frac{1}{n}\sum_{i=1}^{n}\bar{x}^2\\
&= \frac{1}{n}\sum_{i=1}^{n}x_i^2-\frac{1}{n}2\bar{x}\ n\bar{x}+\frac{1}{n}n\bar{x}^2\\
&= \frac{1}{n}\sum_{i=1}^{n}x_i^2-2\bar{x}^2+\bar{x}^2\\
&= \frac{1}{n}\sum_{i=1}^{n}x_i^2-\bar{x}^2
\end{aligned}
$$

Bei nicht-vereinfachter Berechnung: 2 Durchgänge durch den Datensatz
Bei vereinfachter Berechnung: **nur 1 Durchgang** durch den Datensatz!

Dies spart computing-costs bei großen Datenmengen (z.B. EVS mit über 40000 Haushalten)

Beispiel:

a) Temperaturen

i	x_i	x_i^2
1	-6	36
2	18	324
3	12	144
4	3	9
Σ	27	513

$\bar{x} = 27/4 = 6,75$

Vereinfachte Berechnung:
Varianz

$$s^2 = \frac{1}{4}\cdot 513-\left(6,75\right)^2 = 128,25-45,5625 = 82,6875$$

Standardabweichung

$$s = \sqrt{s^2} = 9,0933$$

Nicht vereinfachte Berechnung ($\bar{x} = 6,75$):
Mit $\bar{x} = 6,75$ ergibt sich:

i	x_i	$\left(x_i-\bar{x}\right)$	$\left(x_i-\bar{x}\right)^2$
1	–6	–12,75	162,5625
2	18	11,25	126,5625
3	12	5,25	27,5625
4	3	–3,75	14,0625
Σ	27		330,75

Varianz

$$s^2 = \frac{1}{4} \cdot 330,75 = 82,6875$$

Standardabweichung

$$s = \sqrt{s^2} = 9,0933$$

Zum Vergleich: mittlere absolute Abweichung d = 8,25

b) Computerprogramm zur Varianzberechnung z.B. in FORTRAN:

```
      DO 10 I=1,N
        SUMX=SUMX+X(I)
10    SUMX2=SUMX2+X(I)**2
      VAR=SUMX2/N–(SUMX/N)**2
      SD=SQRT(VAR)
```

Gonick, Smith (1993)

Varianz bei gruppiertem Datenmaterial

Die Berechnung erfolgt nun über die Klassenmitte x_i^*:

$$s^2 = \frac{1}{n}\sum_{i=1}^{k}\left(x_i^* - \overline{x}\right)^2 \cdot n_i$$

vereinfacht:

$$s^2 = \frac{1}{n}\sum_{i=1}^{k}\left(x_i^*\right)^2 \cdot n_i - \overline{x}^2$$

Beispiel:

Monatliches Haushaltsnettoeinkommen BRD 2009,

Tab. III.6: Monatliches Haushaltsnettoeinkommen, Mittelwerte und Streuungsermittlung

x_i^*	n_i	x_i^{*2}	$x_i^{*2} n_i$
250	1.400	62.500	87.500.000
750	6.200	562.500	3.487.500.000
1250	7.500	1.562.500	11.718.750.000
1750	7.200	3.062.500	22.050.000.000
2250	4.800	5.062.500	24.300.000.000
2750	3.800	7.562.500	28.737.500.000
3250	2.400	10.562.500	25.350.000.000
3750	1.700	14.062.500	23.906.250.000
4250	860	18.062.500	15.533.750.000
4750	740	22.562.500	16.696.250.000
5250	270	27.562.500	7.441.875.000
	37.510		204.309.375.000

$$s^2 = \frac{1}{n}\sum_i\left(x_i^*\right)^2 \cdot n_i - \overline{x}^2 = \frac{204.309.375.000}{37.510} - 1.999{,}13^2$$
$$= 5.446.797{,}67 - 3.996.520{,}80 = 1.450.276{,}8$$

$$s = \sqrt{1.450.276{,}8} = 1204{,}27\,EUR$$

Interpretation: Das Haushaltseinkommen weicht im Durchschnitt um 1204,27 EUR vom mittleren Haushaltseinkommen $\overline{x} = 1.999{,}13\,EUR$ ab.
Für die Quartilsabweichung ergab sich ein Wert von 762,50 EUR.

3.5 Variationskoeffizient

Für den Vergleich verschiedener Häufigkeitsverteilungen wird eine 'relative Streuung' (Streuungsmaß/Lagemaß) definiert mit

Gonick, Smith (1993)

Variationskoeffizient

$$V = \frac{s}{\overline{x}} \cdot 100(\%) = \frac{\text{Standardabweichung}}{\text{Mittelwert}} \cdot 100$$

Problem: Bei positiven und negativen Merkmalswerten kann $\overline{x}$ nahe Null sein. Dadurch entstehen 'beliebig' große Werte von V.

Beispiel:

a) Temperaturbeispiel

$$s = 9,09\,[°C]$$

$$\overline{x} = 6,75\,[°C]$$

$$V = \frac{s}{\overline{x}} = \frac{9,09\,[°C]}{6,75\,[°C]} = 1,35 \qquad (\text{dimensionslos !})$$

b) Monatliches Haushaltsnettoeinkommen BRD 2009

$$s = 1204,27\,[EUR]$$

$$\overline{x} = 1.999,13\,[EUR]$$

$$V = \frac{s}{\overline{x}} = \frac{1.204,27}{1.999,13} = 0,60$$

Interpretation: Relative Streuung von b) < a)

Aber: Vorsicht bei der inhaltlichen Interpretation! Es ist besser, Äpfel mit Äpfeln und Birnen mit Birnen zu vergleichen!

3.6 Konzept der Momente, Schiefe und Exzeß

Momente sind Verallgemeinerungen des Varianzkonzeptes.

Momente

Durchschnittliche **potenzierte** Abweichungen der Merkmalswerte von einem Bezugspunkt (a).

Bezugspunkt Null: Momente um Null
Bezugspunkt arithmetisches Mittel: Momente um das arithmetische Mittel

Zentrale Momente

$$m_r^a = \frac{1}{n}\sum_{i=1}^{n}(x_i - a)^r \qquad (\text{zentrale Momente})$$

Das r-te Moment um Null

$$m_r^0 = \frac{1}{n}\sum_{i=1}^{n}(x_i - 0)^r = \frac{1}{n}\sum_{i=1}^{n} x_i^r \qquad \text{(ungruppierte Daten)}$$

$$m_r^0 = \frac{1}{n}\sum_{i=1}^{k}\left(x_i^* - 0\right)^r \cdot n_i = \frac{1}{n}\sum_{i=1}^{k}\left(x_i^*\right)^r \cdot n_i \qquad \text{(gruppierte Daten)}$$

$$\left(\text{für } r = 1: \qquad m_r^0 = \bar{x}\right)$$

Das r-te Moment um das arithmetische Mittel $\bar{x}$

$$m_r = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^r \qquad \text{(ungruppierte Daten)}$$

$$m_r = \frac{1}{n}\sum_{i=1}^{K}\left(x_i^* - \bar{x}\right)^r \cdot n_i \qquad \text{(gruppierte Daten)}$$

Momente höherer Ordnung ergeben die Schiefe (r = 3) und den Exzeß (Kurtosis, Wölbung) (r = 4).

Schiefe
Das 3. zentralen Moment (r = 3) gibt Auskunft über die Symmetrie, bzw. Asymmetrie einer Verteilung.

3. zentrales Moment:

$$m_3 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^3$$

Das 3. zentrale Moment allein ist jedoch kein sehr geeignetes Maß für die Unsymmetrie, da seine Größe von der Streuung und der Maßeinheit der Variable beeinflusst wird. Daher wird das folgende Schiefemaß (skewness) verwendet:

Schiefe:

$$sm_3 = \frac{m_3}{s^3} = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^3}{\left[\sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2}\right]^3}$$

Symmetrie: $sm_3 = 0$

Asymmetrie: je stärker negativ die Maßzahl, desto rechtssteiler (linksschiefer) ist die Verteilung.
je stärker positiv die Maßzahl, desto linkssteiler (rechtsschiefer) ist die Verteilung.

Sinnvoll bei Eingipfligkeit der Verteilung (unimodal).
ET verwendet z.B. sm_3 mit $(n-1)$ statt n.

Wölbung (Kurtosis, Exzeß)
Auskunft über den Grad der Wölbung oder Spitzigkeit einer Verteilung gibt das 4. zentralen Moment (r = 4).

4. zentrales Moment:

$$m_4 = \frac{1}{n}\sum_{i=1}^{n}\left(x_i - \overline{x}\right)^4$$

Das 4. zentrale Moment ist für jede Verteilung positiv.
Kleinerer Werte des 4. zentralen Moments deuten auf eine *spitzere/steiler gewölbte* Verteilung.
Große Werte weisen auf eine *flachere* Verteilung hin.

Um ein maßstabs- und streuungsunabhängiges Maß zu erhalten verwendet man die Kurtosis als Maß für die Wölbung ('thickness of the distribution tails').

Kurtosis:

$$sm_4 = \frac{m_4}{s^4} = \frac{\frac{1}{n}\sum_{i=1}^{n}\left(x_i - \overline{x}\right)^4}{\left[\sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(x_i - \overline{x}\right)^2}\right]^4}$$

ACHTUNG: Durch die relativ großen s^4-Werte dreht sich die Interpretation von sm_4 gegenüber m_4 um:
Nun steht ein *kleinerer* standardisierter Exzeß-Wert für eine *flachere* Verteilung,
große Werte weisen auf eine *spitzere* Verteilung hin.

Da für die Normalverteilung $sm_4 = 3$ ist, erfolgt auch hier manchmal eine Normierung mit

$$sm_4^* = sm_4 - 3$$

Ist $sm_4^* > 0$, dann ist die Verteilung *spitzer/steiler gewölbt* als die Normalverteilung (bei gleicher Varianz und Mittelwert).
Ist $sm_4^* < 0$, dann ist die Verteilung *flacher* als die Normalverteilung (bei gleicher Varianz und Mittelwert).

Beispiele:

a) Monatliches Haushaltsnettoeinkommen, BRD 2009
Bisher:

$$\overline{x} = 1.999{,}13\,EUR$$

$$s = 1.204{,}27\,EUR$$

gruppierte Daten: $m_r = \frac{1}{n}\sum_{i=1}^{k}\left(x_i^* - \overline{x}\right)^r \cdot n_i$

Tab. III.7: Monatliches Haushaltsnettoeinkommen, BRD 2009, Hilfswerte für Schiefe und Exzeß

i	Einkommens-klasse	n_i	x_i^*	$\left(x_i^* - \bar{x}\right)^3 n_i$	$\left(x_i^* - \bar{x}\right)^4 n_i$
1	unter 500	1.400	250	-7,492E+12	1,310E+16
2	500 - unter 1000	6.200	750	-1,208E+13	1,509E+16
3	1000 - unter 1500	7.500	1.250	-3,153E+12	2,362E+15
4	1500 - unter 2000	7.200	1.750	-1,113E+11	2,774E+13
5	2000 - unter 2500	4.800	2.250	7,579E+10	1,901E+13
6	2500 -unter 3000	3.800	2.750	1,609E+12	1,208E+15
7	3000 - unter 3500	2.400	3.250	4,697E+12	5,876E+15
8	3500 - unter 4000	1.700	3.750	9,125E+12	1,598E+16
9	4000 - unter 4500	860	4.250	9,807E+12	2,207E+16
10	4500 - unter 5000	740	4.750	1,540E+13	4,238E+16
11	5000 - unter 5500	270	5.250	9,276E+12	3,016E+16
12	5500 - unter 7000	640	6.250	4,916E+13	2,090E+17
		37.510		7,631E+13	3,572E+17

Quelle: Sozio-ökonomisches Panel (Welle Z (26), 2009), eigene Berechnungen

Schiefe:

$$m_3 = \frac{1}{n}\sum_{i=1}^{k}\left(x_i^* - \bar{x}\right)^3 \cdot n_i$$
$$= 2{,}034\text{E+}09$$

$$sm_3 = \frac{m_3}{s^3} = \frac{2{,}034\text{E+}09}{1{,}747\text{E+}09} = 1{,}165$$

Exzeß (Wölbung):

$$m_4 = \frac{1}{n}\sum_{i=1}^{k}\left(x_i^* - \bar{x}\right)^4 \cdot n_i$$
$$= 9{,}524\text{E+}12$$
$$sm_4 = \frac{m_4}{s^4} = \frac{9{,}524\text{E+}12}{2{,}103\text{E+}12} = 4{,}528$$

$$sm_4^* = m_4 / s^4 - 3 = 4{,}528 - 3 = 1{,}528$$

Damit ist die Normalverteilung weniger flacher

b) Einkommensvergleich BRD (2009) und U.K. (1979-80)

Tab. III.8: Einkommensverteilung für U.K. (1979 - 80)

	Mittelwert $\bar{x}$	Varianz $m_2 = s^2$	Schiefe sm_3	Exzeß sm_4
Haushaltsnettoeinkommen BRD [EUR]	1.999,13	1,450E+06	1,165	4,528
Personal income* U.K. [£]	3.700,-	$4{,}8 \cdot 10^6$	1,45	7,48

*Quelle: Spanos (1986), S. 24 ff, Sozio-ökonomisches Panel (Welle Z (26), 2009), eigene Berechnungen

Es liegen zwar verschiedene Zeitpunkte und verschiedene Währungen (EUR und £) sowie Haushalts- bzw. Personenkonzepte vor, dennoch:

jeweils U.K. **größerer Mittelwert** als BRD
U.K. **größere Streuung** als BRD
U.K. **Verteilung linkssteiler** als BRD.
U.K. **flachere, breitere Verteilung** als BRD

c) ET: F7, F8 Descriptives, Histogramm

für Körpergrößen von 20 Studentinnen (HWOMEN) und 20 Studenten (HMEN)

- Listing
- Descriptive Statistics
- Stem and leaf Plots
- Percentile
- Box and Whisker-Plots
- Frequency Tables
- Histograms

```
DATA LISTING (Current sample)     ↵/ESC        Press ESC to interrupt list.
Observation      HMEN        HWOMEN
      1        182.00      158.00
      2        182.00      164.00
      3        180.00      174.00
      4        187.00      178.00
      5        179.00      163.00
      6        184.00      168.00
      7        174.00      170.00
      8        168.00      165.00
      9        186.00      158.00
     10        172.00      154.00
     11        178.00      169.00
     12        182.00      176.00
     13        186.00      172.00
     14        174.00      174.00
     15        183.00      164.00
     16        192.00      168.00
     17        177.00      169.00
     18        179.00      168.00
     19        176.00      167.00
     20        178.00      163.00


                          Descriptive Statistics
Variable    Mean    Std. Dev.     Skew.   Kurt.   Minimum     Maximum   Cas↵
HMEN      179.95     5.6983      -.020   2.690    168.0      192.0       20
HWOMEN    167.10     6.1976      -.254   2.480    154.0      178.0       20


Stem and Leaf Plot for HMEN                Use ↑ and ↓ to scroll. ESC=exit.
                                             8 lines            TOP of file
Total number of observations =   20
    1   Low  values discarded
    1   high values discarded

Stem width = 100.00

Count   Stem    Leaves
   18      1 .  777777777888888888
```

Stem and Leaf Plot for HWOMEN Use ↑ and ↓ to scroll. ESC=exit.
8 lines TOP of file

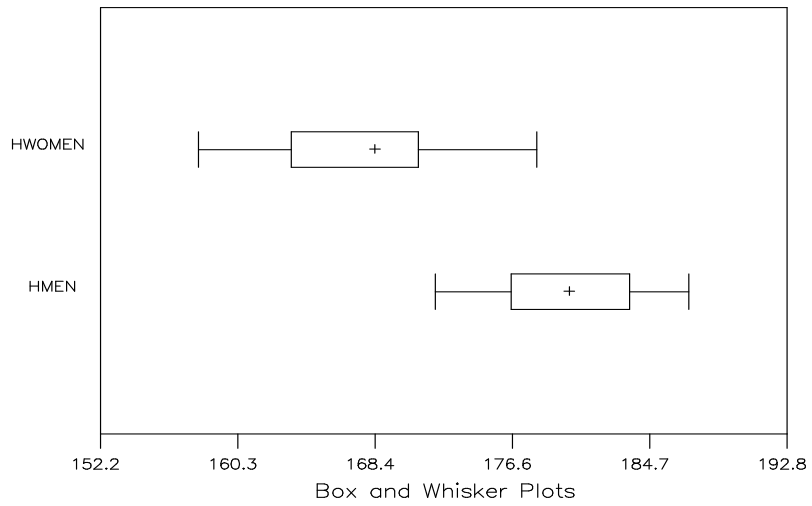
```
Toal number of observations =   20
    1   Low  values discarded
    0   high values discarded
t
Stem width = 100.00

Count   Stem    Leaves
   19      1 .  5566666666666777777
```

Order Statistics for Variables

Percentile	HMEN	HWOMEN
Min.	168.00	154.00
10th	173.00	158.00
20th	175.00	163.00
25th	176.50	163.50
30th	177.50	164.00
40th	178.50	166.00
Med.	179.50	168.00
60th	182.00	168.50
70th	182.50	169.50
75th	183.50	171.00
80th	185.00	173.00
90th	186.50	175.00
Max.	192.00	178.00

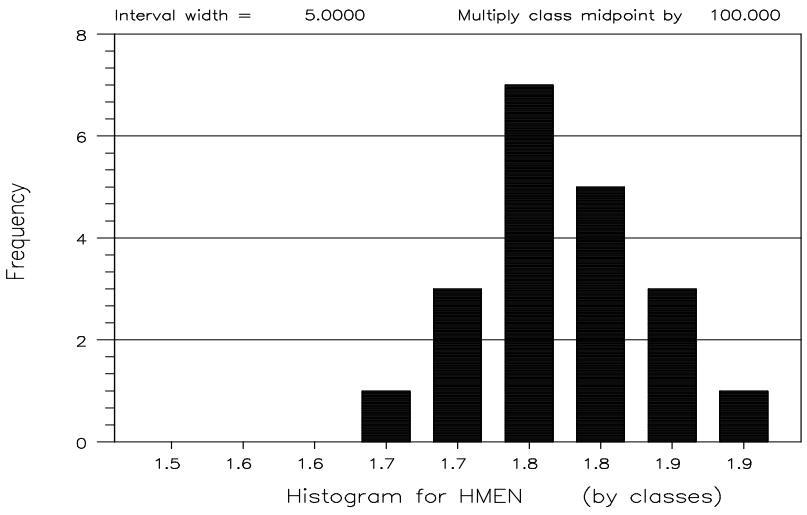
Partition of range Min to Max

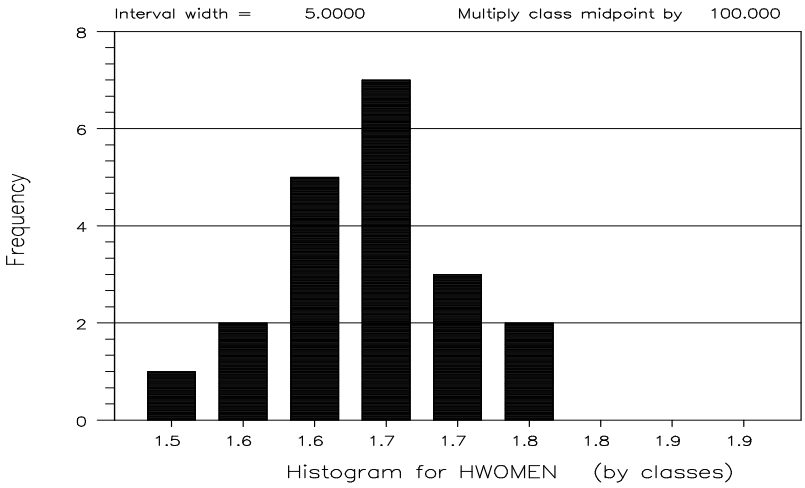
Range of X	HMEN	HWOMEN
Minimum	168.00	154.00
1st.Qrtl	174.00	160.00
Midpoint	180.00	166.00
3rd.Qrtl	186.00	172.00
Maximum	192.00	178.00



Histogram for HMEN computed using 20 observations
Obs. out of range: too low= 0, too high= 0

```
     Individual data      Mean=  179.950, std.dev.=    5.698
                                     Frequency          Cumulative
    Lower Limit    Upper Limit    Total   Relative   Total   Relative
0       150.000        155.000        0      .0000       0      .0000
1       155.000        160.000        0      .0000       0      .0000
2       160.000        165.000        0      .0000       0      .0000
3       165.000        170.000        1      .0500       1      .0500
4       170.000        175.000        3      .1500       4      .2000
5       175.000        180.000        7      .3500      11      .5500
6       180.000        185.000        5      .2500      16      .8000
7       185.000        190.000        3      .1500      19      .9500
8       190.000        195.000        1      .0500      20     1.0000

     Histogram for HWOMEN   computed using   20 observations
     Obs. out of range: too low=   0, too high=   0
     Individual data      Mean=  167.100, std.dev.=    6.198
                                     Frequency          Cumulative
    Lower Limit    Upper Limit    Total   Relative   Total   Relative
0       150.000        155.000        1      .0500       1      .0500
1       155.000        160.000        2      .1000       3      .1500
2       160.000        165.000        5      .2500       8      .4000
3       165.000        170.000        7      .3500      15      .7500
4       170.000        175.000        3      .1500      18      .9000
5       175.000        180.000        2      .1000      20     1.0000
6       180.000        185.000        0      .0000      20     1.0000
7       185.000        190.000        0      .0000      20     1.0000
8       190.000        195.000        0      .0000      20     1.0000
```





4 Konzentration einer Verteilung

Zur Untersuchung der Frage, ob sich bestimmte Merkmale (Einkommen, Vermögen, Firmenumsätze etc.) bei bestimmten anderen Merkmalen (Personen, Haushalte, Firmentypen etc.) konzentrieren; also Frage nach der Gleich-/Ungleichverteilung.

4.1 Konzentration

Die Standardabweichung ist bereits ein Maß für die Konzentration:
s mißt die durchschnittliche Abweichung von der Gleichverteilung des Merkmals i ($\overline{x}$ gibt an, welcher Wert sich ergibt, wenn die Merkmalssumme auf alle Einheiten gleich aufgeteilt würde).

Informativer ist die Abweichung von der Gleichverteilung für jede Klasse!

Für Konzentrationsanalysen werden die Merkmale grundsätzlich erst nach ihrer Größe geordnet.

Gleich-/Ungleichverteilung über Klassen

Für jede Klasse:

Bilde die **Differenz** d_i zwischen der **beobachteten Merkmalssumme** $x_i^* \cdot n_i$ und der **Merkmalssumme bei Gleichverteilung** $\overline{x} \cdot n_i$.

$$d_i = x_i^* \cdot n_i - \overline{x} \cdot n_i \qquad \text{für} \qquad i = 1,2,...,k$$

Anteil an gesamter Merkmalssumme:

$$\tilde{d}_i = \frac{d_i}{\overline{x} \cdot n} = \frac{x_i^* \cdot n_i}{\overline{x} \cdot n} - \frac{n_i}{n} \qquad \text{für} \qquad i = 1,2,...,k$$

Eine Gleichverteilung liegt dann vor, wenn $\tilde{d}_i = 0$ für alle Klassen gilt.

Beispiel:

Monatliches Haushaltsnettoeinkommen BRD 2009

d_i = Differenz zwischen

Einkommensanteil der Haushalte der Klasse i $\left(x_i^* \cdot n_i / \overline{x} \cdot n\right)$ und

Anteil der Haushalte dieser Klasse i an allen Haushalten $\left(n_i / n\right)$

Tab. III.9: Monatliches Haushaltsnettoeinkommen 2009, Konzentration der Verteilung

Einkommens-klasse	x_i^*	n_i	$\frac{n_i}{n}$	$x_i^* \cdot n_i$	$\frac{x_i^* \cdot n_i}{n \cdot \overline{x}}$	$\tilde{d}_i$	$F\left(x_i^o\right)$	$MS\left(x_i^o\right)$
unter 500	250	1.400	0,037	350.000	0,005	-0,033	0,037	0,005
500 - unter 1000	750	6.200	0,165	4.650.000	0,062	-0,103	0,203	0,067
1000 - unter 1500	1250	7.500	0,200	9.375.000	0,125	-0,075	0,403	0,192
1500 - unter 2000	1750	7.200	0,192	12.600.000	0,168	-0,024	0,595	0,360
2000 - unter 2500	2250	4.800	0,128	10.800.000	0,144	0,016	0,722	0,504
2500 -unter 3000	2750	3.800	0,101	10.450.000	0,139	0,038	0,824	0,643
3000 - unter 3500	3250	2.400	0,064	7.800.000	0,104	0,040	0,888	0,747
3500 - unter 4000	3750	1.700	0,045	6.375.000	0,085	0,040	0,933	0,832
4000 - unter 4500	4250	860	0,023	3.655.000	0,049	0,026	0,956	0,881
4500 - unter 5000	4750	740	0,020	3.515.000	0,047	0,027	0,976	0,928
5000 - unter 5500	5250	270	0,007	1.417.500	0,019	0,012	0,983	0,947
5500 - unter 7000	6250	640	0,017	4.000.000	0,053	0,036	1,000	1,000
		37.510	1,00	74987500,00 $= n \cdot \overline{x}$	1,000	0,00		

Quelle: Sozio-ökonomisches Panel (Welle Z (26), 2009), eigene Berechnungen

Interpretation:

Konzentration (Ungleichheit) dadurch, dass $\tilde{d}_i$ bis unter 2000 negativ, für die übrigen Klassen positiv ist.

Die beiden letzten Spalten dieser Tabelle sind besonders interessant. Aus

$$\tilde{d}_i = \underbrace{\frac{x_i^* \cdot n_i}{\overline{x} \cdot n}}_{\text{Beitrag zu kumulierter relativer Merkmalssumme}} - \underbrace{\frac{n_i}{n}}_{\text{Verteilungsfunktion}}$$

folgt:

Kumulierte relative Merkmalssumme: $$MS\left(x_i^o\right) = \frac{\sum_{i=1}^{i^o} x_i^* \cdot n_i}{\overline{x} \cdot n}$$

Verteilungsfunktion: $$F\left(x_i^o\right) = h\left(x \le x_i^o\right) = \sum_{i=1}^{i^o} \frac{n_i}{n}$$

Interpretation:

Auf $F\left(x_i^o\right)[\%]$ aller Einheiten mit $x < x_i^o$ entfallen $MS\left(x_i^o\right)[\%]$ der gesamten Merkmalssumme.

Beispiel:

Monatliches Haushaltsnettoeinkommen 2009, Einkommensklasse i = 3
Auf 40,3 % aller Haushalte (mit einem Einkommen unter 1500 EUR) fallen nur 19,2 % des Gesamteinkommens aller Haushalte.

oder: Einkommensklasse i = 11
Auf 98,3 % aller Haushalte (mit einem Einkommen unter 5500 EUR) fallen 94,7 % des Gesamteinkommens aller Haushalte.

4.2 Lorenzkurve und Gini-Koeffizient

Zur zusammenfassenden grafischen und quantitativen Beschreibung der Konzentration einer Verteilung wird die Lorenzkurve und der Gini-Koeffizient verwendet.

Lorenzkurve

Für die Lorenzkurve überträgt man die Wertepaare $\left[F\left(x_i\right), MS\left(x_i\right)\right]$ in ein Koordinatensystem.
Lorenzkurve = Streckenzug, der [0,0] mit allen Wertepaaren $\left[F\left(x_i\right), MS\left(x_i\right)\right]$ verbindet.

Gleichverteilung:
Gleichverteilung liegt dann vor, wenn $F\left(x_i\right) = MS\left(x_i\right)$ für alle i ist. Es herrscht dann keine Konzentration vor, da alle Punkte auf der Diagonalen eines Quadrates liegen. Die Diagonale verläuft von [0,0] bei: $\left[\left(F\left(x_i\right), MS\left(x_i\right)\right)\right] = \left[\left(F\left(x_i\right), MS\left(x_i\right)\right)\right]$ bis [1,1].

Voraussetzungen für den Vergleich zweier Verteilungen:

Die Lorenzkurven dürfen sich nicht schneiden!
Es sollten zusätzliche Informationen verwendet werden, damit keine Fehlinterpretationen entstehen.

Beispiel:

Lorenzkurve für das monatliche Haushaltsnettoeinkommen in der BRD 2009 (Abb. III.11)

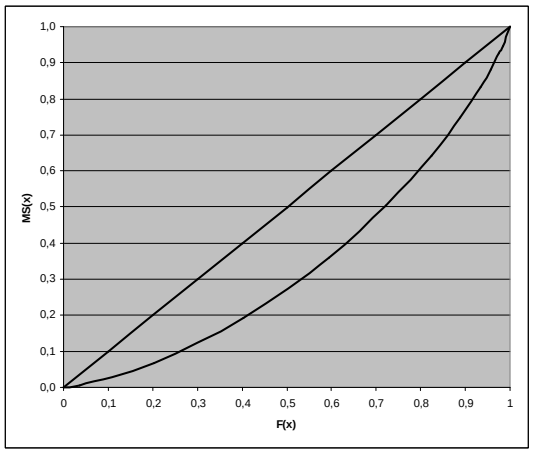


Abb. III.11: Lorenzkurve für das monatliche Haushaltsnettoeinkommen 2009 in der BRD
Quelle: Sozio-ökonomisches Panel (Welle Z (26), 2009), eigene Berechnunge

Gini-Koeffizient

Die Fläche zwischen der Gleichverteilungsgeraden und der Lorenzkurve wird als Maß für die Konzentration verwendet.

$$G = \frac{\text{Fläche zwischen Lorenzkurve und Gleichverteilungsgerade}}{\text{Fläche des Dreiecks unter der Gleichverteilungsgeraden}}$$

Je kleiner die Fläche (bzw. G), desto gleichverteilter.

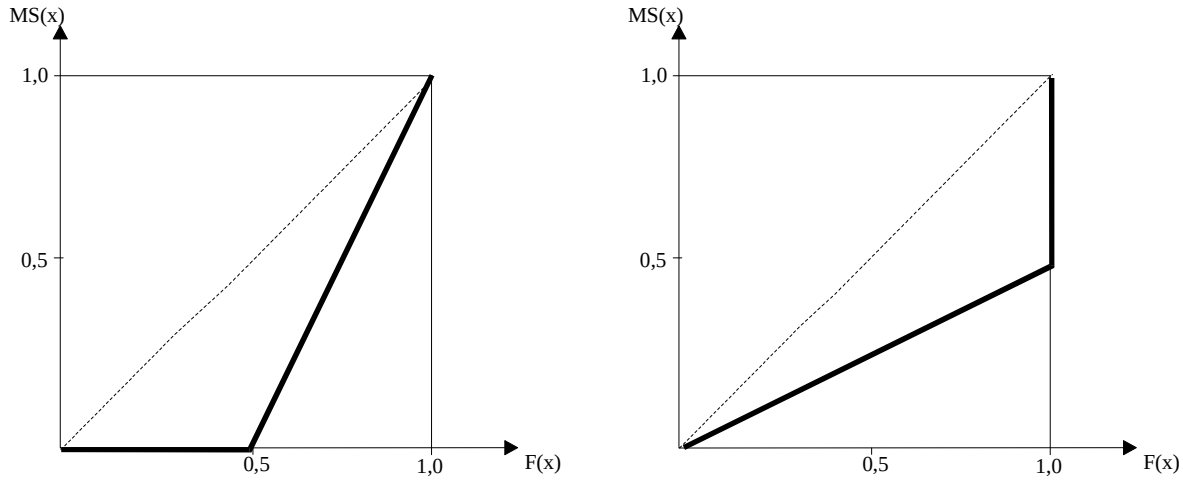
$$G = \sum_{i=1}^{k} \left\{ \left[F\left(x_{i-1}\right) + F\left(x_i\right) \right] \cdot \frac{n_i \cdot x_i^*}{n \cdot \overline{x}} \right\} - 1$$

ungruppiert (x_i **geordnet!)**

$$G = \frac{2 \cdot \sum_{i=1}^{n} i \cdot x_i - (n+1) \cdot \sum_{i=1}^{n} x_i}{n \cdot \sum_{i=1}^{n} x_i}$$

Beispiele

a) Gleicher Gini-Koeffizient bei verschiedenen Sachverhalten:



A **B**

Abb. III.12: Gleiche Gini-Koeffizienten bei verschiedenem Sachverhalten

A: Auf die eine Hälfte der Haushalte entfällt ein Haushaltsnettoeinkommen von (z.B.) Null, während die andere Hälfte alle das gleiche Haushaltsnettoeinkommen haben.

B: Die Hälfte des gesamten Einkommens entfällt gleichmäßig auf alle Haushalte mit Ausnahme eines einzigen Haushalts, der die andere Hälfte des Gesamteinkommens bezieht.

b) Gini-Koeffizienten und zusammenfassende deskriptive Informationen für das monatliche Haushaltsnettoeinkommen 1983, 1984,1992 und 2009 in der BRD

Zur Einkommens- und Verbrauchsstatistik vgl. auch Abschnitt II.6.1.

Tab. III.11: Gini-Koeffizienten und zusammenfassende deskriptive Informationen für das monatliche Haushaltsnettoeinkommen 1983, 1984, 1992 und 2009

	EVS (1983)	SOEP (1983)	SOEP (1991)	SOEP (2009)
Zahl der Beobachtungen	42750	5587	6431	9768
Hochger. Merkmalsumme (Mrd. DM)	75,35	61,88	111,06	83,36
Hochger. Zahl d. Bezugseinheiten	22545383	24599576	34763559	37509999
Hochger. arithmetisches Mittel (DM)	3342	2514	3195	2230
Varianz	$4{,}4648\ 10^{6}$	$5{,}2203\ 10^{6}$	$3{,}738\ 10^{6}$	$1{,}45\ 10^{6}$
Schiefe	$1{,}03\ 10^{9}$	$2{,}927\ 10^{9}$	$0{,}5598\ 10^{10}$	$2{,}03\ 10^{9}$
Exzeß	$2{,}18\ 10^{13}$	$2{,}12\ 10^{13}$	$4{,}4\ 10^{13}$	$9{,}52\ 10^{12}$
Spannweite (DM)	24836	82185	20750	29850
Quartilsabstand (DM)	1250	800	1050	762,5
Median (DM)	2917	2100	2800	1924
Gini-Koeffizient	0,32851	0,31921	0,30149	0,32
N-tils-Verteilung (in %)				
0% - 25%	9,69	10,12	10,27	10,17
25% - 50%	17,67	18,09	18,12	17,79
50% 75%	26,80	26,08	26,70	26,24
75% 100%	45,02	44,78	44,56	45,80
Ant. d. untersten 20% (in %)	7,08	7,30	7,45	7,93
Ant. der obersten 5% (in %)	14,27	15,00	13,01	14,93
Randgruppenrelation 90/10	8,06	8,22	6,99	6,87

Quelle: (Hansen (1974), S. 18); Die Berechnungen auf der Basis der EVS 1983 wurden freundlicherweise durch Herrn Jürgen Faik an der Professur für Sozialpolitik, Prof. Dr. Hauser, Universität Frankfurt, vorgenommen (vgl. auch Merz und Faik (1992)); Sozio-ökonomisches Panel (Welle I (1), 1984; Welle I (9), 1992; Welle Z (26), 2009), eigene Berechnungen, hochgerechnete Werte

c) Einkommensverteilung im Längsschnitt aus Daten des Sozio-ökonomischen Panels (vgl. Abb. III.13)

Tabelle 1: Regelsätze nach § 22 Bundessozialhilfegesetz und Bedarfsgewichte

Zeitraum	Personen					
	Erwachsene		Kinder			
	HHV[1)]	Andere	Alter 0-6	Alter 7-10	Alter 11-14	Alter 15-21
Mindestsätze in DM						
1.7.1982-30.6.1983	338	270	152	220	254	304
1.7.1983-30.6.1984	345	276	155	224	259	311
1.7.1984-30.6.1985	356	285	160	232	267	321
1.7.1985-30.6.1986	384	307	173	250	288	346
1.7.1986-30.6.1987	394	315	177	256	295	354
Bedarfsgewicht in %	1,0	0,8	0,45	0,65	0,75	0,9

Anmerkung: Rechnerischer Durchschnitt für das Bundesgebiet; Regelsätze ohne Mehrbedarf

Quelle: Nachrichtendienst des Deutschen Vereins für öffentliche und private Fürsorge; verschiedene Jahrgänge

1) HHV = Haushaltsvorstand

Tabelle 2: Kennziffern der Verteilung des Nettoeinkommens von Haushalten und der Nettowohlstandsposition von Haushalten und Personen (NWP errechnet) für die Jahre 1983 bis 1986

	1983	1984	1985	1986
Haushalte				
Arithmetisches Mittel	2919	2991	3019	3070
Zentralwert	2548	2543	2560	2612
Gini-Koeffizient	0,324	0,338	0,334	0,338
Quintilsverteilung in %				
1.Quintil	6,89	6,59	6,90	6,65
2.Quintil	12,33	11,94	12,02	11,93
3.Quintil	17,48	17,15	17,10	17,10
4.Quintil	24,20	24,17	23,99	24,20
5.Quintil	39,08	40,15	39,99	40,12
Anteil der oberen 5%	13,47	14,25	14,43	14,38
Personen				
Arithmetisches Mittel	1434	1493	1516	1558
Zentralwert	1287	1333	1346	1403
Gini-Koeffizient	0,259	0,267	0,268	0,268
Quintilsverteilung in %				
1.Quintil	9,61	9,48	9,38	9,15
2.Quintil	14,15	13,84	13,86	14,02
3.Quintil	17,99	17,90	17,77	17,96
4.Quintil	22,76	22,91	22,80	22,89
5.Quintil	35,49	35,98	36,18	35,97
Anteil der oberen 5%	12,48	12,90	13,13	12,82

Quelle: Berntsen, R. (1991)

Nettowohlstandsposition:

$$\text{NWP} = \frac{\sum \text{Einkünfte } (\text{HH})}{\sum \text{Bedarfsgewichte}^*}$$

(* aus Regelsätzen nach § 22 Bundessozialhilfegesetz)

Tabelle 3: Verteilung der Wohlstandsposition von Personen nach Vielfachen der durchschnittlichen Wohlstandsposition - Vergleich der Jahre 1983 mit 1986 (Wanderungsbilanz)

	Wohlstandsposition 1986								
Wohlstandsposition 1983	unter 0,5 %	0,50 bis 0,75 %	0,75 bis 1,00 %	1,00 bis 1,25 %	1,25 bis 1,50 %	1,50 bis 1,75 %	1,75 bis 2,00 %	2,00 und mehr %	Gesamt %
Nettowohlstandsposition									
unter 0,5 %	39,3	42,6	8,7	5,3	2,1	[0,9]	[0,5]	[0,6]	100
0,50 bis unter 0,75 %	16,1	49,6	24,9	6,1	2,2	(1,0)	[0,1]	[0,1]	100
0,75 bis unter 1,00 %	5,8	22,6	43,1	18,8	7,7	(0,9)	(0,7)	(0,4)	100
1,00 bis unter 1,25 %	3,9	9,4	26,1	36,9	14,1	4,7	2,0	3,0	100
1,25 bis unter 1,50 %	(2,1)	8,1	9,6	25,7	30,4	14,2	7,0	2,9	100
1,50 bis unter 1,75 %	(2,7)	(4,5)	9,1	14,4	23,3	27,3	9,3	9,4	100
1,75 bis unter 2,00 %	[0,4]	(5,0)	(6,0)	9,6	12,4	25,3	24,6	16,6	100
2,00 und mehr %	[1,5]	(2,7)	(4,8)	9,8	10,0	12,1	12,8	46,2	100

Erläuterung: ( ) = Fallzahl unter 30 Personen;
[ ] = Fallzahl unter 10 Personen

Quelle: Berntsen, R. (1991)

Tabelle 4: Kurzfristige relative Veränderungsklassen der Wohlstandsposition von Personen 1983 bis 1986

Veränderungsklassen	Kurzfristige Veränderungsraten* Anteile in %		
	1984 zu 1983	1985 zu 1984	1986 zu 1985
Nettowohlstandsposition			
Relative Aufstiege			
50 % u.m.	5,7	6,0	5,5
25 bis unter 50 %	8,8	10,0	8,0
10 bis unter 25 %	13,6	14,6	14,5
0 bis unter 10 %	19,5	21,2	24,0
Relative Abstiege			
0 bis unter 10 %	22,4	21,0	23,7
10 bis unter 25 %	19,6	16,7	13,6
25 bis unter 50 %	8,4	7,7	8,1
50 % u.m.	2,0	2,8	2,7

* Relative Veränderung zwischen zwei Jahren in Veränderungsklassen

Quelle: Berntsen, R. (1991)

Abb. III.13: Einkommensverteilung im Längsschnitt in der BRD 1983 bis 1986 aus Daten des Sozio-ökonomischen Panels

Wenn man einen Wechsel der Wohlstandsposition so definiert, dass mehr als 10 % Auf- oder Abstieg eintreten muß, dann haben 1984 zu 1983 mehr als 50 % ihre Wohlstandsposition verändert.

Keyconcepts

Häufigkeits-/Verteilungsfunktion
Modus, Median
Arithmetrisches Mittel, Geometrisches Mittel, Harmonisches Mittel
Spannweite
p-Quantile
Mittlere absolute Abweichung
Varianz und Standardabweichung
Konzept von Momente, Schiefe und Exzeß
Konzentration (Lorenzkurve, Gini-Koeffizient)

III Statistische Analyse mehrerer Merkmale

Analyse mehrerer Merkmale, Messung von Zusammenhängen mittels Korrelationsrechnung und Regressionsrechnung

Mehrdimensionale Betrachtung gemeinsam auftretender Merkmale zur Analyse der Zusammenhänge zwischen mehreren sozioökonomischen Merkmalen einer statistischen Masse

Sozioökonomische Merkmale einer Person
Arbeitszeit, Einkommen, Alter, Geschlecht, Haushaltsgröße, Haushaltszusammensetzung etc.

Zwei zentrale Fragestellungen:

- Wie stark ist der Zusammenhang zwischen den Variablen?

  Beispiele:

  Privater Konsum - Volkseinkommen über die Jahre
  Gewinn - Umsatz über Betriebe
  Benzinverbrauch - Geschwindigkeit über Meßpunkte

  → **Korrelationsrechnung** (nicht gerichtete Analyse)

- Gerichtete Analyse: In welcher funktionalen Weise können die Abhängigkeiten zwischen den Variablen beschrieben werden? → Kausalanalyse

  Beispiele:

  Privater Konsum = f(Volkseinkommen)
  Arbeits(zeit)angebot = f(Lohnsatz, Alter, Ausbildung,...)
  Gewinn = f(Werbung, Marktkonzentration,...)

  → **Regressionsrechnung**

1 Zweidimensionale Häufigkeitsverteilungen und ihre Darstellung

1.1 Allgemeine Grundbegriffe und Darstellungsweisen

Für jede statistische Enheit (wie z.B. Person, Haushalt, Betrieb) werden die Merkmalsausprägungen von **zwei Merkmalen x und y** (z.B. Alter x und Geschlecht y einer Person) erhoben und tabellarisch dargestellt: → Zweidimensionale Tabelle

Jeder Kombination von (x_i, y_j) wird die Anzahl (absolute Häufigkeit) n_{ij} zugeordnet:

$$n(x_i, y_j) = n_{ij}$$

x_i und y_i könne dabei nominal, ordinal oder metrisch skaliert sein.

Allgemeiner Aufbau einer zweidimensionalen Tabelle

	y_1	y_2	$\cdots$	y_j	$\cdots$	y_m	Zeilen-summen
x_1	n_{11}	n_{12}	$\cdots$	n_{1j}	$\cdots$	n_{1m}	$n_{1\cdot}$
x_2	n_{21}	n_{22}	$\cdots$	n_{2j}	$\cdots$	n_{2m}	$n_{2\cdot}$
$\vdots$	$\vdots$	$\vdots$		$\vdots$		$\vdots$	$\vdots$
x_i	n_{i1}	n_{i2}	$\cdots$	n_{ij}	$\cdots$	n_{im}	$n_{i\cdot}$
$\vdots$	$\vdots$	$\vdots$		$\vdots$		$\vdots$	$\vdots$
x_k	n_{k1}	n_{k2}	$\cdots$	n_{kj}	$\cdots$	n_{km}	$n_{k\cdot}$
Spalten-summen	$n_{\cdot 1}$	$n_{\cdot 2}$	$\cdots$	$n_{\cdot j}$	$\cdots$	$n_{\cdot m}$	$n(=n_{\cdot\cdot})$

Absolute Häufigkeit

$$n(x_i, y_j) = n_{ij}$$

Relative Häufigkeit

$$h(x_i, y_j) = \frac{n_{ij}}{n}$$

Beispiel:

Privathaushalte in der BRD (Mai 1987) nach Familienstand (x) und Geschlecht (y) des Haushaltsvorstandes (HHV) (in 1000)

Tab. IV.1: Privathaushalte in der BRD (Mai 1987) nach Familienstand (x) und Geschlecht (y) des Haushaltsvorstandes (HHV) (in 1000)

		Geschlecht des HHV (y)		Zeilensumme
Familienstand	(x)	männlich (y_1)	weiblich (y_2)	von x_i: $\sum_{j=1}^{2} n_{ij}$
ledig	(x_1)	2.755	2.403	5.158
verheiratet	(x_2)	14.929	680	15.609
verwitwet	(x_3)	694	3.988	4.682
geschieden	(x_4)	744	1.211	1.955
Spaltensumme von y_j: $\sum_{i=1}^{4} n_{ij}$		19.122	8.282	27.404

Quelle: Statistisches Jahrbuch 1989, S. 56, Volkszählung 1987

$n_{32} = 3.988, \quad n_{2\cdot} = 15.609, \quad n_{\cdot 1} = 19.122, \quad n = 27.404$

n_{23} = gibt es nicht!

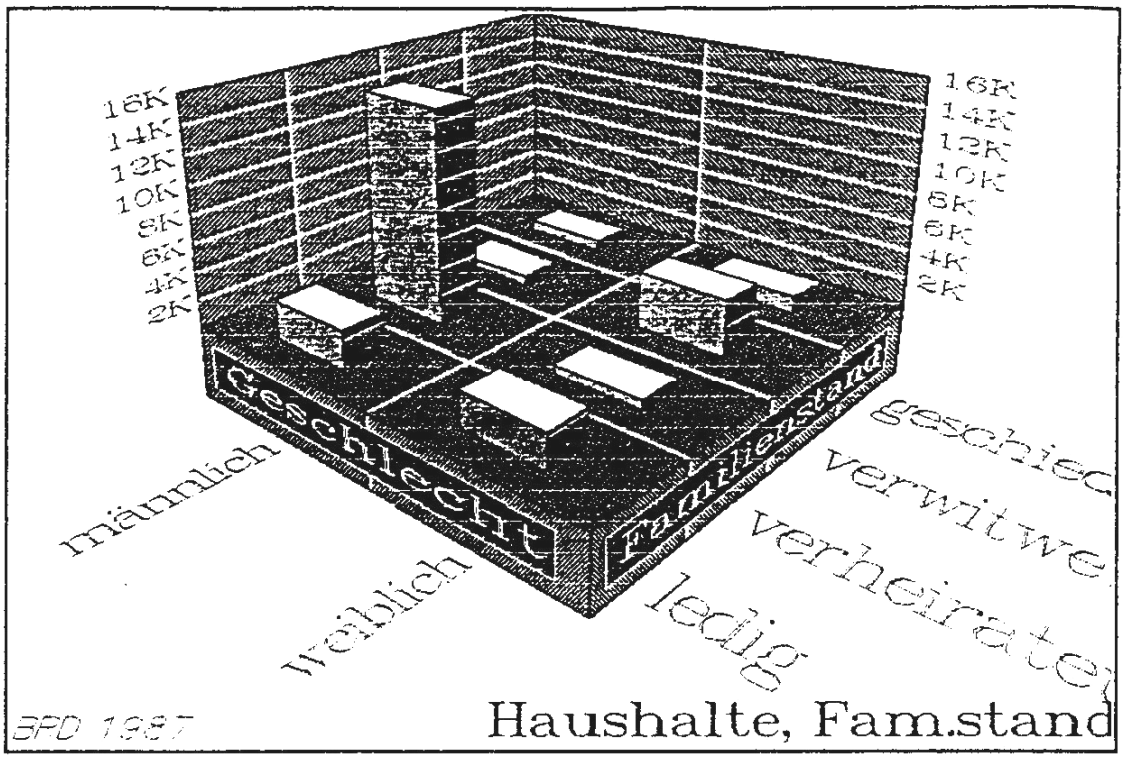


Abb. IV.1: Privathaushalte in der BRD (Mai 1987) nach Familienstand und Geschlecht des Haushaltsvorstands (vgl. Tab. IV.1)

1.2 Randverteilungen

Randverteilungen = eindimensionale Verteilungen eines Merkmals, marginale Häufigkeiten

Randverteilung des 1. Merkmals (x) (Zeilensummen)

$$n_{i\cdot} = \sum_{j=1}^{m} n_{ij} \qquad \text{(marginale absolute Häufigkeiten)}$$

$$h_{i\cdot} = \frac{n_{i\cdot}}{n} \qquad \text{(marginale relative Häufigkeiten)}$$

Randverteilungen des 2. Merkmals (y) (Spaltensummen)

$$n_{\cdot j} = \sum_{i=1}^{k} n_{ij} \qquad \text{(marginale absolute Häufigkeiten)}$$

$$h_{\cdot j} = \frac{n_{\cdot j}}{n} \qquad \text{(marginale relative Häufigkeiten)}$$

Beispiel:

Privathaushalte BRD 1987 nach Familienstand (x) und Geschlecht (y) (vgl. Tab IV.1)

Familienstand:

$n_{1.} = 5158$	$h_{1.} = 0,1882$
$n_{2.} = 15609$	$h_{2.} = 0,5696$ *
$n_{3.} = 4682$	$h_{3.} = 0,1709$
$n_{4.} = 1955$	$h_{4.} = 0,0713$
$n = 27404$	$h_{..} = 1,0$

*56,96% sind verheiratet

Geschlecht:

$n_{.1} = 19122$	$h_{.1} = 0,6978$
$n_{.2} = 8282$	$h_{.2} = 0,3022$
$n = 27404$	$h_{..} = 1,0$

1.3 Bedingte Verteilungen

Neben des Bezugs auf die Gesamtsumme aller Merkmalsausprägungen n auch Bezüge auf die jeweilige Randverteilung

Bedingte Verteilung

Bezug auf die Gesamtheit **einer** Zeile oder Spalte.

Bei gegebener Ausprägung des einen Merkmals werden den Ausprägungen des anderen Merkmals relative Häufigkeiten zugeordnet:

eindimensionale Häufigkeitsverteilung von x bei gegebenem y_j, wobei $x_i = \text{Wirkung}$, $y_j = \text{Ursache}$:

$$h\left(x_i \middle| y_j\right) = \frac{n\left(x_i, y_j\right)}{n\left(y_j\right)} = \frac{n_{ij}}{n_{.j}}$$

eindimensionale Häufigkeitsverteilung von y bei gegebenem x_i, wobei $y_j = \text{Wirkung}$, $x_i = \text{Ursache}$:

$$h\left(y_j \middle| x_i\right) = \frac{n\left(x_i, y_j\right)}{n\left(x_i\right)} = \frac{n_{ij}}{n_{i.}}$$

Die bedingten Verteilungen liefern nur dann mehr Informationen, wenn sie sich unterscheiden. Sind alle bedingten Verteilungen gleich, dann sind sie auch identisch mit der Randverteilung.

Statistische Unabhängigkeit

Merkmale sind statistisch voneinander unabhängig, wenn alle bedingten Verteilungen **gleich** der entsprechenden Randverteilung sind. Die Verteilung des Merkmals x ist dann unabhängig von spezieller Ausprägung des Merkmals y:

$$h(x_i|y_1) = h(x_i|y_2) = ... = h(x_i|y_m) = h(x_i) = \frac{n_{i\cdot}}{n}$$

Weiter gilt bei statistischer Unabhängigkeit:

$$\frac{n_{ij}}{n_{\cdot j}} = \frac{n_{i\cdot}}{n} \rightarrow n_{ij} = \frac{n_{i\cdot} \cdot n_{\cdot j}}{n}$$

$$\rightarrow \quad \frac{n_{ij}}{n} = \frac{n_{i\cdot}}{n} \cdot \frac{n_{\cdot j}}{n} \quad \text{oder}$$

$$h(x_i, y_j) = h(x_i) \cdot h(y_j)$$

Aussagen

1. Ist x von y statistisch unabhängig, so ist auch y von x statistisch unabhängig (symmetrische Beziehung).
2. Sind x und y statistisch unabhängig, so sind die bedingten Häufigkeitsverteilungen gleich der zugehörigen Randverteilung.

Beispiel:

Wohnbevölkerung der BRD (1987) nach der Beteiligung am Erwerbsleben und dem Geschlecht (in 1.000)

Tab. IV.2: Wohnbevölkerung der BRD (1987) nach der Beteiligung am Erwerbsleben und dem Geschlecht (in 1.000), Häufigkeiten (n_{ij})

Beteiligung am Erwerbsleben x	Geschlecht y männlich	weiblich	Zeilensumme
Erwerbspersonen	17.834	11.160	28.994
Nicht-Erwerbspersonen	11.489	20.594	32.083
Spaltensumme	29.323	31.754	61.077

Quelle: Statistisches Jahrbuch der BRD 1989, S. 89, Volkszählung 1987

n = 61.077, n_{22}=20.594 (weibliche Nicht-Erwerbspersonen)

Häufigkeiten $h(x_{i,}y_j)$ (in %):

Beteiligung am Erwerbsleben x	Geschlecht y männlich	weiblich	Zeilensumme
Erwerbspersonen	29,2	18,3	47,5
Nicht-Erwerbspersonen	18,8	33,7	52,5
Spaltensumme	48,0	52,0	100,0

$h(x_2, y_1) = \frac{n_{21}}{n} = 18{,}8\ \%$ sind männliche Nicht-Erwerbspersonen.

$h(x_1) = \frac{n_{1\cdot}}{n} = 47{,}5\ \%$ der Wohnbevölkerung sind Erwerbspersonen.

Bedingte Verteilung:

Frage: Wie gliedert sich die Beteiligung der männlichen und weiblichen Personen am Erwerbsleben auf? $\rightarrow h(x_i \,|\, Geschlecht)$

von 100 Personen, die das Geschlecht ... hatten, waren	Geschlecht y männlich	weiblich	Randverteilung
Erwerbspersonen	60,8	35,2	47,5
Nicht-Erwerbspersonen	39,2	64,8	52,5
Spaltensumme	100,0	100,0	100,0

$$h(x_2 \,|\, y_1) = h(Nicht-Erwerbspersonen \,|\, alle\ Männer) = \frac{n_{21}}{n_{\cdot 1}} = 39{,}2\ \%$$

$$h(x_1 \,|\, y_2) = h(Erwerbspersonen \,|\, alle\ Frauen) = \frac{n_{12}}{n_{\cdot 2}} = 35{,}2\ \%$$

Frage: Wie gliedert sich der Erwerbsstatus auf das Geschlecht auf? $\rightarrow (h \,|\, y_j \,|\, Erwerbsstatus)$

von 100 Personen, die den Erwerbsstatus ... hatten, waren	Geschlecht y männlich	weiblich	Zeilensumme
Erwerbspersonen	61,5	38,5	100,0
Nicht-Erwerbspersonen	35,8	64,2	100,0
Randverteilung	48,0	52,0	100,0

$$h(y_1 \,|\, x_1) = h(Männer \,|\, alle\ Erwerbspersonen) = \frac{n_{11}}{n_{1\cdot}} = 61{,}5\ \%$$

$$h(y_2 \,|\, x_2) = h(Frauen \,|\, alle\ Nichterwerbspersonen) = \frac{n_{22}}{n_{2\cdot}} = 64{,}2\ \%$$

Wäre die Beteiligung am Erwerbsleben unabhängig vom Geschlecht, so müßte die Randverteilung gleich den prozentualen Besetzungszahlen für die einzelnen Geschlechter sein.

Fiktive Verteilung der Beteiligung am Erwerbsleben bei Unabhängigkeit vom Geschlecht:

von 100 Personen, die das	Geschlecht y		
Geschlecht ... hatten, waren	männlich	weiblich	Randverteilung
Erwerbspersonen	47,5	47,5	47,5
Nicht-Erwerbspersonen	52,5	52,5	52,5
Spaltensumme	100,0	100,0	100,0

Statistisch unabhängig?

$$h(x_i|y_1) = h(x_i|y_2) = h(x_i) \overset{?}{=} \frac{n_{i\cdot}}{n} \quad (i = 1,2)$$

$$h(x_1|y_1) = 60{,}8\,\% \quad \neq \quad \tilde{h}(x_1|y_1) = 47{,}5\,\% \quad \Rightarrow$$

Die Beteiligung am Erwerbsleben ist **nicht unabhängig** vom Geschlecht.

---

2 Korrelationsrechnung

2.1 Zusammenhangsmaße

Zusammenhang zwischen den Merkmalsausprägungen verschiedener Merkmale; aus der Vielzahl der Maßzahlen des Zusammenhangs (Kontingenz, Assoziation, Korrelation) werden folgende drei Konzepte näher betrachtet:
Korrelation zwischen

- nominal skalierten Merkmalen
  Kontingenzanalyse und Kontingenzkoeffizient
- ordinal skalierten Merkmalen
  Rangkorrelation nach Spearman
- metrisch skalierten Merkmalen
  Bravais-Pearson-Korrelationskoeffizient

2.2 Korrelation zwischen nominal skalierten Merkmalen: Kontingenzanalyse und Kontingenzkoeffizient

Wie die Analyse im letzten Abschnitt gezeigt hat, gilt bei Abhängigkeit der Merkmale

$$n_{ij} \neq \frac{n_{i\cdot} n_{\cdot j}}{n} \ .$$

Damit besteht eine Differenz zwischen beobachteter Häufigkeit n_{ij} und fiktiver Häufigkeit bei Unabhängigkeit $\tilde{n}_{ij}$:

$$\tilde{n}_{ij} = \frac{n_{i\cdot} n_{\cdot j}}{n}$$

Quadratische Kontingenz χ^2 (Chi-Quadrat)

$$\chi^2 = \sum_{i=1}^{k} \sum_{j=1}^{m} \frac{\left(n_{ij} - \tilde{n}_{ij}\right)^2}{\tilde{n}_{ij}},$$

wobei

n_{ij} = beobachtete Häufigkeit

$\tilde{n}_{ij}$ = Häufigkeit bei Unabhängigkeit

$\chi^2 = 0 \Rightarrow$ statistischeUnabhängigkeit

Nachteil: χ^2 ist nicht normiert und kann daher unbegrenzt große Werte annehmen. Deshalb verwendet man besser:

Kontingenzkoeffizient K^*

$$K^* = \sqrt{\frac{\chi^2}{n + \chi^2} \cdot \frac{M}{M-1}},$$

wobei

M = min (k, m) mit

k = Zeilenanzahl

m = Spaltenanzahl

Wertebereich: $0 \le K^* \le 1$, damit ist K^* normiert zwischen 0 und 1.

Beispiel:

Gliederung der Wohnbevölkerung der BRD 1987 nach Beteiligung am Erwerbsleben und nach Geschlecht:

$$\tilde{n}_{11} = \frac{n_{1.} n_{.1}}{n} = \frac{28.994 \cdot 29.323}{61.077} = 13.920$$

$$\tilde{n}_{12} = \frac{n_{1.} n_{.2}}{n} = \frac{28.994 \cdot 31.754}{61.077} = 15.074$$

$$\tilde{n}_{21} = \frac{n_{2.} n_{.1}}{n} = \frac{32.083 \cdot 29.323}{61.077} = 15.403$$

$$\tilde{n}_{22} = \frac{n_{2.} n_{.2}}{n} = \frac{32.083 \cdot 31.754}{61.077} = 16.680$$

Kontingenztabellen

Beobachtete absolute Häufigkeit n_{ij}:

Beteiligung am Erwerbsleben n_{ij}	Geschlecht männlich	weiblich	Zeilensumme
Erwerbspersonen	17.834	11.160	28.994
Nicht-Erwerbspersonen	11.489	20.594	32.083
Spaltensumme	29.323	31.754	61.077

Fiktive absolute Häufigkeit Ausgang bei statistischer Unabhängigkeit $\tilde{n}_{ij}$:

Beteiligung am Erwerbsleben n_{ij}	Geschlecht männlich	weiblich	Zeilensumme
Erwerbspersonen	13.920	15.074	28.994
Nicht-Erwerbspersonen	15.403	16.680	32.083
Spaltensumme	29.323	31.754	61.077

Für k = 2 und m = 2:

$$\chi^2 = \sum_{i=1}^{2}\sum_{j=1}^{2}\frac{\left(n_{ij}-\tilde{n}_{ij}\right)^2}{\tilde{n}_{ij}} = 1100,53+1016,28+994,57+918,43 = 4029,81$$

Für M = min (2,2) = 2:

$$K^* = \sqrt{\frac{\chi^2}{n+\chi^2}\cdot\frac{M}{M-1}} = \sqrt{\frac{4029,81}{61077+4029,81}\cdot\frac{2}{2-1}} = \sqrt{0,123791} = 0,35 \qquad (0 \le K^* \le 1)$$

Interpretation: Beteiligung am Erwerbsleben ist geschlechtsabhängig (Vergleich mit anderen Jahren, Ländern ist aussagekräftiger).

2.3 Korrelation zwischen ordinal-skalierten Merkmalen: Rangkorrelationskoeffizient nach Spearman

Ordinal-skalierte Merkmale, daher Rangnummern (-plätze)

R_i: Rangnummern des 1. Merkmals (x)

R_i': Rangnummern des 2. Merkmals (y)

Jeder statistischen Einheit i $(i = 1,...,n)$ werden Rangnummern beider Merkmalsausprägungen R_i und R_i' zugeordnet. Gibt es übereinstimmende Beobachtungswerte für mehrere statistische Einheiten, dann wird hier das arithmetische Mittel der diesen Werten zuzuordnenden Nummern als Rangzahl zugeordnet.

Spearman'scher Rangkorrelationskoeffizient

$$r_{Sp} = 1 - \frac{6\sum_{i=1}^{n} \left(R_i - R_i'\right)^2}{(n-1)n(n+1)},$$

Normiert auf $\left[-1 \le r_{Sp} \le +1\right]$

$r_{Sp} = -1$: Die Ränge verhalten sich völlig gegenläufig.
$\left(R_i = n+1-R_i';\ \ i=1,...,n\right)$;
negativer Zusammenhang

$r_{Sp} = 0$: kein Zusammenhang

$r_{Sp} = +1$: Die Ränge verhalten sich völlig gleichläufig.
$\left(R_i = R_i';\ \ i=1,...,n\right)$
positiver Zusammenhang

Beispiel:

Für zehn Angestellte wurden organisatorische Geschicklichkeit (x) und Arbeitssorgfalt (y) ermittelt.

Rangziffern (-plätze):

Angestellter i		1	2	3	4	5	6	7	8	9	10
x:	R_i	7	3	9	10	1	5	4	6	2	8
y:	R_i'	3	9	10	8	7	1	5	4	2	6

$$r_{Sp} = 1 - \frac{6\sum_{i=1}^{10} \left(R_i - R_i'\right)^2}{(10-1)\cdot 10\cdot(10+1)} = 1 - \frac{6\cdot 118}{9\cdot 10\cdot 11} = 0,28$$

Interpretation: Es liegt eine schwach gleichläufige (positive) Korrelation vor.

2.4 Korrelation zwischen metrisch-skalierten Merkmalen: Bravais-Pearson-Korrelationskoeffizient

Metrisch-skalierte Merkmale in ungruppierter und gruppierter Form

Ungruppiertes Datenmaterial:

Gegeben: n Beobachtungspaare von Merkmalsausprägungen

$$(x_1, y_1), (x_2, y_2), (x_3, y_3), ..., (x_n, y_n)$$

Mittelwerte

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

$$\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$$

Varianzen

$$s_x^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

$$s_y^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2$$

Kovarianz

$$s_{xy} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})\cdot(y_i - \bar{y}) = \frac{1}{n}\sum_{i=1}^{n} x_i \cdot y_i - \bar{x}\cdot\bar{y} = \overline{xy} - \bar{x}\cdot\bar{y}$$

Die Kovarianz stellt eine Beziehung zwischen x und y her (vgl. Abb. IV.2):

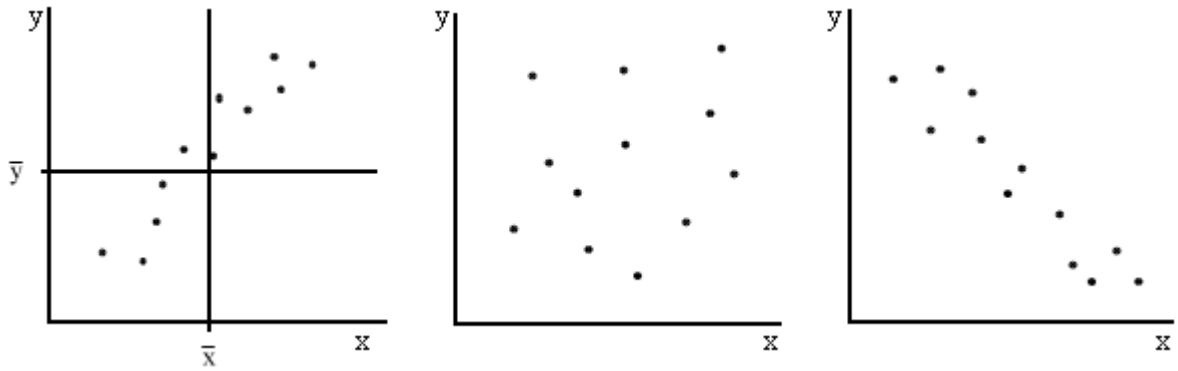


Abb. IV.2: Kovarianzen

Kreuzprodukte: $(x_i - \bar{x})(y_i - \bar{y})$ positiv, ≈ 0 oder negativ

s_{xy} wird umso größer, je stärker die Wertepaare überwiegen, bei denen große x-Werte mit großen y-Werten und kleine x- mit kleinen y-Werten gekoppelt sind.

Die Kovarianz ist ein Maß für die Stärke des Zusammenhangs zwischen zwei Merkmalen x und y. Normiert ergibt sich:

Bravais-Pearson-Korrelationskoeffizient

$$r = \frac{s_{xy}}{s_x s_y}$$

(Normierung auf das Produkt der Standardabweichungen)

r liegt im Intervall $[-1 \le r \le +1]$

Interpretation:

$r = -1$	extrem starker negativer Zusammenhang
$r = 0$	keine Korrelation
$r = +1$	extrem starker positiver Zusammenhang

Beispiel:

Umsatz und Werbeausgaben eines Industrieunternehmens

Tab. IV.3: Umsatz und Werbeausgaben der Firma IXWHYZET

Jahr	Umsatz y (Mio. EUR)	Werbeausgaben x (Mio. EUR)
1996	17,0	1,4
1997	17,6	1,7
1998	17,5	1,6
1999	18,1	1,8
2000	18,7	2,0
2001	19,1	1,9
2002	19,0	2,0
2003	20,5	2,2
2004	21,8	2,0
2005	21,3	2,1
2006	26,5	2,5
2007	25,8	3,0
2008	26,3	2,8
2009	27,8	3,2
2010	30,0	3,0

Werbeausgaben: $\bar{x} = 2,21 \quad s_x^2 = 0,29$

Umsatz: $\bar{y} = 21,80 \quad s_y^2 = 17,40$

Kovarianz: $s_{xy} = 2,20$

Korrelationskoeffizient: $r = \frac{2,20}{0,54 \cdot 4,17} = 0,977$

Es besteht also ein starker positiver Zusammenhang, d.h. hohe (geringe) Werbeausgaben korrelieren mit hohen (geringen) Umsätzen.

ET: Zusammenhangsanalyse
ET: Histogram, Scatter

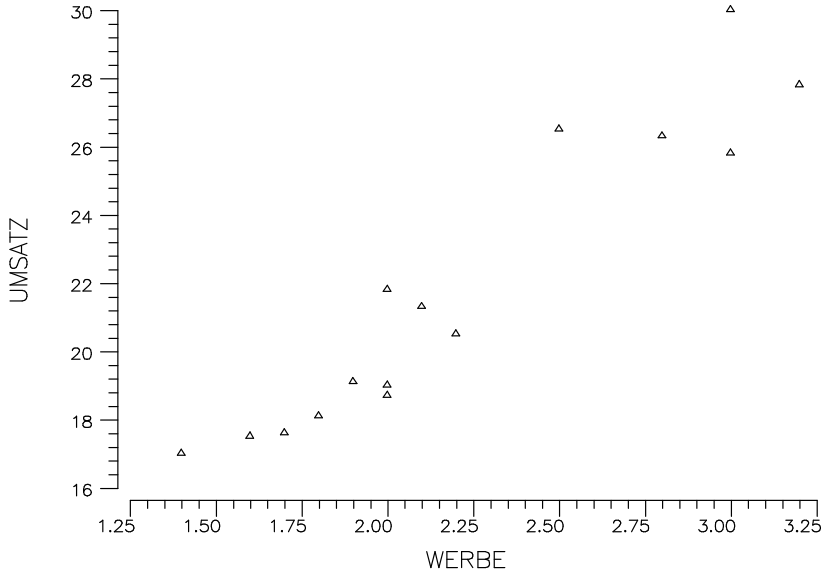


Abb. IV.3: ET: Scatterplot der Umsätze und Werbeausgaben der Firma IXWHYZET

ET: Histogram, Crosstab

```
        WERBE       Crosstabulation    Chi-squared[  0]=   .0000, P= .00000
UMSATZ      0     1     2     3     4    Total
  0         3     4     0     0     0     7
  1         0     2     1     0     0     3
  2         0     0     0     0     0     0
  3         0     0     0     2     1     3
  4         0     0     0     0     2     2
Total       3     6     1     2     3    15


Classes          UMSATZ  /Out of range=   0     WERBE   /Out of range=   0
  0              16.9900 to     19.6000          1.3900 to      1.7600
  1              19.6000 to     22.2000          1.7600 to      2.1200
  2              22.2000 to     24.8000          2.1200 to      2.4800
  3              24.8000 to     27.4000          2.4800 to      2.8400
  4              27.4000 to     30.0100          2.8400 to      3.2100
```

Gruppiertes Datenmaterial:

Die Daten werden in Klassen eingeteilt und die Klassenmitten x_i^* und y_i^* ermittelt.

Mittelwerte

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{k} x_i^* \cdot n_{i\cdot}$$

$$\overline{y} = \frac{1}{n}\sum_{j=1}^{m} y_j^* \cdot n_{\cdot j}$$

Varianzen

$$s_x^2 = \frac{1}{n}\sum_{i=1}^{k}\left(x_i^* - \overline{x}\right)^2 \cdot n_{i\cdot} = \frac{1}{n}\sum_{i=1}^{k}\left(x_i^*\right)^2 \cdot n_{i\cdot} - \overline{x}^2$$

$$s_y^2 = \frac{1}{n}\sum_{j=1}^{m}\left(y_j^* - \overline{y}\right)^2 \cdot n_{\cdot j} = \frac{1}{n}\sum_{j=1}^{m}\left(y_j^*\right)^2 \cdot n_{\cdot j} - \overline{y}^2$$

Kovarianz

$$s_{xy} = \frac{1}{n}\sum_{i=1}^{k}\left(x_i^* - \overline{x}\right)(y_i^* - \overline{y}) \cdot n_{i\cdot} = \frac{1}{n}\sum_{i=1}^{k} x_i^* y_i^* \cdot n_{i\cdot} - \overline{xy}$$

Korrelationskoeffizient

$$r = \frac{s_{xy}}{s_x \cdot s_y} \qquad -1 \le r \le 1$$

Beispiel:

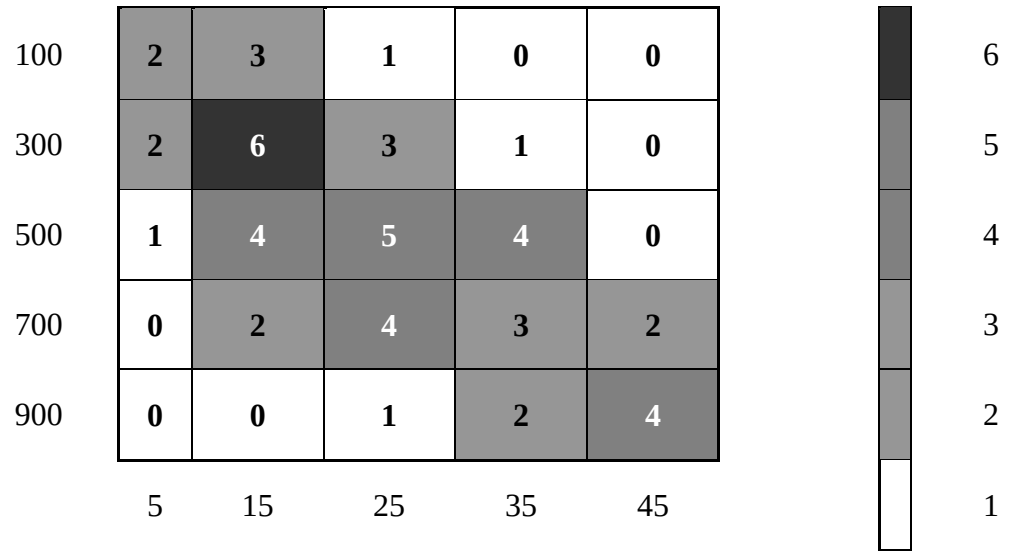
Aufwendungen für Forschung und Entwicklung FuE (R&D = RESEARCH AND DEVELOPMENT) von Unternehmen (Tabelle IV.4)

Tab. IV.4: Aufwendungen für Forschung und Entwicklung (FuE) von Unternehmen

Umsatz x	Aufwendungen für Forschung und Entwicklung (FuE) y: 5 (y_1^*)	15 (y_2^*)	25 (y_3^*)	35 (y_4^*)	45 (y_5^*)	Zeilensumme
100 (x_1^*)	2	3	1	-	-	$6 = n_{1\cdot}$
300 (x_2^*)	2	6	3	1	-	$12 = n_{2\cdot}$
500 (x_3^*)	1	4	5	4	-	$14 = n_{3\cdot}$
700 (x_4^*)	-	2	4	3	2	$11 = n_{4\cdot}$
900 (x_5^*)	-	-	1	2	4	$7 = n_{5\cdot}$
Spaltensumme	5 $n_{\cdot 1}$	15 $n_{\cdot 2}$	14 $n_{\cdot 3}$	10 $n_{\cdot 4}$	6 $n_{\cdot 5}$	$50 = n$

Aufwendungen für FuE

(in 10.000 DM)



Quelle: Tabelle FuE

Abb. IV.4: Aufwendungen für FuE: Spektraldarstellung

$$\bar{x} = \frac{1}{50}\left(100 \cdot 6 + 300 \cdot 12 + 500 \cdot 14 + 700 \cdot 11 + 900 \cdot 7\right) = 504$$

$$\bar{y} = \frac{1}{50}\left(5 \cdot 5 + 15 \cdot 15 + 25 \cdot 14 + 35 \cdot 10 + 45 \cdot 6\right) = 24,4$$

$$s_x^2 = \frac{1}{50}\left(100^2 \cdot 6 + 300^2 \cdot 12 + 500^2 \cdot 14 + \cdot 700^2 \cdot 11 + 900^2 \cdot 7\right) - \left(504^2\right) = 59984$$

$$s_x = 244,916$$

$$s_y^2 = \frac{1}{50}\left(5^2 \cdot 5 + 15^2 \cdot 15 + 25^2 \cdot 14 + 35^2 \cdot 10 + 45^2 \cdot 6\right) - \left(24,4^2\right) = 137,64$$

$$s_y = 11,732$$

$$s_{xy} = 1/50\ (100 \cdot 5 \cdot 2 + 100 \cdot 15 \cdot 3 + 100 \cdot 25 \cdot 1 + 300 \cdot 5 \cdot 2 + 300 \cdot 15 \cdot 6 + 300 \cdot 25 \cdot 3 + 300 \cdot 35 \cdot 1 +$$
$$500 \cdot 5 \cdot 1 + 500 \cdot 15 \cdot 4 + 500 \cdot 25 \cdot 5 + 500 \cdot 35 \cdot 4 + 700 \cdot 15 \cdot 2 + 700 \cdot 25 \cdot 4 + 700 \cdot 35 \cdot 3 +$$
$$700 \cdot 45 \cdot 2 + 900 \cdot 25 \cdot 1 + 900 \cdot 35 \cdot 2 + 900 \cdot 45 \cdot 4) - 504 \cdot 24,4 = 1922,4$$

$$r = \frac{1922,4}{244,916 \cdot 11,732} = 0,669$$

Hier liegt also eine positive Korrelation vor.

Keyconcepts

Mehrdimensionale Verteilung
Randverteilungen
Bedingte Verteilungen
Kontingenzanalyse und Kontingenzkoeffizient
Rangkorrelationskoeffizient nach Spearman
Bravais-Pearson-Korrelationskoeffizient

Statistik für alle

Übungs- und Klausuraufgaben mit Lösungen

A Übungsaufgaben mit Lösungen

Aufgabenblatt 1: Allgemeine Grundlagen

1 Im Rahmen der Statistik I-Vorlesung findet seit letztem Semester regelmäßig eine Umfrage
zu Wohnsituation der Studierenden statt. In dieser werden die Teilnehmer der Vorlesung zu Themen wie ihrer Wohnform, der Größe ihrer Wohnung und ähnlichem befragt.
Klären Sie für diese Analyse die folgenden Begriffe:

- statistische Einheit (bzw. Merkmalsträger),
- Merkmal (beispielhaft)
- Merkmalsausprägung (beispielhaft)

2 Geben Sie die Skalierung der folgenden Merkmale an:

a) Geschlecht
b) Studiengang (BWL, Uwi, Kuwi, etc.)
c) Wohnform (bspw. WG, Einzelwohnung, etc.)
d) Entfernung der Wohnung von der Universität (drei Ausprägungen: „nah“, „weit“, „sehr weit“)
e) Wohnungsgröße (in m²)
f) Zahl der Zimmer
g) Miete (in EUR)
h) Miete (drei Ausprägungen: „niedriger als die Durchschnittsmiete“, „gleich der Durchschnittsmiete“, „höher als die Durchschnittsmiete“)
i) Miete pro m² Wohnfläche
j) Durchschnittliche Zimmergröße
k) Zahl der Mitbewohner
l) Zufriedenheit mit der Wohnung (drei Ausprägungen „sehr zufrieden“, „geht so“, „bin weg, sobald ich *irgendetwas* anderes finde“)
m) Durchschnittliche Temperatur im Sommer (in °C)

3 Welche der in Aufgabe 2 genannten Merkmale sind stetig bzw. diskret?

4 Üben Sie den Umgang mit Summenzeichen anhand folgender Beispiele:

i	**1**	**2**	**3**	**4**	**5**
x_i	2	4	6	8	10
y_i	20	40	80	160	320

a) $\sum_{i=1}^{5} x_i$ b) $\sum_{j=2}^{4} y_j$

c) $\sum_{j=1}^{n} x_2 y_j$ d) $\sum_{i=1}^{5} x_i + y_2$

e) $\sum_{i=1}^{5} (2x_i + 2)$ f) $\sum_{i=1}^{5} \sum_{j=1}^{2} x_i y_j$

5 Im Folgenden finden Sie wirtschaftliche Daten eines fiktiven Landes geteilt nach Männern und Frauen sowie nach den beiden Regierungsbezirken Norden und Süden.

Erwerbstätige; In Klammern: abhängige Erwerbstätige	Männer (Bevölkerung: 40,0 Mio.)	Frauen (Bevölkerung: 44,0 Mio.)
Norden (Fläche: 0,9 Mio. km²)	13,0 Mio. (80%)	8,0 Mio. (90%)
Süden (Fläche: 1,2 Mio. km²)	17,0 Mio. (85%)	9,2 Mio. (95%)

Erwerbslose; In Klammern: davon Arbeitslose	Männer	Frauen
Norden	1,3 Mio. (75%)	0,75 Mio. (30%)
Süden	1,36 Mio. (85%)	1,0 Mio. (40%)

a) Berechnen und interpretieren Sie das Geschlechterverhältnis.
b) Berechnen und interpretieren Sie die Bevölkerungsdichte.
c) Berechnen und interpretieren Sie die allgemeine Erwerbsquote für die Männer.[1]
d) Wie viele abhängige Erwerbstätige gibt es im Süden?

6 Ordnen Sie folgende Personen in das Erwerbskonzept ein:

a) Peter Neururer, 50 Jahre, bis Ende Mai 2005 Trainer des Vfl Bochum, sucht eine neue Stelle als Bundesligatrainer, nicht bei der Agentur für Arbeit gemeldet
b) Stefan M., 27, nach abgeschlossenem Philosophie und Anthropologiestudium mit einem Zeitvertrag über drei Jahre bei einer Tageszeitung als Sportreporter beschäftigt
c) Herrmann K., 41, Angestellter der Stadt Hamburg, seit 4 Monaten mit schwerer Grippe im Krankenhaus, angeblich kein Simulant
d) Josef Ackermann, 57, Sprecher des Vorstands der Deutschen Bank AG
e) Paul R, 40, ehemaliger Bäcker, wegen einer Mehlstauballergie berufsunfähig, bezieht Arbeitslosengeld II und sucht über die Agentur für Arbeit eine Stelle als Lagerarbeiter
f) Peter R, 40, Bäcker in einer Großbäckerei, möchte sich beruflich verändern und sucht privat und über die Agentur für Arbeit eine Stelle als Lagerarbeiter
g) Hannelore B., 34, Hausfrau, macht in der Abendschule ihr Abitur nach

[1] Erwerbsquote = Erwerbspersonen / Wohnbevölkerung
Erwerbspersonen = Erwerbstätige + Erwerbslose

Lösungen zu Aufgabenblatt 1:
Grundlagen, Wirtschafts- und Sozialstatistik

1

- statistische Einheit/Merkmalsträger: der einzelne (befragte) Student
- Merkmal: bspw. die Wohnform
- Ausprägungen (gleiches Beispiel): WG, Einzelwohnung, Wohnheimzimmer, bei Eltern…

2 und 3

a) Geschlecht ⇒ männlich, weiblich ⇒ NOMINAL (diskret)
b) Studiengang ⇒ NOMINAL (diskret)
c) Wohnform ⇒ WG etc. ⇒NOMINAL (diskret)
d) Entfernung zu Uni ⇒ 3 Kategorien ⇒ ORDINAL (diskret)
e) Wohnungsgröße ⇒ VERHÄLTNIS (stetig)
f) Zahl der Zimmer ⇒ ABSOLUT (diskret) natürliche Einheit und Nullpunkt
g) Miete (in EUR) ⇒ VERHÄLTNIS (appr. stetig), natürlicher Nullpunkt
h) Miete (3 Ausprägungen) ⇒ ORDINAL (diskret)
i) Miete pro m² ⇒ VERHÄLTNIS (approximativ stetig)
j) Durchschnittliche Zimmergröße ⇒ VERHÄLTNIS (stetig)
k) Zahl der Mitbewohner ⇒ ABSOLUT (diskret)
l) Zufriedenheit ⇒ ORDINAL (diskret)
m) Durschnittstemperatur ⇒ INTERVALL (stetig)

4 Summenzeichen

i/j	**1**	**2**	**3**	**4**	**5**
x_i	2	4	6	8	10
y_j	20	40	80	160	320

a) $2+4+6+8+10=30$
b) $40+80+160=280$
c) $4\cdot 20+4\cdot 40+4\cdot 80+4\cdot 160+4\cdot 320=4\cdot(20+40+80+160+320)=2480$
d) $30+40=70$ (Hinweis: 30 ist das Ergebnis aus a))
e) $6+10+14+18+22=70$
f)
$(2\cdot 20+2\cdot 40)+(4\cdot 20+4\cdot 40)+(6\cdot 20+6\cdot 40)+(8\cdot 20+8\cdot 40)+(10\cdot 20+10\cdot 40)=1800$

5

a) $GV = \frac{Anzahl \quad Frauen}{Anzahl \quad Männer} = \frac{44 \quad Mio.}{40 \quad Mio.} = 1,1$

$\Rightarrow$ auf 1000 Männer kommen 1100 Frauen

b) $BD = \frac{Bevölkerung}{Fläche} = \frac{84 \quad Mio.}{1,2 \; Mio.\, km^2 + 0,9 \quad Mio.\, km^2} = 40$ Personen/km²

$\Rightarrow$ auf einem km² leben durchschnittlich 40 Personen

c) $EQ = \frac{\text{Erwerbspersonen}}{\text{Wohnbevölkerung}} = \frac{\text{Erwerbstätige} + \text{Erwerbslose}}{\text{Wohnbevölkerung}} = \frac{(13+17)+(1{,}3+1{,}36)}{40} = 0{,}8165$

$\Rightarrow$ Von 1000 <u>Männern</u> sind 817 Erwerbspersonen

d) Abh. EW-Tätige = $0,85 \cdot 17 \; Mio. + 0,95 \cdot 9,2 \; Mio. = 23,19 \; Mio.$

6

a) stille Reserve
b) Erwerbstätig
c) Erwerbstätig
d) Erwerbstätig
e) Arbeitslos
f) Erwerbstätig
g) Nicht-Erwerbsperson

Aufgabenblatt 2:
Statistische Analyse eines einzelnen Merkmals

1 In der autofreien Stadt Klauingen steigt die Fahrraddiebstahlquote immer weiter an. Die Arbeitsgruppe „Statistik“ der Klauinger Gesamtschule, die sich seit je her mit dem Fahrverhalten der Einwohner beschäftigt, nahm dieses zum Anlass für eine neue Umfrage.
Sie befragten 20 Bewohner nach der Anzahl der ihnen gestohlener Fahrräder in den letzten 5 Jahren.
Die Gruppe kam zu folgendem Ergebnis:

0	8	5	6	1
5	2	7	4	6
4	3	1	5	4
7	5	3	6	5

a) Auf welcher Skala wird dieses Merkmal gemessen? Handelt es sich um ein stetiges oder diskretes Merkmal?
b) Ermitteln sie die absoluten und relativen Häufigkeiten und die Verteilungsfunktion anhand einer Tabelle.
c) Wie vielen der Befragten wurden weniger als 5 Räder entwendet?
d) Berechnen Sie den Median (Z) und den Modus (D).
e) Bestimmen Sie die Spannweite (R).

2 Des Weiteren interessierte sich die Statistikgruppe für die im letzten Jahr zurückgelegten Kilometer der Stadtbewohner. Sie kamen zu folgenden erstaunlichen Ergebnissen:

50	107	590	690	745
498	345	93	444	203
655	765	277	480	455
561	401	132	540	478

a) Bestimmen Sie die absoluten und relativen Häufigkeiten unter Berücksichtigung folgender Klassen:
$0 \le x < 200;\ 200 \le x < 400;\ 400 \le x < 600;\ 600 \le x < 800.$
b) Ermitteln Sie des Weiteren die Verteilungs- und die Dichtefunktion.
c) Errechnen Sie die modale Klasse, den Median sowie das arithmetische Mittel.
d) Errechnen Sie das obere und untere Quartil, sowie die Quartilsabweichung und zeichnen sie ein “Box and Whiskers“ Plot.
e) Ermitteln Sie außerdem die Varianz und die Standardabweichung.
f) Errechnen und interpretieren sie die standarisierte Schiefe und die standarisierte Wölbung.

3 Wahr oder Falsch?

a) Der Fechnerschen Lageregel zur Folge gilt für eine asymmetrische Verteilung: Arithmetisches Mittel = Median = Modus.
b) Ein Merkmal des arithmetischen Mittels ist es, dass die Summe der quadrierten Abweichungen der Merkmalswerte vom arithmetischen Mittel gleich 0 ist.
c) Das geometrische Mittel wird zur Berechnung von Durchschnittsgeschwindigkeiten herangezogen, da es in diesem Zusammenhang genauer ist, als das arithmetische Mittel.
d) Bei gruppiertem Datenmaterial ist der Modus immer gleich dem Mittelwert.
e) Eine Verteilungsfunktion kann auch einen Wert größer 1 annehmen.
f) Je größer der Exzess, desto flacher ist die Verteilung.
g) Je stärker negativ das dritte Moment (Schiefe) ist, desto linkssteiler ist die Verteilung.

Lösungen zu Aufgabenblatt 2:
Statistische Analyse eines einzelnen Merkmals

1

a) Metrische Absolutskala / diskretes Merkmal

b)

i	x_i	n_i	h_i	$F(x_i)$
1	0	1	0,05	0,05
2	1	2	0,10	0,15
3	2	1	0,05	0,20
4	3	2	0,10	0,30
5	4	3	0,15	0,45
6	5	5	0,25	0,70
7	6	3	0,15	0,85
8	7	2	0,10	0,95
9	8	1	0,05	1,00
		n = 20		

c) gesucht: $h\,(x < 5) = h\,(x \le 4) = F(4) = 0,45 = 45\%$

d) <u>Modus (D):</u>
Bei metrisch skalierten, ungruppierten Merkmalen, ist der Modus der Merkmalswert
x, bei dem die relative Häufigkeit ihr Maximum annimmt.
$D = 5$

<u>Median (Z):</u>
Der Median halbiert das Datenmaterial, so dass 50% darüber und 50% darunter liegen.
hier ist n = 20 (wichtig: 20 ist eine gerade Zahl)

$$Z = \frac{1}{2}\left[x_{(n/2)} + x_{(n+1/2)} \right]$$

$$x_{(n/2)} = x_{(10)} = 5$$

$$x_{(n+1/2)} = x_{(11)} = 5$$

$$Z = \frac{1}{2}[5 + 5] = 5$$

e) $R = x_{max} - x_{min} = 8 - 0 = 8$

2

a) und b)

Klasse	**Kilometer**						
I	$x_i^u \le x_i < x_i^o$	Δx_i	n_i	$h(x_i)$	$F(x_i^o)$	$f(x_i)$	x_i^*
1	$0 \le x < 200$	200	4	0,2	0,2	0,001	100
2	$200 \le x < 400$	200	3	0,15	0,35	0,00075	300
3	$400 \le x < 600$	200	9	0,45	0,8	0,00225	500
4	$600 \le x < 800$	200	4	0,2	1	0,001	700

c) Modale Klasse:
ist gleich der Klasse mit der größten Häufigkeitsdichte:
Klasse $i = 3$ $(400 \le x < 600)$

Median:
halbiert das Datenmaterial:

$$Z = x_i^u + \frac{F(z) - F(x_i^u)}{h(x_i)} \cdot \Delta x_i$$

$$= 400 + \frac{0,5 - 0,35}{0,45} \cdot 200 = 466,67$$

Arithmetisches Mittel

$$\overline{x} = \sum_{i=1}^{k} x_i^* \cdot h(x_i)$$

$$\overline{x} = 100 \cdot 0,2 + 300 \cdot 0,15 + 500 \cdot 0,45 + 700 \cdot 0,2 = 430$$

d) Unteres Quartil:

$$x_{0,75} = x_i^u + \frac{F(z) - F(x_i^u)}{h(x_i)} \cdot \Delta x_i = 400 + \frac{0,75 - 0,35}{0,45} \cdot 200 = 577,78$$

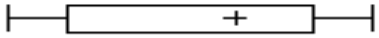
Oberes Quartil:

$$x_{0,25} = x_i^u + \frac{F(z) - F(x_i^u)}{h(x_i)} \cdot \Delta x_i = 200 + \frac{0,25 - 0,2}{0,15} \cdot 200 = 266,67$$

Quartilsabweichung:

$$QA = \frac{1}{2} \cdot (x_{0,75} - x_{0,25}) = \frac{1}{2} \cdot (577,78 - 266,67) = 155,56$$

Box and Whisker Plot:



(-199,9) 50 266,7 466,67 577,7 765 (1044,4)

e) Varianz (vereinfachte Formel)

$$s^2 = \sum_{i=1}^{n} (x_i - \overline{x})^2 \cdot h(x_i)$$

Klasse i	$h(x_i)$	x_i *	$(x_i - \overline{x})$	$(x_i - \overline{x})^2$	$(x_i - \overline{x})^2 \cdot h(x_i)$
1	0,2	100	-330	108900	21780
2	0,15	300	-130	16900	2535
3	0,45	500	70	4900	2205
4	0,2	700	270	72900	14580
					41100

$s^2 = 41.100$

$s = 202,731$

Die gefahrenen Kilometer weichen im Durchschnitt 202,731 km von Mittelwert 430 km ab.

f) Standardisierte Schiefe = m_3/s^3 (Asymmetriemaß)
Standardisierte Wölbung = m_4/s^4 (Wölbungsmaß)

$s^3 = 202,731^3 = 8.332.258,385$

$s^4 = 202,7314 = 1.689.198.357$

$$m^3 = \sum_{i=1}^{n} \left(x_i^* - \overline{x}\right)^3 \cdot h(x_i) = -3.426.000$$

$$m^4 = \sum_{i=1}^{n} \left(x_i^* - \overline{x}\right)^4 \cdot h(x_i) = 3.488.370.000$$

Standardisierte Schiefe: $\frac{-3.426.000}{8.332.215,385} = -0,4112$

Standardisierte Wölbung: $\frac{3.488.370.000}{1.689.198.357} = 2,0651$

Da die standardisierte Schiefe kleiner 0 ist, handelt es sich um eine rechtssteile bzw.
um eine linksschiefe Verteilung. Die standardisierte Wölbung ist gleich 2,0651. Dieser Wert lässt sich normieren $sm_4 = m_4/s^4 - 3 = -0,9349$. Damit lässt sich nun sagen, dass die hier vorliegende Verteilung spitzer ist, als die Normalverteilung.

3

a) Falsch.
Der Fechnerschen Lageregel zur Folge gilt für eine asymmetrische Verteilung: arithmetisches Mittel ≠ Z ≠ D . D < Z < arithmetisches Mittel = linkssteile Verteilung und arithmetisches Mittel < Z< D. (nur bei einer Uni-modalen Verteilung sinnvoll).

b) Falsch.
Die Summe der quadrierten Abweichungen der Merkmalswerte vom arithm. Mittel ist ein Minimum.

c) Falsch.
Das geometrische Mittel wird bei Wachstumsraten herangezogen, bei der Berechnung von Durchschnittsgeschwindigkeiten hilft das harmonische Mittel.

d) Falsch.

e) Falsch.
Die Verteilungsfunktion besteht aus den aufsummierten Wahrscheinlichkeiten und kann somit nicht über 1 steigen.

f) Richtig.

g) Falsch.
Je stärker negativ das dritte Moment ist, desto rechtssteiler ist die Verteilung.

Aufgabenblatt 3:
Konzentration und statistische Analyse mehrerer Merkmale

1 Sie wollen das Nettoeinkommen der Deutschen analysieren. Eine Befragung (Quelle: Allbus 2004) ergab folgende Einkommensverteilung:

Monatliches Nettoeinkommen in EUR	0-500	500-1000	1000-2000	2000-10000
Anzahl der Personen	440	735	1000	325

a) Ermitteln Sie die kumulierte relative Merkmalssumme und tragen Sie diese mit der kumulierten Häufigkeit zusammen in einem Diagramm (Lorenzkurve) ab.
b) Beurteilen Sie, ob die deutsche Einkommensverteilung gleichverteilt ist. Berechnen Sie hierzu ein geeignetes statistisches Maß.

2 Ihr Kommilitone ist Politikstudent und soll den Zusammenhang zwischen dem politischen Interesse und der Wahlabsicht analysieren. Das politische Interesse x wurde als ordinales Merkmal erfasst (1:sehr stark; 2:stark; 3:mittel; 4:wenig; 5:überhaupt nicht). Bei der Wahlabsicht y wurde nach folgenden Parteien gefragt:
1: CDU/CSU
2: SPD
3: FDP
4: Die Grünen
5: Sonstige/Nichtwähler

	$\mathbf{y_1}$	$\mathbf{y_2}$	$\mathbf{y_3}$	$\mathbf{y_4}$	$\mathbf{y_5}$	$n_{i\bullet}$
$\mathbf{x_1}$	59	41	16	27	41	184
$\mathbf{x_2}$	146	77	33	56	93	
$\mathbf{x_3}$	295	143	44	69	171	722
$\mathbf{x_4}$	135	62	14	24	131	366
$\mathbf{x_5}$	38		4	4	76	142
$n_{\bullet j}$	673	343	111		510	1819

a) Ermitteln Sie die fehlenden Werte in der Tabelle sowie die Randverteilungen.
b) Berechnen und interpretieren Sie die Ausdrücke $n_{4\bullet}$, $h_{\bullet 5}$, $h(y_3 \mid x_2)$, $h(y_2)$, $h(x_4, y_3)$

3 Da Sie völlig begeistert über die im „Allbus 2004“ erhobenen Daten sind, wollen Sie den Zusammenhang zwischen einigen Merkmalen analysieren. Welches Zusammenhangsmaß wählen Sie?
a) Wahlabsicht – Politisches Interesse
b) Wahlabsicht – Nettoeinkommen
c) Politisches Interesse – Nettoeinkommen
d) Nettoeinkommen – Alter in Jahren

4 Sie interessieren sich für den Zusammenhang zwischen dem Alter und dem Nettoeinkommen in Deutschland. Aus dem Allbus-Datenmaterial nehmen Sie eine zufällige Stichprobe von sechs Personen. Die Angaben für das Alter und das Nettoeinkommen dieser sechs Personen sind in folgender Tabelle dargestellt:

Person	1	2	3	4	5	6
Alter in Jahren (X)	19	20	33	47	69	76
Nettoeinkommen in EUR (Y)	450	500	250	2600	950	400

a) Berechnen und interpretieren Sie den Korrelationskoeffizienten nach Bravais-Pearson. Warum ist dieser am besten für dieses Datenmaterial geeignet?

b) Berechnen und interpretieren Sie nun den Rangkorrelationskoeffizienten nach Spearman. Warum ist dieser Koeffizient nicht für dieses Datenmaterial geeignet?

5 In einer weiteren Analyse wollen Sie nun analysieren, ob sich das Wahlverhalten zwischen Männern und Frauen unterscheidet. Folgende Daten liegen Ihnen vor:

	CDU/CSU	SPD	FDP	Die Grünen	Sonstige/ Nichtwähler	Summe
männlich	390	210	100	120	280	1100
weiblich	320	170	50	100	260	900
Summe	710	380	150	220	540	2000

e

Besteht ein Zusammenhang zwischen dem Geschlecht und dem Wahlverhalten? Berechnen Sie hierzu einen geeigneten Korrelationskoeffizienten und interpretieren Sie das Ergebnis.

Lösungshinweise:

1b) Gini = 0,46725

4a) Kovarianz = 2.850
Standardabweichung(X) = 22,286
Standardabweichung(Y) = 807,259

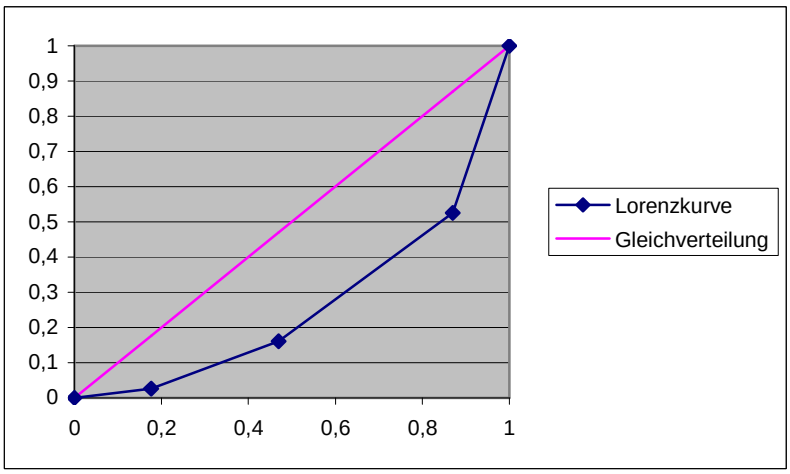
4b) Transformieren Sie die Daten zunächst in Ränge.
r(Spearman) = 0,0857

5) Hier ist nach dem Kontingenzkoeffizienten gefragt. Zunächst müssen die Häufigkeiten bei Unabhängigkeit ermittelt werden. Das M der Formeln aus S.174 im Skript ist das Minimum aus der Anzahl der Zeilen und der Anzahl der Spalten. Da es hier 2 Zeilen und 5 Spalten gibt, ist M = 2. Als Ergebnis erhält man K = 0,1019: Der Zusammenhang ist also schwach.

Lösungen zu Aufgabenblatt 3:
Konzentration und die statistische Analyse mehrerer Merkmale

1 a)

i	n_i	x_i^*	h_i	$F(x_i)$	$x_i^* \cdot n_i$	$x_i^* \cdot n_i \;/\; n \cdot \overline{x}$	$MS(x_i)$
1: 0-500	440	250	0,176	0,176	110000	0,0268	0,0268
2: 500-1000	735	750	0,294	0,470	551250	0,1341	0,1609
3: 1000-2000	1000	1500	0,400	0,870	1500000	0,3649	0,5257
4: 2000-10000	325	6000	0,130	1,000	1950000	0,4743	1,0000
Summe	2500		1,000		4111250	1,0000	



b)

Gini berechnen:

$$G = \sum_{i=1}^{k} \left(F(x_{i-1}) + F(x_i)\right) \cdot \frac{x_i^* \cdot n_i}{n \cdot \overline{x}} - 1$$

$$= (0 + 0{,}176) \cdot 0{,}0268$$
$$+(0{,}176 + 0{,}470) \cdot 0{,}1341$$
$$+(0{,}470 + 0{,}870) \cdot 0{,}3649$$
$$+(0{,}870 + 1{,}00) \cdot 0{,}4743$$
$$-1$$
$$= 0{,}46725$$

2

	y_1	y_2	y_3	y_4	y_5	$n_{i\bullet}$
x_1	59	41	16	27	41	184
x_2	146	77	33	56	93	405
x_3	295	143	44	69	171	722
x_4	135	62	14	24	131	366
x_5	38	20	4	4	76	142
$n_{\bullet j}$	673	343	111	180	512	1819

$n_{4\bullet} = 366$ (366 Personen haben wenig politisches Interesse)
$h_{\bullet 5} = 510/1819 = 0{,}2804$ (28,04 Prozent aller Personen wählen Sonstige/sind Nichtwähler)
$h(y_3 \mid x_2) = 33/405 = 0{,}0815$ (8,15 Prozent mit starkem politischen Interesse wählen die FDP)
$h(y_2) = 343/1819 = 0{,}1886$ (18,86 Prozent aller Personen wählen die SPD)
$h(x_4 , y_3) = 14/1819 = 0{,}0077$ (0,77 Prozent haben wenig politisches Interesse und wählen die FDP)

3

a) Wahlabsicht (nominal) – Politisches Interesse (ordinal)
Kontingenzkoeffizient
b) Wahlabsicht (nominal) – Nettoeinkommen (metrisch)
Kontingenzkoeffizient
c) Politisches Interesse (ordinal) – Nettoeinkommen (metrisch)
Spearman Rangkorr.koeff.
d) Nettoeinkommen (metrisch) – Alter in Jahren (metrisch)
Bravais-Pearson

4

a)

dazu Berechnung der Kovarianz:

$$s_{xy} = \frac{1}{n}\sum_{i=1}^{n} x_i \cdot y_i - \bar{x} \cdot \bar{y}$$

Tabelle mit Hilfswerten:

Monat	x	y	$x \cdot y$	x^2	y^2
1	19	450	8.550	361	202.500
2	20	500	10.000	400	250.000
3	33	250	8.250	1.089	62.500
4	47	2.600	122.200	2.209	6.760.000
5	69	900	62.100	4.761	810.000
6	76	400	30.400	5.776	160.000
Summe:	264	5.100	241.500	14.596	8.245.000

$$\bar{x} = \frac{1}{6} \cdot 264 = 44$$

$$\bar{y} = \frac{1}{6} \cdot 5.100 = 850$$

$$s_{xy} = \frac{1}{n} \sum_{i=1}^{n} x_i \cdot y_i - \bar{x} \cdot \bar{y}$$

$$= \frac{1}{6} \cdot 241.500 - 44 \cdot 850 = 2.850$$

Die Kovarianz misst, ob ein linearer Zusammenhang zwischen Variablen besteht. Das Vorzeichen gibt die Richtung des Zusammenhangs an: es besteht also ein positiver Zusammenhang.
Die Stärke des Zusammenhangs lässt sich anhand der Kovarianz allerdings nicht bemessen, da ihr Wertebereich nicht normiert ist. Hierfür benötigt man den Bravais-Pearson-Korrelationskoeffizienten.

$$r = \frac{s_{xy}}{s_x \cdot s_y}$$

$$s_x = \sqrt{\frac{1}{n} \cdot \sum_{i=1}^{n} x_i^2 - \bar{x}^2} = \sqrt{\frac{1}{6} \cdot 14596 - 44^2} = 22,286$$

$$s_y = \sqrt{\frac{1}{n} \cdot \sum_{i=1}^{n} y_i^2 - \bar{y}^2} = \sqrt{\frac{1}{6} \cdot 8.245.000 - 850^2} = 807,259$$

$$r = \frac{2850}{22,286 \cdot 807,259} = 0,1584$$

Es liegt also ein schwacher positiver Zusammenhang vor.

b) Umwandlung des Datensatzes in Ränge:

Person	1	2	3	4	5	6
Alter in Jahren (X)	19	20	33	47	69	76
Nettoeinkommen in EUR (Y)	450	500	250	2600	900	400
R(X)	1	2	3	4	5	6
R(Y)	3	4	1	6	5	2
R(X)-R(Y)	-2	-2	2	-2	0	4
[R(X)-R(Y)]²	4	4	4	4	0	16

$$\text{r(Spearman)} = 1 - \frac{6 \sum_{i=1}^{n} [R(x) - R(y)]^2}{(n-1)n(n+1)} = 1 - \frac{6 \cdot (4+4+4+4+0+16)}{5 \cdot 6 \cdot 7} = 0,0857$$

Es liegt also ebenfalls ein schwacher positiver Zusammenhang vor. Allerdings sollte man den BP-Koeffizienten berechnen, weil bei der Spearman–Methode durch die Transformation in Ränge Informationen verloren gehen.

5

Zunächst Berechnung der Häufigkeiten bei Unabhängigkeit (in Klammern)

	CDU/CSU	SPD	FDP	Die Grünen	Sonstige/ Nichtwähler	Summe
männlich	390 (390,5)	210 (209)	100 (82,5)	120 (121)	280 (297)	1100
weiblich	320 (319,5)	170 (171)	50 (67,5)	100 (109)	260 (243)	900
Summe	710	380	150	220	540	2000

Berechnung von CHI-QUADRAT:

$$\chi^2 = \sum_{i=1}^{k}\sum_{j=1}^{m}\frac{(n_{ij} - \tilde{n}_{ij})^2}{\tilde{n}_{ij}} = \frac{0,5^2}{390,5} + \frac{0,5^2}{319,5} + \frac{1^2}{209} + \frac{1^2}{171} + \frac{17,5^2}{82,5} + \frac{17,5^2}{67,5} + \frac{1^2}{121} + \frac{1^2}{109} + \frac{17^2}{297} + \frac{17^2}{243} = 10,441$$

Kontingenzkoeffizient:

$$K = \sqrt{\frac{\chi^2}{n + \chi^2} \cdot \frac{M}{M - 1}} = \sqrt{\frac{10,441}{2000 + 10,441} \cdot \frac{2}{2 - 1}} = 0,1019$$

M: Minimum aus [Anzahl der Zeilen & Anzahl der Spalten]

Es besteht also nur ein schwacher Zusammenhang zwischen dem Geschlecht und der Wahlabsicht.

B Klausur mit Lösung

Prof. Dr. Joachim Merz

Statistik I – Deskription
Klausur zum Wintersemester 2005 / 2006
25.1.2006

Aufgabe 1: Allgemeines, Wirtschafts- und Sozialstatistik

a) ☒ Welche der folgenden Aussagen ist/ sind richtig?

A: Das arithmetische Mittel kann als Lageparameter für ordinalskalierte Merkmale herangezogen werden.

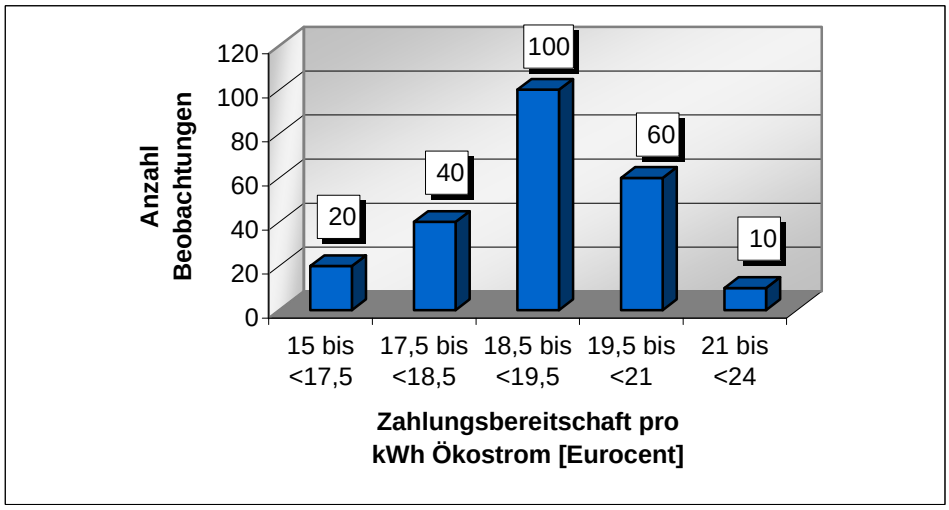
B: Für die Aussage „Die Waschkraft ist 3 mal größer als bei einem herkömmlichen Waschmittel“ ist mindestens eine Verhältnisskala nötig.

C: Falls für vergebene ‚Beliebtheitspunkte’ ein konstanter Wertabstand gilt, handelt es sich hierbei um eine Ordinalskala.

D: Der Gini-Koeffizient setzt mindestens ein ordinales Skalenniveau voraus.

Aufgabe 2: Eindimensionale Häufigkeitsverteilung

Sie untersuchen für die Marktforschungsabteilung der OELGEMOELLER ENERGIE AG die maximale Zahlungsbereitschaft für Ökostrom aus Solar- und Windkraftanlagen. Folgende Verteilung wurde ermittelt:



a) ⌦ Das arithmetische Mittel für diese Verteilung beträgt:
A: 19,065 cent B: 18,39 cent C: 19,2 cent D: 20,140 cent

b) ⌦ Die Standardabweichung dieser Verteilung beträgt:
A: 3,122 cent B: 1,70 cent C: 1,31 cent D: 18,903 cent

c) Erstellen Sie ein Box-and-Whisker-Diagramm für die vorliegenden Daten. Berechnen Sie hierfür zunächst alle notwendigen Parameter und beschriften Sie Ihren Plot mit diesen.

d) Berechnen Sie die Schiefe für die Verteilung der Zahlungsbereitschaft. Im Vorjahr betrug die Schiefe m_3=-0,52. Welche Schlüsse können Sie aus dem vergleich der aktuellen Schiefe mit der des Vorjahres ziehen?

e) ⌦ Welche der folgenden Aussagen ist/ sind richtig?
A: Ein p-Quantil ist die relative Häufigkeit eines bestimmten Merkmals.
B: Das harmonische Mittel wird zur Mittelung von Wachstumsraten herangezogen.
C: Zwischen $x_{0,25}$ und $x_{0,75}$ liegen 50% aller Merkmalsträger.
D: Bei einer symmetrischen Verteilung sind Modus und Median gleich groß.

Aufgabe 3: Konzentration 18 P.

Sie möchten die Informationen über die Zahlungsbereitschaft für Ökostrom der OELGEMOELLER ENERGIE AG (Siehe Aufgabe 2) hinsichtlich der Verteilung und Konzentration genauer analysieren.

a) Berechnen Sie die Verteilungsfunktion der Stromkunden sowie die kumulierte relative Merkmalssumme für die Zahlungsbereitschaft und stellen Sie Ihre Ergebnisse in einer passenden Grafik dar.

b) Berechnen Sie den Gini-Koeffizienten für die Verteilung der Zahlungsbereitschaft. Welche Aussagen können Sie aufgrund Ihres Ergebnisses machen?

c) ⌦ Welchen Wert *könnte* der Gini-Koeffizient *maximal* annehmen, wenn bei der vorliegenden Stichprobe die Konzentration immer weiter zunehmen würde?
A: 0,500 B: 0,824 C: 0,996 D: 1,000

d) ⌦ Welche der folgenden Aussagen ist/ sind richtig?
A: Die Lorenzkurve kann niemals über der Gleichverteilungsgeraden liegen.

B: Zwei Lorenzkurven verschiedener Verteilungen lassen sich nur vergleichen, wenn sich diese nicht schneiden.
C: Der Gini-Koeffizient ist die Fläche unterhalb der Lorenzkurve.
D: Das erste Dezil ist betragsmäßig immer kleiner (oder höchstens gleich groß) als das letzte Dezil.

Aufgabe 4: Zweidimensionale Häufigkeiten und Korrelation 20 P.

In einer weiteren Umfrage unter den Kunden der OELGEMOELLER ENERGIE AG über die Zahlungsbereitschaft für Ökostrom haben Sie Einzeldaten erhoben. Ihnen liegen Informationen über die Zahlungsbereitschaft und über zusätzliche soziodemografische Merkmale vor:

Zahlungs-bereitschaft [Eurocent]	Monatliches Nettoeinkommen [EUR]	Kinder im Haushalt	Alter	Wohngegend
15,5	1.400	Nein	35	Ländlich
21,0	2.500	Ja	48	Kleinstadt
20,0	2.600	Ja	42	Stadt
19,5	2.100	Nein	39	Ländlich
16,5	1.200	Ja	32	Ländlich
17,5	1.600	Ja	24	Stadt
17,0	1.100	Nein	21	Stadt
21,5	2.000	Nein	50	Kleinstadt
16,0	2.000	Nein	32	Stadt

a) Erstellen Sie auf Grundlage des Datenmaterials *eine* Kreuztabelle, die die Verteilung der absoluten Häufigkeiten der Stromkunden zwischen Altersgruppen („Personen bis 40 Jahren“ und „Personen über 40 Jahren“) und dem Wohnort darstellt. Berechnen Sie hierfür die Randverteilungen.

b) Berechnen und interpretieren Sie die Werte h(Stadt, alt), h(Stadt | jung) und h(alt).

c) ⌦ Welche der folgenden Aussagen ist/ sind richtig?
   A: Die Korrelation zwischen Alter und Zahlungsbereitschaft ist mit dem Bravais-Pearson Korrelationskoeffizienten zu berechnen.
   B: Im Gegensatz zur quadratischen Kontingenz (χ^2) kann der normierte Kontingenzkoeffizient K* die Richtung eines Zusammenhangs bestimmen.
   C: Der Bravais-Pearson-Korrelationskoeffizient ist grundsätzlich höher als der Spearman'sche Rangkorrelationskoeffizient.
   D: Der Zusammenhang zwischen der Wohngegend und der Zahlungsbereitschaft kann mit dem Rangkorrelationskoeffizienten nach Spearman berechnet werden.

d) Berechnen und interpretieren Sie auf der Grundlage des gegebenen Datenmaterials ein geeignetes Korrelationsmaß für die Zahlungsbereitschaft und das Einkommen der befragten Kunden.

Klausurlösung zur Klausur WS 05/06

Aufgabe 1: Allgemeines, Wirtschafts- und Sozialstatistik

b) A: falsch
B: richtig
C: falsch
D: falsch

Aufgabe 2: Ordinale Häufigkeitsskala

a) $16,25 \cdot 0,087 + 18 \cdot 0,174 + 19 \cdot 0,435 + 20,25 \cdot 0,261 + 22,5 \cdot 0,043 = 19,065$

b)

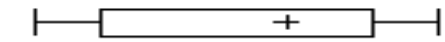
$$s = \sqrt{s^2} = \sqrt{\sum x^{2*}_i \cdot h_i - \overline{x}^2}$$

$$s = \sqrt{264,06 \cdot 0,087 + 324 \cdot 0,174 + 361 \cdot 0,435 + 410,06 \cdot 0,261 + 506,25 \cdot 0,043 - 19,065^2} = 1,31$$

c) $$x_{0,25} = 17,5 \cdot \frac{0,25 - 0,08696}{0,1739} \cdot 1 = 18,4375$$

$$x_{0,50} = 18,5 \cdot \frac{0,5 - 0,26087}{0,434} \cdot 1 = 19,05$$

$$x_{0,75} = 19,5 \cdot \frac{0,75 - 0,69565}{0,2609} \cdot 1,5 = 19,8125$$



16,34 18,437 19,05 19,81 21,87

(Zusatzinfo: $1,5 \cdot \text{Boxbreite} = 1,5 \cdot 1,375 = 2,06$)

d) $$m_3 = \frac{1}{n} \cdot \sum \left(x^*_i - \overline{x} \right)^3 \cdot n_i$$

$$m_3 = \frac{1}{230} \cdot \left[(16,25 - 19,065)^3 \cdot 20 + (18 - 19,065)^3 \cdot 40 + ... \right] = \frac{1}{230} \cdot 10,4 = 0,045$$

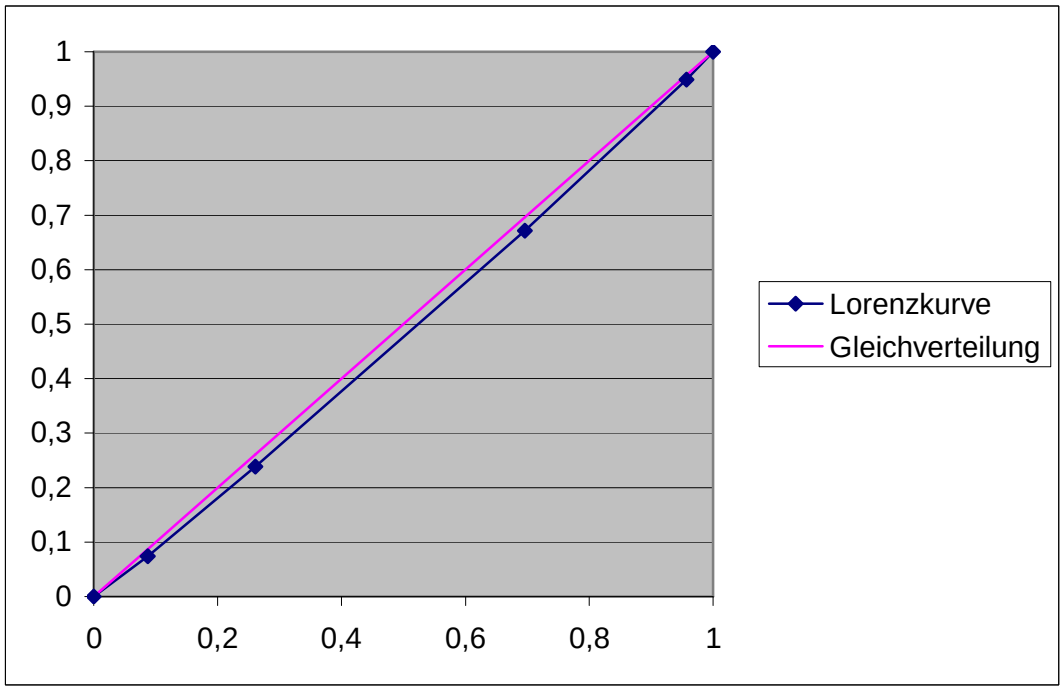
$0,045 > -0,52$ jetzt linkssteil, früher rechssteil

e) A: falsch
B: falsch
C: richtig
D: richtig

Aufgabe 3: Konzentration

a)

	Δx	x^*	n	h	$F(x)$	$x^* \cdot n$	$\frac{x^* \cdot n}{n \cdot \overline{x}}$	$MS(x)$
1	2,5	16,25	20	0,087	0,087	325	0,0741	0,0741
2	1	18	40	0,174	0,261	720	0,1642	0,2383
3	1	19	100	0,435	0,696	1.900	0,4333	0,6716
4	1,5	20,25	60	0,261	0,957	1.215	0,2771	0,9487
5	3	22,5	10	0,043	1	225	0,0513	1
			230	1		4.385		



b) $$G = \left[\sum \left(F(x_{i-1}) + F(x_i) \cdot \frac{x^* \cdot n}{n \cdot \overline{x}} \right) \right] - 1$$

$$G = [(0 + 0,087) \cdot 0,0741 + (0,087 + 0,261) \cdot 0,1642 + ...] - 1 = 0,0367$$

Der Gini ist nahe 0 und daher ist die Verteilung ziemlich eng an der Gleichverteilung

c) C

d) A: richtig
B: richtig
C: falsch
D: richtig

Aufgabe 4: Zweidimensionale Häufigkeiten und Korrelation

a) X = Alter der Personen
Y = Wohnort

	Ländlich	Kleinstadt	Stadt	
≤ 40 Jahre	3	0	3	6
> 40 Jahre	0	2	1	3
	3	2	4	9

b) h(Stadt, alt) = $\frac{1}{9} = 0,11$ aller Befragten sind > 40 und leben in der Stadt.

h(Stadt | jung) = $\frac{3}{6} = 0,5$ der jungen Leute leben in einer Stadt

h(alt) = $\frac{3}{9} = 0,33$ sind über 40 Jahre alt.

c) A: richtig
B: falsch
C: falsch
D: richtig

d) Beide metrisch skaliert – Bravais-Pearson

$$r = \frac{785,19}{1082,41} = 0,725$$

$\overline{x} = 18,278$ $s_x = 2,12$

$\overline{y} = 1833,33$ $s_y = 509,9$

$$r = \frac{785,19}{1082,41} = 0,725$$

Anhang:
Formelsammlung ‚Statistik für alle'

I Allgemeine Grundlagen

Summen, Doppelsummen

Häufig hat man es mit Summen endlich vieler Sumanden zu tun. Um die Schreibweise zu vereinfachen, können sie mit dem griechischen Sigma Σabgekürzt werden.

Definition: Das Summenzeichen steht als als Wiederholungszeichen für die fortgesetzte Addition:

$$\sum_{i=k}^{m} a_i = a_k + a_{k+1} + ... + a_m, \qquad \begin{array}{l} i,k,m \in N \\ k < m \end{array}$$

wobei:
i = Summationsindex
k = untere Summationsgrenze
m= obere Summationsgrenze
a_i = allg. Summationsglied

Beispiele:

a) $\sum_{i=1}^{3} i^2 = 1^2 + 2^2 + 3^2$

b) $\sum_{1}^{4} c = c + c + c + c = 4c$

c) $\sum_{i=1}^{4} x^i = x^1 + x^2 + x^3 + x^4$

Zerlegungsregeln für einfache Summen

1. Summe gleicher Summanden

$$\sum_{i=k}^{m} a = (m-k+1)a \quad , \sum_{i=1}^{n} a = na$$

2. Summen mit gleicher Summationsvorschrift

$$\sum_{i=k}^{m} (a_i + b_i) = \sum_{i=k}^{m} a_i + \sum_{i=k}^{m} b_i$$

3. Summen mit additiven Konstanten

$$\sum_{i=k}^{m}(a_i + c) = \sum_{i=k}^{m} a_i + (m-k+1)c$$

4. Summen mit multiplikativen Konstanten

$$\sum_{i=k}^{m} ca_i = c\sum_{i=k}^{m} a_i$$

5. Summenzerlegung

$$\sum_{i=k}^{m} a_i = \sum_{i=k}^{l} a_i + \sum_{i=l+1}^{m} a_i \qquad , \mathrm{k} \le \mathrm{l} \le \mathrm{m}$$

II Eindimensionale Häufigkeitsverteilung

Qualitative (nominalskalierte) Merkmale

Ausprägung eines qualitativen Merkmals	A_i
absolute Häufigkeit eines Merkmals	$n_i = n(A_i)$
Anzahl der verschiedenen Ausprägungen	k
Anzahl der Beobachtungen	$n = \sum_{i=1}^{k} n_i$
relative Häufigkeit eines Merkmales (Häufigkeitsverteilung)	$h(A_i) = \frac{n_i}{n}$

Quantitative (metrisch skalierte) Merkmale

Diskrete Merkmale

Merkmalswert	x_i
absolute Häufigkeit eines Merkmals	$n_i = n(x_i)$
Anzahl der verschiedenen Merkmalswerte	k
Anzahl der Beobachtungen	$n = \sum_{i=1}^{k} n_i$
relative Häufigkeit eines Merkmalswertes (Häufigkeitsfunktion, - verteilung)	$h(x_i) = \frac{n_i}{n}$

kumulierte absolute Häufigkeit $n(x \le x_i) = \sum_{j=1}^{i} n_j$

kumulierte relative Häufigkeit $h(x \le x_i) = \sum_{j=1}^{i} h(x_j)$

Verteilungsfunktion $F(x_i) = h(x \le x_i) = \sum_{j=1}^{i} h(x_j), \quad i = 1, \ldots, k$

Stetige Merkmale

Merkmalswert x

Klassenuntergrenze der Merkmalsklasse i x_i^u

Klassenobergrenze der Merkmalsklasse i x_i^o

Klassenbreite $\Delta x_i = x_i^o - x_i^u$

absolute Häufigkeit der in der Klasse i liegenden Merkmalswerte $n_i = n(x_i^u \le x < x_i^o)$

Anzahl der Klassen k

Anzahl der Beoachtungen $n = \sum_{i=1}^{k} n_i$

relative Häufigkeit der in der Klasse i liegenden Merkmalswerte (Häufigkeitsverteilung, $i = 1, \ldots, k$) $h(x_i) = h(x_i^u \le x < x_i^o) = \frac{n_i}{n}$

normierte relative Häufigkeit (Dichtefunktion, $i = 1, \ldots, k$) $f(x_i) = \frac{n_i}{n \Delta x_i}$

kumulierte relative Häufigkeit $h(x \le x_i^o) = \sum_{j=1}^{i} h(x_j)$

Interpolation innerhalb der Klasse i $F(x) = F(x_i^u) + \frac{x - x_i^u}{\Delta x_i} \cdot h(x_i)$

III Lageparameter

Häufigster Wert (Modus)

- ungruppiertes Datenmaterial

$$D = x_i \left| \frac{n(x_i)}{n} = \max \right.$$

- gruppiertes Datenmaterial

$$D = x_i^* \left| \frac{n_i}{n\Delta x_i} = \max \right.$$

Median (Zentralwert)

- ungruppiertes Datenmaterial

falls n ungerade $\quad Z = x_{\frac{n+1}{2}}$

falls n gerade $\quad Z = \frac{1}{2}(x_{\frac{n}{2}} + x_{\frac{n}{2}+1})$

- gruppiertes Datenmaterial: Der Median ist nur approximativ mit Hilfe der Verteilungsfunktion erhältlich. Es gilt: $h(x \le Z) = F(Z) = 0,5$

lineare Interpolation bei metrisch skalierten, stetigen Merkmalen:

$$Z = x_i^u + \frac{F(Z) - F(x_i^u)}{h_i} \Delta x_i$$

Arithmetisches Mittel

- ungruppiertes Datenmaterial

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

- gruppiertes Datenmaterial

bekannte Gruppenmittel $\quad \bar{x} = \frac{1}{n}\sum_{i=1}^{k} \bar{x}_i n_i = \sum_{i=1}^{k} \bar{x}_i \cdot h(x_i)$

unbekannte Gruppenmittel $\quad \bar{x} = \frac{1}{n}\sum_{i=1}^{k} x_i^* n_i \quad \text{mit} \quad x_i^* = \frac{1}{2}(x_i^u + x_i^o)$

Geometrisches Mittel

$$GM = \sqrt[n]{\prod_{i=1}^{n} x_i}, \qquad (x_i > 0) \qquad \log GM = \frac{1}{n}\sum_{i=1}^{n} \log x_i$$

Harmonisches Mittel

$$HM = \frac{n}{\sum_{i=1}^{n} \frac{1}{x_i}}$$

IV Streuungsparameter

Spannweite (range) R

$$R = x_{max} - x_{min} = x_{(n)} - x_{(1)}$$

Quartilsabweichung

$$QA = \frac{1}{2}\ (x_{0,75} - x_{0,25})$$

p-Quantile

Interpolationsformel bei gruppiertem Datenmaterial:

$$x_p = x_i^u + \frac{F(x_p) - F(x_i^u)}{\frac{n_i}{n}} \cdot \Delta x_i$$

Mittlere absolute Abweichung

- ungruppiertes Datenmaterial

$$d = \frac{1}{n}\sum_{i=1}^{n}\left|x_i - \overline{x}\right|$$

- gruppiertes Datenmaterial

$$d = \frac{1}{n}\sum_{i=1}^{k}\left|x_i^* - \overline{x}\right| \cdot n_i = \sum_{i=1}^{k}\left|x_i^* - \overline{x}\right| \cdot h_i, \qquad x_i^* = \text{Klassenmitte der Klasse i}$$

Varianz

- ungruppiertes Datenmaterial

$$s^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2 = \frac{1}{n}\sum_{i=1}^{n}x_i^2 - \overline{x}^2$$

- gruppiertes Datenmaterial

$$s^2 = \frac{1}{n}\sum_{i=1}^{k}(x_i^* - \overline{x})^2 \cdot n_i = \frac{1}{n}\sum_{i=1}^{k}(x_i^*)^2 \cdot n_i - \overline{x}^2 \qquad x_i^* = \text{Klassenmitte der Klasse i}$$

Standardabweichung

$$s = \sqrt{s^2}$$

Variationskoeffizient

$$V = \frac{s}{\overline{x}} \cdot 100(\%)$$

Konzept der Momente

Durchschnittliche potenzierte Abweichungen der Merkmalswerte um einen Bezugspunkt a:

Bezugspunkt Null (a = 0)	Momente um Null
Bezugspunkt arithmetisches Mittel (a = $\bar{x}$)	Momente um das arithmetische Mittel

- ungruppiertes Datenmaterial

$$m_r^a = \frac{1}{n}\sum_{i=1}^{n}(x_i - a)^r$$

- gruppiertes Datenmaterial

$$m_r^a = \frac{1}{n}\sum_{i=1}^{k}(x_i^* - a)^r \cdot n_i, \qquad x_i^* = \text{Klassenmitte der Klasse i}$$

Standardisierte Schiefe

Momente 3. Ordnung (r=3) ergeben die Schiefe. Die Schiefe (skewness) ist ein Asymmetriemaß

$$sm_3 = \frac{m_3}{s^3} = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^3}{\left[\sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2}\right]^3}$$

Exzeß (Kurtosis, Wölbung)

Momente 4. Ordnung (r=4) ergeben die Wölbung

$$sm_4 = \frac{m_4}{s^4} = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^4}{\left[\sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2}\right]^4}$$

V Konzentration

Die Merkmale werden für Konzentrationsanalysen grundsätzlich nach ihrer Größe geordnet

*V*erteilungsfunktion:
Abszissenwerte der Lorenzkurve

$$F(x_j) = \sum_{i=1}^{j}\frac{n_i}{n}$$

kumulierte relative Merkmalssumme:
Ordinatenwerte der Lorenzkurve

$$MS(x_j) = \frac{\sum_{i=1}^{j} x_i^* \cdot n_i}{n\bar{x}}$$

Gini Koeffizient:

Gruppiert: $$G = \left[\sum_{i=1}^{k} \left\{ \left[F\left(x_{i-1}\right) + F\left(x_i\right) \right] \cdot \frac{n_i \cdot x_i^*}{n \cdot \bar{x}} \right\} \right] - 1$$

Ungruppiert (x_i geordnet!) $$G = \frac{2 \cdot \sum_{i=1}^{n} i \cdot x_i - (n+1) \cdot \sum_{i=1}^{n} x_i}{n \cdot \sum_{i=1}^{n} x_i}$$

VI Zweidimensionale HK-Verteilung/Korrelationsrechnung

Zweidimensionale Häufigkeitsverteilung – Darstellung

Ausprägung des 1. Merkmals (x) x_i $i = 1, \ldots, k$

Ausprägung des 2. Merkmals (y) y_j $j = 1, \ldots, m$

absolute Häufigkeit des Merkmalspaares (x_i, y_j)

$$n_{ij} = n(x_i, y_j)$$

Anzahl der Beobachtungen

$$n = \sum_{i=1}^{k} \sum_{j=1}^{m} n_{ij}$$

relative Häufigkeit des Merkmalspaares (x_i, y_j)

$$h(x_i, y_j) = \frac{n_{ij}}{n}$$

Randverteilung des 1. Merkmals (x) (Zeilensumme)

marginale absolute Häufigkeiten $$n_{i\cdot} = \sum_{j=1}^{m} n_{ij}$$

marginale relative Häufigkeiten $$h_{i\cdot} = \frac{n_{i\cdot}}{n}$$

Randverteilung des 2. Merkmals (y) (Spaltensummen)

marginale absolute Häufigkeiten $$n_{\cdot j} = \sum_{i=1}^{k} n_{ij}$$

marginale relative Häufigkeiten $$h_{\cdot j} = \frac{n_{\cdot j}}{n}$$

Häufigkeitsverteilung von x bei gegebenem y (bedingte Verteilung)

$$h(x_i \mid y_j) = \frac{n_{ij}}{n_{\cdot j}}$$

Häufigkeitsverteilung von y bei gegebenem x (bedingte Verteilung)

$$h(y_j \mid x_i) = \frac{n_{ij}}{n_{i\cdot}}$$

Korrelationsrechnung

Häufigkeit bei Unabhängigkeit

$$\tilde{n}_{ij} = \frac{n_{i\cdot} \cdot n_{\cdot j}}{n}$$

Quadratische Kontingenz

$$\chi^2 = \sum_{i=1}^{k} \sum_{j=1}^{m} \frac{(n_{ij} - \tilde{n}_{ij})^2}{\tilde{n}_{ij}}$$

Kontingenzkoeffizient

$$K^* = \sqrt{\frac{\chi^2}{n + \chi^2} \cdot \frac{M}{M-1}} \qquad M = \min\ (k, m)$$

Rangkorrelationskoeffizient nach Spearman

R_i Rangnummer des 1. Merkmals (ordinalskaliert)

R_i' Rangnummer des 2. Merkmals (ordinalskaliert)

$$r_{sp} = 1 - \frac{6 \cdot \sum_{i=1}^{n} (R_i - R_i')^2}{(n-1)n(n+1)}, \qquad (-1 \le r_{sp} \le +1)$$

Kovarianz

- ungruppiertes Datenmaterial

$$s_{xy} = \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x}) \cdot (y_i - \bar{y}) = \frac{1}{n} \sum_{i=1}^{n} x_i \cdot y_i - \bar{x} \cdot \bar{y}$$

- gruppiertes Datenmaterial

$$s_{xy} = \frac{1}{n} \sum_{i=1}^{k} \left(x_i^* - \bar{x}\right)(y_i^* - \bar{y}) \cdot n_{i\cdot} = \frac{1}{n} \sum_{i=1}^{k} x_i^* y_i^* \cdot n_{i\cdot} - \overline{xy}$$

Bravais-Pearson-Korrelationskoeffizient

$$r = \frac{s_{xy}}{s_x s_y}, \qquad (-1 \le r \le +1)$$

Literatur

A EINIGE STANDARDWERKE

Anderson, David, R., Sweeney, Dennis, J., Williams, Thomas, A., Freeman, Jim und Essie Shoesmith (2007), Statistics for Business and Economics, Thomson Publisher, London (mit CD)

Anderson, O., Popp, W., Schaffranek, M., Stenger, H. und K. Szameitat (1988), Grundlagen der Statistik, Springer-Verlag, 2. Auflage, Berlin

Bamberg, G. und F. Baur (2002), Statistik, R. Oldenbourg Verlag, 12. Auflage, München

Bleymüller, J., Gehlert, G. und H. Gülicher (2004), Statistik für Wirtschaftswissenschaftler, 14. Auflage, Vahlen, München

Blossfeld, H.-P., Hamerle, A. und K. U. Mayer (1986), Ereignisanalyse: Statistische Theorie und Anwendung in den Wirtschafts- und Sozialwissenschaften, Campus-Verlag, Frankfurt/New York

Bortz, Jürgen (2004, 2010), Statistik für Human- und Sozialwissenschaftler, Springer-Verlag, Berlin

Buttler G. und N. Fickel (2002), Statistik mit Stichproben, Rowohlt Taschenbuch Verlag, Reinbeck bei Hamburg

Fahrmeier, L., Künstler, R., Pigeot, I. und G. Tutz (2004, 2009), Statistik - Der Weg zur Datenanalyse, 5./7. verbesserte Auflage, Springer-Verlag, Berlin

Ferschl, F. (1985), Deskriptive Statistik, Physica-Verlag, 3., korrigierte Auflage, Würzburg

Grohmann, H. (1986a), Statistik - Allgemeine Methodenlehre I (ohne Wahrscheinlichkeitsrechnung), 2. Auflage, dipa-Verlag, Frankfurt a.M.

Hansen, G. (1985), Methodenlehre der Statistik, 3. Auflage, München, Vahlen

Hartung, J., Elpelt, B. und K.-H. Klösener (2005), Statistik: Lehr- und Handbuch der angewandten Statistik, 14., unwesentlich veränderte Auflage, R. Oldenbourg Verlag, München

Hochstädter, D. (1996), Statistische Methodenlehre, 8., überarbeitete Auflage, Verlag Harri Deutsch, Frankfurt a.M.

Hujer, R. (2001), Statistik - Manuskript zur Vorlesung, Frankfurt a.M.

Hujer, R. und R. Cremer (1998), Methoden der empirischen Wirtschaftsforschung, 2. Auflage, Vahlen, München

Kellerer, H. (1976), Statistik im modernen Wirtschafts- und Sozialleben, 14. Auflage, Rowohlt, Reinbek bei Hamburg

Kommission zur Verbesserung der informationellen Infrastruktur zwischen Wissenschaft und Statistik (Hrsg.) (2001), Wege zu einer besseren informationellen Infrastruktur. Gutachten der vom Bundesministerium für Bildung und Forschung eingesetzten Kommission zur Verbesserung der informationellen Infrastruktur zwischen Wissenschaft und Statistik, Baden-Baden.

Kreyszig, E. (1989), Statistische Methoden und ihre Anwendungen, 7. Auflage, Vandenhoeck & Ruprecht, Göttingen

Krug, W. und M. Nourney (2001), Wirtschafts- und Sozialstatistik: Gewinnung von Daten, R. Oldenbourg Verlag, 6. Auflage, München

Kunz, D. (1987), Praktische Wirtschaftsstatistik, Kohlhammer, Stuttgart.

Lippe von der, P. (1996), Wirtschaftsstatistik, 5., völlig neubearbeitete und erweiterte Auflage, Lucius & Lucius, Stuttgart

Litz, H.P. (2003), Statistische Methoden in den Wirtschafts- und Sozialwissenschaften, 3., vollständig überarbeitete und erweiterte Auflage, R. Oldenbourg Verlag, München, Wien

Merz, J. (2011), Statistik I - Deskription, Skriptum zur Vorlesung, 10. erweiterte Auflage, Lüneburg

Mittag, H.J. und D. Stemann, Statistik – Beschreibende Statistik und explorative Datenanalyse, 5., verbesserte und erweiterte Auflage, Fachbuchverlag Leipzig im Carl-Hanser-Verlag, Leipzig

Moore, D.S. (1997), Statistics – Concepts and Controversies, 5. Auflage, W.H. Freeman and Company, New York

Neubauer, W., Bellgardt, E. und A. Behr (2002), Statistische Methoden, 2. Auflage, Vahlen, München

Pfanzagl, J. (1983), Allgemeine Methodenlehre der Statistik I, 6., verbesserte Auflage, Walter de Gruyter, Berlin/New York

Sachs, L. (2004), Angewandte Statistik - Anwendung statistischer Methoden, 11., überarbeitete und aktualisierte Auflage, Springer Verlag, Berlin

Scharnbacher, K. (2004), Statistik im Betrieb - Lehrbuch mit praktischen Beispielen, 14., aktualisierte Auflage, Gabler Verlag, Wiesbaden

Schira, J. (2009), Statistische Methoden der VWL und BWL Theorie und Praxis, 3. aktualisierte Auflage, Pearson Studium, München

Schlittgen, R. (2003), Einführung in die Statistik, 10., durchgesehene Auflage, R. Oldenbourg Verlag, München, Wien

Schwarze, J. (2005, 2009), Grundlagen der Statistik I - Beschreibende Verfahren, 10./11. Auflage, Verlag Neue Wirtschafts-Briefe, Herne/Berlin

Wetzel, W. (1971, 1973), Statistische Grundausbildung für Wirtschaftswissenschaftler, Teil I und Teil II, Walter de Gruyter, Berlin

Yamane, T. (1981), Statistik - Ein einführendes Lehrbuch, Teil I und Teil II, Fischer Taschenbuch Verlag, Frankfurt

Zöfel, P. (1992), Statistik in der Praxis, 3., überarbeitete und ergänzte Auflage, UTB 1293, Stuttgart

B BÜCHER MIT ÜBUNGSAUFGABEN

Bamberg, G. und F. Baur (2000, 2007), Statistik Arbeitsbuch: Übungsaufgaben, Fallstudien, Lösungen, 6./8. Auflage, R. Oldenbourg Verlag, München

Bihn, W.R. und K.A. Schäffer (1986), Übungsaufgaben zur Grundausbildung in Statistik für Wirtschaftswissenschaftler, J.C. Witsch Nachf., Köln

Fahrmeier, L., Künstler, R., Pigeot, I., Tutz, G., Caputo, A. und S. Lang (2003), Arbeitsbuch Statistik, 3., überarbeitete und erweiterte Auflage, Springer-Verlag, Berlin

Hartung, J. und B. Heine (1999), Statistik Übungen: Deskriptive Statistik, 6. Auflage, R. Oldenbourg Verlag, München

Hochstädter, D. (1993), Aufgaben mit Lösungen zur statistischen Methodenlehre, 2. Auflage, Verlag Harri Deutsch, Frankfurt

Lippe von der, P. (2006), Formeln, Aufgaben, Klausurtraining in Statistik, R. Oldenbourg Verlag, 7. Auflage, München/Wien

Merz, J. (2014), Statistik I - Deskription, Übungs- und Klausuraufgaben mit Lösungen, 12. Auflage, Lüneburg

Spiegel, M.R. (2003), Statistik, 1. Auflage, Mc Graw Hill, Düsseldorf

C FORMEL- UND TABELLENWERKE

Bihn, E.R. und K.A. Schäffer (1987), Formeln und Tabellen zur Grundausbildung in Statistik für Wirtschaftswissenschaftler, J.C. Witsch Nachf., Köln

Bleymüller, J. und G. Gehlert (1999), Statistische Formeln, Tabellen und Programme, 9. Auflage, Verlag Franz Vahlen, München

Bohley, P. (1998), Formeln, Rechenregeln und Tabellen zur Statistik, R. Oldenbourg Verlag, 7. Auflage, München

Rinne, H. (1997), Statistische Formelsammlung, 2. Auflage, Verlag Harri Deutsch, Frankfurt

Vogel, F. (2000), Beschreibende und schließende Statistik - Formeln, Definitionen, Erläuterungen, Stichwörter und Tabellen, R. Oldenbourg Verlag, 12. Auflage, München

D WEITERFÜHRENDE LITERATUR

Allgemeines Statistisches Archiv (1992), Band 76

Arminger, G. und F. Müller (1990), Lineare Modelle zur Analyse von Paneldaten, Opladen

Backhaus, K., Erichson, B., Pinke, W., Schuchard-Fischer, Chr. und R. Weiber (2006), Multivariate Analysemethoden, 11., überarbeitete Auflage, Springer Verlag, Berlin/Tokyo

Bates, B.C., Z.W. Kundzewicz, S. Wu, J.P. Palutikof, Eds. (2008), Climate Change and Water. Technical Paper of Ingovernmental Panel on Climate Change, IPCC Secretariat, 210 pp., Geneva

Berntsen, R. (1991), Dynamik in der Einkommensverteilung. Eine empirische Längsschnittuntersuchung der Strukturen der Einkommensverteilung privater Haushalte in der Bundesrepublik Deutschland, Dissertation, Frankfurt

Diewald, M. (1984), Das 'SPES-Indikatoren-Tableau' 1976 - Fortschreibung bis zum Jahr 1982, Sfb 3-Arbeitspapier Nr. 150, Frankfurt/Mannheim

Esenwein-Rothe (1978), Modelle für eine Bevölkerungsprojektion und die Grenzen der Aussagekraft, in: Jahrbuch für Nationalökonomie und Statistik, Band 193, Heft 1

Esser, H., Grohmann, H., Müller, W. und K.A. Schäffer (1989), Mikrozensus im Wandel - Untersuchungen und Empfehlungen zur inhaltlichen und methodischen Gestaltung, Forum der Bundesstatistik, Bd. 11, Metzler-Poeschel Verlag, Stuttgart

Galler, H.P. and G. Wagner (1986), The Microsimulation Model of the Sfb 3 for the Analysis of Economic and Social Policies, in: Orcutt, G.H., Merz, J. and H. Quinke (eds.), Microanalytic Simulation Models to Support Social and Financial Policy, S. 227-247, North Holland, Amsterdam

Galler, H.P. und N. Ott (1994), Das dynamische Mikrosimulationsmodell des Sonderforschungsbereichs 3, in: Mikroanalytische Grundlagen der Gesellschaftspolitik, Deutsche Forschungsgemeinschaft, Bd. 2, Akademie Verlag, Berlin

Glatzer, W. und W. Zapf (Hrsg.) (1984), Lebensqualität in der Bundesrepublik, Frankfurt/New York

Goldberger, A.S. (1991), A Course in Econometrics, Harvard University Press, London

Greene, W. (1991), Econometric Analysis, Macmillan Publishing Company, New York

Greene, W. (1992), ET - The Econometrics Toolkit, Version 3.0, Econometric Software, Inc., New York

Grohmann, H. (1986b), Bevölkerungs- und Wirtschaftsstatistik, dipa-Verlag, 2. Auflage, Frankfurt a.M.

Habich, R., und H.-H. Noll unter Mitarbeit von W. Zapf (1993), Soziale Indikatoren und Sozialberichterstattung - Internationale Erfahrungen und gegenwärtiger Forschungsstand, Berlin/Mannheim

Hamer, G. und C. Stahmer (1992), Integrierte Volkswirtschaftliche- und Umweltgesamtrechnung (I): Konzeption, in: ZfU, Heft 1, S. 85-117; Integrierte Volkswirtschaftliche- und Umweltgesamtrechnung (II): (Zahlen-)Beispiel und Realisierungsmöglichkeiten, in: ZfU, Heft 2, S. 237-256

Hansen, G. (1993), Quantitative Wirtschaftsforschung, Franz Vahlen, München

Hsiao, C. (1986), Analysis of Panel Data, Cambridge (Mass.)

Huff, D. (1978), How to Lie with Statistics, Penguin Books, Harmondsworth (UK)

Johnson, J.D. (1992), Applied Multivariate Data Analysis – Volume II: Categorical and Multivariate Methods, Springer Verlag, New York

Krämer, W. (1991), So lügt man mit Statistik, Campus-Verlag, Frankfurt/New York

Krupp, H.-J. und W. Zapf (1977), Sozialpolitik und Sozialberichterstattung, Frankfurt/New York

Leipert, C. (1975), Unzulänglichkeiten des Sozialprodukts in seiner Eigenschaft als Wohlstandsmaß, Tübingen

Leipert, C. (1989), Die heimlichen Kosten des Fortschritts, Frankfurt

Maddala, G.S., Rao, C.R. und H.D. Vinod (Hrsg.) (1993), Econometrics, 11. Auflage, North-Holland, New York

Malinvaud, E. (1980), Statistical Methods in Econometrics, 3. Auflage, American Elsevier, New York

Mátyás, L. und P. Sevestre (1992), The Econometrics of Panel Data, Handbook of Theory and Applications, Kluwer academic publishers, Dortrecht

Merz (2002), Freie Berufe im Wandel der Märkte, FFB-Schriftenreihe, Band 13, Nomos Verlag, Baden-Baden

Merz, J. (1980a), Die Ausgaben privater Haushalte - Ein mikroökonometrisches Modell für die Bundesrepublik Deutschland, Frankfurt/New York

Merz, J. (1980b), Prognosegüte und Spektraleigenschaften ökonomischer Modelle, in: Stöppler, S. (Hrsg.), Dynamische ökonomische Systeme - Analyse und Steuerung, 2. Auflage, Gabler-Verlag, Wiesbaden, S. 31-66

Merz, J. (1987), Mathematik II für Wirtschaftswissenschaftler, Skriptum zur Vorlesung, Johann Wolfgang Goethe-Universität Frankfurt, Frankfurt a.M.

Merz, J. (1991a), Markt- und nichtmarktmäßige Aktivitäten privater Haushalte - Theoretischer Ansatz, repräsentative Mikrodaten, mikroökonometrische Analyse und Mikrosimulation wirtschafts- und sozialpolitischer Maßnahmen für die Bundesrepublik Deutschland, Frankfurt a.M.

Merz, J. (1991b), Microsimulation - A Survey of Principles, Developments and Applications, in: International Journal of Forecasting, 7, S. 77-104

Merz, J. (1994a), Statisches Mikrosimulationsmodell: Mainframe und PC-Version, in: Hauser, R., Ott, N. und G. Wagner (eds.), Deutsche Forschungsgemeinschaft: Mikroanalytische Grundlagen der Gesellschaftspolitik - Erhebungsverfahren, Analysemethoden und Mikrosimulation, Akademie Verlag, Berlin

Merz, J. (1994b), Microdata Adjustment by the Minimum Information Loss Principle, Forschungsinstitut Freie Berufe der Universität Lüneburg, FFB-Diskussionspapier Nr. 10, Lüneburg

Merz, J. (2008), Statistik II - Wahrscheinlichkeitsrechnung und induktive Statistik, Skriptum zur Vorlesung, 7. verbesserte Auflage, Lüneburg

Merz, J. und H. Stolze (2006), Adjust for Windows Version 1.1 - A Software Package to Achieve Representative Microdata by the Minimum Information Loss Principle - Manual, FFB-Dokumentation Nr. 13, Forschungsinstitut Freie Berufe, Fakultät II Wirtschafts-, Verhaltens- und Rechtswissenschaften, Universität Lüneburg, Lüneburg

Merz, J. und J. Faik (1992), Equivalence Scales Based on Revealed Preference Consumption Expenditure Microdata - The Case of West Germany, Forschungsinstitut Freie Berufe der Universität Lüneburg, FFB-Diskussionspapier Nr. 3, Lüneburg

Merz, J., Helberger, C. und H. Schneider (1985), Nebenerwerbstätigkeitsumfrage 1984, Dokumentation, bearbeitet von Klaus Wolff, Frankfurt

Merz, J., Stolze, H. und M. Zwick (2002), Professions, entrepreneurs, employees and the new German tax (cut) reform 2000 - A MICSIM microsimulation analysis of distributional impacts, Forschungsinstitut Freie Berufe der Universität Lüneburg, FFB-Diskussionspapier Nr. 34, Lüneburg

Mueller, U. (1993), Bevölkerungsstatistik und Bevölkerungsdynamik, Walter de Gruyter, Berlin/New York

Noll, H.-H. (1990), Sozialindikatorenforschung in der Bundesrepublik - Konzepte, Forschungsansätze und Perspektiven, in: Timmermann, H. (Hrsg.), Lebenslagen, Sozialindikatorenforschung in beiden Teilen Deutschlands, S. 69-87, Saarbrücken

Noll, H.-H. (1993), Lebensbedingungen in der Europäischen Gemeinschaft gleichen sich nur langsam an - ökonomische und soziale Indikatoren im EG-Vergleich, in: ZUMA, Informationsdienst Soziale Indikatoren (ISI), Heft 4, S. 11-15

Noll, H.-H. (Hrsg.) (1997), Sozialberichterstattung in Deutschland - Konzepte, Methoden und Ergebnisse für Lebensbereiche und Bevölkerungsgruppen, Juventa, Weinheim und München

Nullau, B. u.a. (1969), Das Berliner Verfahren, DIW- Beiträge zur Strukturforschung, Berlin

Orcutt, G., Merz, J. and H. Quinke (eds.) (1986), Microanalytic Simulation Models to Support Social and Financial Policy, North Holland, Amsterdam

Rapin, H. (Hrsg.) (1990), Der private Haushalt - Daten und Fakten, Stiftung 'Der Private Haushalt', Campus, Frankfurt a.M.

Rinne, H. (1994), Wirtschafts- und Bevölkerungsstatistik - Erläuterungen, Erhebungen, Ergebnisse, R. Oldenbourg Verlag, München/Wien

Rockwell, R. (1986/87), Prospects for Social Reporting in the United States: A Receding Horizion, in: The Tocqueville Review, Vol. 8, S. 251-262

Schlittgen, R. und B.H.J. Streitberg (1987), Zeitreihenanalyse, 2. Auflage, R. Oldenbourg Verlag, München/Wien

Schwarze, J. (2006), Grundlagen der Statistik II – Wahrscheinlichkeitsrechnung und induktive Statistik, Verlag Neue Wirtschafts-Briefe, 8. Auflage, Herne/Berlin

Sheldon, E. B. and R. Park (1975), Social Indicators, in: Science, American Association for the Advancement of Science, Vol. 188, S. 693-699

Sheldon, E. B. and W.E. Moore (eds.) (1968), Indicators of Social Change, Concepts and Measurement, New York

Spanos, A. (1986), Statistical Foundation of Econometric Modelling, Cambridge

Stahmer, C. (1992), Integrierte Volkswirtschaftliche- und Umweltgesamtrechnung, in: Wirtschaft und Statistik, Heft 9, S. 577-593

Statistisches Bundesamt (1980), Fachserie 14, Reihe 7.3, Finanzen und Steuern, Lohnsteuern, Wiesbaden

Statistisches Bundesamt (1990), Fachserie 1, Reihe 4.1.1, Bevölkerung und Erwerbstätigkeit, Wiesbaden

Statistisches Bundesamt (Hrsg.) (1988), Das Arbeitsgebiet der Bundesstatistik, W. Kohlhammer, Stuttgart/Mainz

Statistisches Bundesamt (Hrsg.) (1992). Datenreport 1992, in Zusammenarbeit mit dem Wissenschaftszentrum für Sozialforschung, Berlin und dem Zentrum für Umfragen, Methoden und Analysen, Mannheim, Wiesbaden

Statistisches Bundesamt, Statistisches Jahrbuch, verschiedene Jahrgänge, Wiesbaden

Statistisches Bundesamt, Zahlen, Fakten, Trends: Monatlicher Pressedienst, Wiesbaden

Steger, A. (1980), Haushalte und Familien bis zum Jahre 2000 ,Campus, Frankfurt a.M./New York).

Stobernack, M. (1989), Die Bedeutung der Arbeitslosenversicherung für Arbeitslosigkeit und Arbeitsangebot unter Einbeziehung eines empirischen Arbeitsangebotsvergleichs zwischen der Bundesrepublik und den USA, Berlin.

United Nations Development Programme (UNDP) (1991), Human Development Report 1991, Oxford

Vogel, J. (1990), Social Indicators - A Swedish Perspective, in: Journal of Public Policy, Vol. 9, S. 439-444

Yang, M.C.K. and D. Robinson (1986), Understanding and Learning Statistics by Computer, World Scientific, Singapore

Zapf, W. (1972), Zur Messung der Lebensqualität, in: Zeitschrift für Soziologie, 1. Jg., S. 353-367

Zapf, W. (1977), Einleitung in das SPES-Indikatorensystem, in: Zapf, W. (Hrsg.), Lebensbedingungen in der Bundesrepublik. Sozialer Wandel und Wohlfahrtsentwicklung, S. 11-27, Frankfurt/New York

Zapf, W. (1990), Einleitung, in: WZB-AG Sozialberichterstattung (Hrsg.), Sozialreport 1990, Dokumentation eines Workshops am Wissenschaftszentrum Berlin für Sozialforschung, Arbeitspapier P90-102, Berlin

E SONSTIGE LITERATUR

Berger-Schmitt, R. (2002), Unterschiede in den Lebensbedingungen in der Europäischen Union kaum verringert, in: Informationsdienst Soziale Indikatoren (ISI), Ausgabe 27, Januar 2002, S. 2

Club of Rome (1991), Der Blick in die Zukunft, in: natur, Heft 9, S. 31-32

Der siebte Tag (30.11.2002), Wochenendbeilage zur Hannoverschen Allgemeinen Zeitung, November 2002, Madsack-Verlag, Hannover

Die ZEIT (02.09.1994), Nr. 36, ZEITVERLAG, Hamburg

Gonick, L. and W. Smith (1993), The Cartoon Guide to Statistics, HarperCollins Publishers, New York

Kuhn, T. (1970), Structure of Scientific Revolutions, 2. Auflage, Chicago

Meadows, D., Randers, D. und J. Randers (1992), Die neuen Grenzen des Wachstums - Die Lage der Menschheit: Bedrohung und Zukunftschancen, Stuttgart

Merz, J., Rauberger, T.K. und A. Rönnau (1994), Freie Berufe in Rheinland-Pfalz und in der Bundesrepublik Deutschland: Struktur, Entwicklung und wirtschaftliche Bedeutung, Schriften des Forschungsinstituts Freie Berufe der Universität Lüneburg Nr. 7, Lüneburg

Myers, N. (Hrsg.) (1985), gaia - der öko-Atlas unserer Erde, Fischer Verlag, Frankfurt

Schwarze, J. (1977), Bibliographie zur Statistik in der Weiterbildung, 2. Auflage, Pädagogische Arbeitsstelle des Deutschen Volkshochschulverbandes, Holzhausenstr. 21, 6000 Frankfurt, Frankfurt/Bonn

Sozio-ökonomisches Panel (Welle I (1) 1984, Welle I (9), 1992), des Sonderforschungsbereichs 3, Frankfurt/Mannheim und des DIW, Berlin

Statistisches Bundesamt (1986), Volkszählung '87: Zehn Minuten, die allen helfen - Materialien, Abschnitt 2.2, Wiesbaden

Statistisches Bundesamt (2003), Energieverbrauch und Luftemissionen des Sektors Verkehr, Band 12 der Schriftenreihe zu den Umweltökonomischen Gesamtrechnungen - Kurzfassung

Statistisches Bundesamt (2005), Verbraucherpreisindex und Index der Einzelhandelspreise - Jahresdurchschnitte ab 1948, Wiesbaden

Statistisches Bundesamt (Hrsg.) (1988), Das Arbeitsgebiet der Bundesstatistik 1988, Mainz, Kohlhammer

United Nations (2005), Population Challenges and Development Goals, Department of Economic and Social Affairs, Population Division, New York

United Nations Development Program (UNDP) (Hrsg.), Human Development Report 1991, New York, Oxford, 1991,Oxford University Press

United Nations Development Program (UNDP) (Hrsg.), Human Development Report 2003, New York, Oxford, 2003,Oxford University Press

Wolffs, M. (2002), Bevölkerung zwischen Dynamik und Stillstand - Demographische Entwicklungen im Längsschnitt, Sankt Augustin, Arbeitspapier der Konrad-Adenauer-Stiftung e.V.

ZEIT-Punkte (1994), Weltbevölkerung - Wird der Mensch zur Plage?, Nr. 4/1994, ZEIT-VERLAG, Hamburg