bookboon.com

Einfach Lernen! Statistik

Hans Diefenbacher; Andreas Frank

Hans Diefenbacher, Andreas Frank

Einfach Lernen! Statistik

Einfach Lernen! Statistik
1. Auflage

ISBN 87-7681-188-3

Inhalt

	Vorwort	**7**
1	**Einleitung**	**8**
1.1	Zum Aufbau des Buches	8
1.2	Einige Hinweise zur Arbeit mit dem Buch	9
2	**Sammeln und Präsentieren von statistischen Daten**	**10**
2.1	Warum Statistik?	10
2.2	Daten sammeln	13
2.3	Ordnen von Daten	17
2.4	Darstellung	22

3	**Deskriptive Statistik**	**30**
3.1	Mittelwerte	30
3.2	Konzentrationsmaße	33
3.3	Streuung, Schiefe, Wölbung	37
3.4	Regression und Korrelation	41
3.5	Indikatoren und Indices	47
3.6	Zeitreihen	53
4	**Wahrscheinlichkeitsrechnungen**	**61**
4.1	Grundbegriffe	61
4.2	Rechnen mit Wahrscheinlichkeiten	63
4.3	Kombinatorik	65
4.4	Wahrscheinlichkeitsverteilungen	69

5	**Induktive Statistik**	**77**
5.1	Grundlagen der Stichprobentheorie	77
5.2	Punktschätzungen	78
5.3	Verteilungen	81
5.4	Intervallschätzungen	82
5.5	Grundlagen des Testens von Hypothesen	85
5.6	Multivariate Regression und Korrelation	89
6	**Anhang**	**92**
6.1	Abkürzungsverzeichnis	92
6.2	Fachwörter englisch – deutsch	93
6.3	Statistische Tafeln	96
6.5	Weiterführende Literatur	103
6.6	Internet-Quellen	108
7	**Noten**	**109**

Vorwort

Das Buch ist in Gemeinschaftsarbeit von Hans Diefenbacher und Andreas Frank entstanden. Die Entwürfe der Kapitel 1, 2, 4 und sowie 6.4 bis 6.6 stammen von Hans Diefenbacher, Kapitel 3 und 5 sowie 6.1 bis 6.3 von Andreas Frank.

Die Anregung zu diesem Buch stammt von Kristian Buus vom Ventus-Verlag, dem wir herzlich für seine freundliche Betreuung und seine Offenheit für Veränderungen des Konzepts danken. Bedanken möchten wir uns auch bei Franziska Strohmaier, die Teile des Manuskripts Korrektur gelesen und das Layout überarbeitet hat.

Heidelberg, August 2006

Hans Diefenbacher
Andreas Frank

1 Einleitung

1.1 Zum Aufbau des Buches

Warum ein weiteres Buch über Statistik? Gibt es nicht schon genügend davon? Gerade in der letzten Zeit sind eine Reihe ausgezeichneter Einführungen und Kompendien erschienen, und zwar auf den unterschiedlichsten Anspruchsniveaus: von Büchern, die den Leserinnen und Lesern die Angst vor der unvermeidlichen Mathematik nehmen wollen, bis hin zu Büchern, die die Sprache der Mathematik als selbstverständlich voraussetzen. Unser Anhang 7.4. stellt einige Hinweise auf weiterführende Literatur zusammen. Außerdem veralten die Grundlagen der Statistik nicht, sodass sogar einige ziemlich „betagte" Lehrbücher auch heute noch mit großem Gewinn verwendet werden können, weil ihr didaktisches Konzept vorzüglich ist.

Wenn wir uns trotzdem entschieden haben, das vorliegende Buch zu schreiben, dann geschah dies vor allem aus zwei Gründen:

- zum einen wollten wir ein kurzes Kompendium zusammenstellen, das in einer Art Grundkurs wesentliche Bausteine des statistischen Arbeitens vorstellt. Die einzelnen Themenbereiche sollten dabei nicht erschöpfend behandelt werden, den Leserinnen und Lesern aber den Weg zu eigener Weiterarbeit eröffnen;
- zum anderen möchten wir mit diesem Grundkurs den „methodischen Boden" vor allem für den Bereich der angewandten Wirtschafts-, Umwelt- und Sozialstatistik bereiten, ein Bereich, dem in vielen wirtschafts- und sozialwissenschaftlichen Ausbildungsgängen an deutschen Universitäten ein – im Verhältnis zum Studium der theoretischen Statistik – unseres Erachtens viel zu geringes Gewicht beigemessen wird.

Das Buch beginnt daher mit einem Kapitel, in dem wir allgemeine methodische Grundlagen des statistischen Arbeitens zusammenstellen (Kap. 2). Danach folgen Kapitel über deskriptive Statistik (Kap. 3), über Wahrscheinlichkeitsrechnungen (Kap. 4) und über induktive Statistik (Kap. 5). Die zum grundlegenden Handwerkszeug der Statistikerin[1] gehörenden Bausteine erklären wir jeweils in vier kurzen Unterabschnitten mit:

- einer allgemeinen Erläuterung des Konzepts,
- der Darstellung der jeweils anzuwendenden Formel,
- einer Beispielrechnung und
- einer Diskussion von Möglichkeiten und Grenzen der Anwendung des jeweiligen Konzepts.

In einem zweiten Band werden wir eine Zusammenstellung von Übungsaufgaben und weiteren Anwendungsbeispielen präsentieren, mit der der Leser die hier gebotenen Arbeitsschritte eingehender vertiefen kann.

1.2 Einige Hinweise zur Arbeit mit dem Buch

Natürlich muss jeder, der sich mit Statistik beschäftigt, seinen eigenen Zugangsweg finden; der eine bevorzugt das Lernen anhand von anwendungsorientierten Beispielen, die andere begreift Statistik vielleicht eher als ein Spezialgebiet der Mathematik. Wer am Ende jedoch mit den statistischen Methoden tatsächlich selbst arbeiten möchte, dem sei in der Tat geraten, sich mit den Rechenschritten auch möglichst praktisch vertraut zu machen – mit anderen Worten: selbst so lange Beispiele zu rechnen, bis einem der Sinn der jeweiligen Rechenschritte geläufig ist und die Formel nicht mehr als „black box" erscheint. Erst dann können Rechenergebnisse in der Regel auch sicher interpretiert werden.

Der ältere der beiden Verfasser dieses Buches, der Statistik noch in der Taschenrechner-Zeit gelernt hat, die dem PC-Zeitalter vorausging, bekam von seinem Statistik-Lehrer an der Universität Heidelberg, Rolf Wagenführ, empfohlen, ab und zu auch einmal Ergebnisse per Hand (!) mit Hilfe eigens aufzustellender Arbeitstabellen auszurechnen. Dafür hat Wagenführ in seinen Lehrbüchern zahlreiche Beispiele gegeben.[2] So vorsintflutlich dieser Ratschlag klingt, als didaktisches Hilfsmittel zum besseren Verständnis der statistischen Arbeitsweise hat er sich sehr bewährt – auch und gerade heute, wo Computer-Programme wie Excel oder SPSS so leicht zu bedienen sind.

Ein zweiter Ratschlag kann ebenfalls allen gegeben werden, die beginnen, sich mit Statistik aktiv auseinanderzusetzen: Nehmen Sie das Bonmot – „Ich glaube nur jenen Statistiken, die ich selbst gefälscht habe"[3] – nicht ganz ernst, aber beachten Sie es in einem übertragenen Sinne. Wissenschaftliche Arbeiten, Zeitungen, Radiosendungen, Politikerreden – die Welt ist voll von statistischen Aussagen und Beweisführungen.

Sehen und hören Sie sich diese Aussagen, sozusagen als Begleitstudium, immer wieder mit wachen Augen und Ohren an: Wie ist die entsprechende Aussage zustande gekommen? Auf welcher Datengrundlage wird sie beruhen? Wie sind die Daten erhoben worden? Wie sind die Daten geordnet und strukturiert worden? Welche Rechenmethoden, welche Indikatoren sind verwendet worden? Und nicht zuletzt: Wie wird das Ergebnis präsentiert, etwa mit welcher Form der graphischen Aufbereitung? Was wird *nicht* gesagt?

In der Tat ist die Welt auch voll von schlechten Statistiken, die zu fehlerhaften Interpretationen verleiten – manchmal nur aufgrund handwerklicher Fehler, manchmal aber auch als Ergebnis des Wunsches, bestimmte Dinge so zu präsentieren, wie der jeweilige Autor sie gerne sehen *möchte* und nicht so, wie es eine einwandfreie statistische Analyse erfordern würde. Wer ein halbwegs sicheres Gespür für diese Probleme entwickelt hat, der hat den statistischen Grundkurs erfolgreich absolviert.

2 Sammeln und Präsentieren von statistischen Daten

2.1 Warum Statistik?

Angewandte Statistik gibt es seit rund viereinhalb Jahrtausenden, theoretische Statistik seit rund drei Jahrhunderten.[4] In der Alltagssprache wird heute jede halbwegs geordnete Übersicht, die Zahlen zur Darstellung ihrer Informationen enthält oder diese graphisch umsetzt, als Statistik bezeichnet. Das heißt, dass das Wort zum einen als Sammelbegriff für die *Ergebnisse statistischen Arbeitens* verwendet wird. Zum anderen bezeichnet dieses Wort aber einen weiteren Inhalt: Es ist auch Sammelbegriff für die *Verfahren,* mit deren Hilfe wir zu diesen Ergebnissen gelangen.

Wird Statistik in diesem Sinne als Wissenschaft oder – bescheidener ausgedrückt – als Methodenlehre verstanden, dann kann man ihre Aufgabe als *Quantifizierung von Massenerscheinungen* bezeichnen, die in einem weiteren Schritt selbst interpretiert oder anderen als Grundlage eigener Interpretationen zur Verfügung gestellt wird.[5] Durch eine solch allgemeine Definition wird zugleich auch deutlich, dass statistische Methoden in allen empirisch orientierten Wissenschaften von Bedeutung sind. Statistik umfasst damit als Methodenlehre ein klar abgegrenztes Gebiet, ihr Einsatzbereich ist jedoch transdisziplinär in einem umfassenden Sinn. Die Statistik ist, wie Menges schreibt, „zur charakteristischen Methode des modernen Wissenschaftsbetriebes geworden".[6]

Die wissenschaftlichen Fragen, deren Beantwortung eine Quantifizierung von Massenerscheinungen notwendig machen, beziehen sich auf sehr verschiedene Untersuchungsgegenstände. Fragen und Untersuchungsgegenstände machen dabei ganz unterschiedliche methodisch Vorgehensweisen erforderlich. Für viele Fragen genügen die Verfahren der deskriptiven, also der beschreibenden Statistik. Das kann zum Beispiel immer dann der Fall sein, wenn die Grundgesamtheit einer bestimmten, zu untersuchenden statistischen Masse vollständig bekannt ist – etwa das Alter aller Studierenden an einer bestimmten Hochschule. Hier lassen sich dann mit Hilfe der einschlägigen Methoden die entsprechenden Grundgesamtheiten darstellen und interpretieren; so kann man in diesem Fall zum Beispiel Mittelwerte, Extremwerte oder Häufigkeitsverteilungen feststellen.

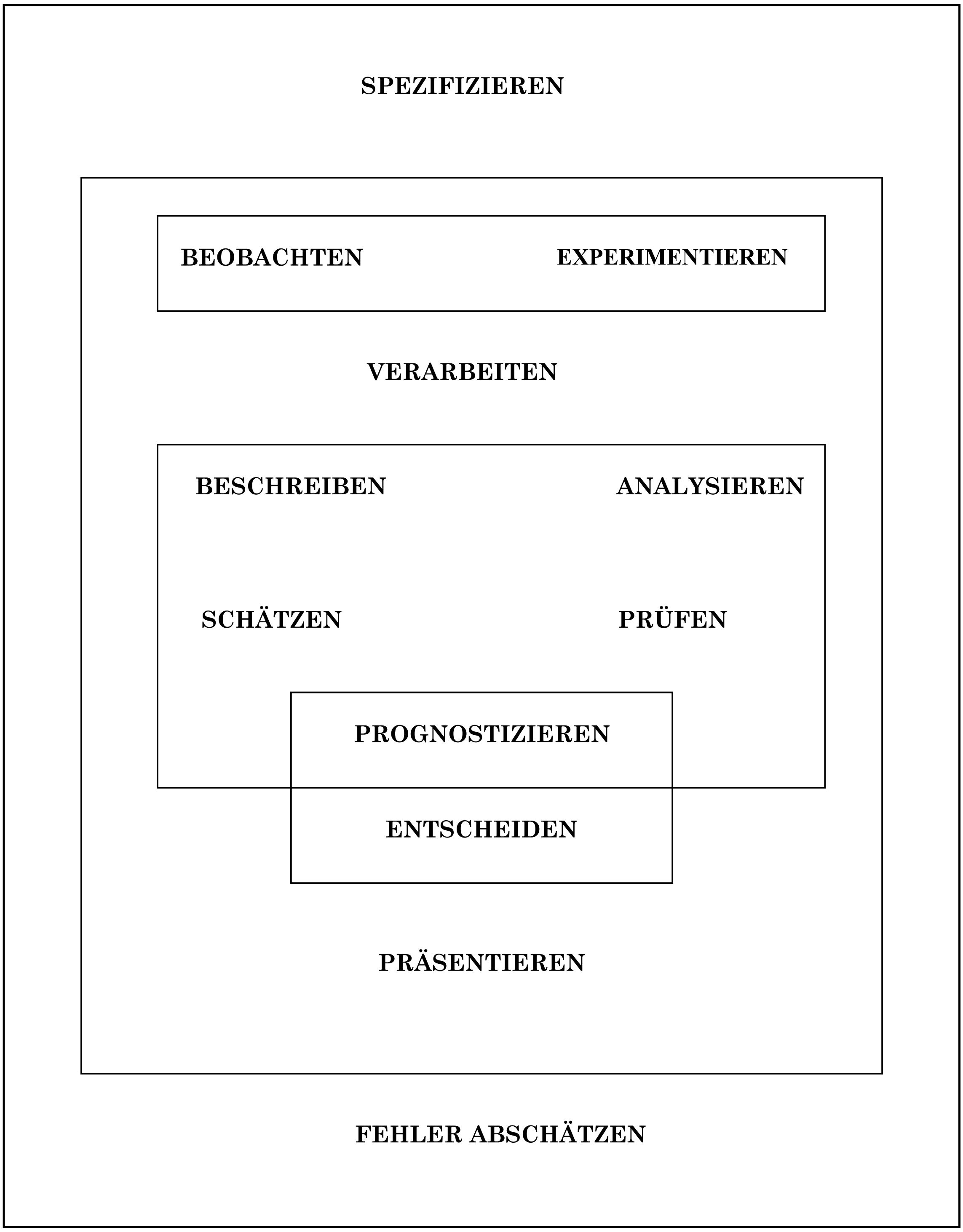
Die Methoden der induktiven, also der schließenden Statistik kommen darüber hinaus dann zum Einsatz, wenn nur ein Teil der Grundgesamtheit bekannt ist, deren Analyse für die Gesamtheit charakteristisch sein soll – Befragungen von Kunden oder Wählern oder Qualitätskontrollen bei Lebensmitteln sind Beispiele für diesen Fall. Aus Beobachtungen eines Teils der Grundgesamtheit soll nun auf die Eigenschaften der Grundgesamtheit als Ganzes geschlossen werden, Um hier möglichst präzise Aussagen treffen zu können, muss die Auswahl der Teilmenge nach statistischen Methoden erfolgen. Außerdem erlaubt es die Statistik, die Genauigkeit der mit diesen Verfahren getroffenen Aussagen zu bestimmen.

Mit diesen grundlegenden Definitionen können wir schon einige wesentliche Aussagen über die Arbeitsmethoden der Statistik treffen:

- Zunächst ist es erforderlich, die zu beantwortende Frage möglichst genau zu spezifizieren.
- Danach werden die Daten erhoben, die für die Beantwortung der Frage notwendig sind. Das kann durch Beobachten im weitesten Sinne geschehen, aber auch mit Hilfe der Durchführung von Experimenten. Die Datenerhebung kann sich unter Umständen nur auf Teilmengen der Gesamtdaten beziehen.
- Die erhobenen Daten werden aufbereitet, analysiert, die Ergebnisse der Untersuchungen geprüft. Sind nur Teilmengen der Daten erhoben worden, ist es erforderlich, Ergebnisse mit Hilfe geeigneter Methoden zu schätzen.
- Bei manchen Fragen sollen zukünftige Entwicklungen aufgrund vergangener Trends prognostiziert werden. Es kann auch erforderlich sein, in Situationen unvollkommener Information oder ungewisser Ereignisse Entscheidungshilfen zu geben, welche Handlungsalternativen bevorzugt werden sollten.
- Nicht zuletzt müssen die gewonnenen Ergebnisse – und bei wissenschaftlichen Untersuchungen auch der eingeschlagene Weg der Analyse! – möglichen Interessenten präsentiert werden.
- Und schließlich gehört ebenfalls zu jeder statistischen Analyse eine bestimmte Form der Selbstvergewisserung – nämlich eine Abschätzung möglicher Fehler, die zu Ungenauigkeiten oder sogar fehlerhaften Ergebnissen führen können.

Abbildung 2.1. fasst die Elemente der statistischen Untersuchung noch einmal in einem Überblick zusammen.

Schaubild 2.1: Elemente statistischen Arbeitens



Quelle: nach Menges (1972), op.cit., i.

2.2 Daten sammeln

2.2.1 Statistische Einheiten und Merkmale

Träger der Daten ist die statistische Einheit, also jene Objekte, deren Merkmale je nach Fragestellung von Interesse sind. Die statistische Einheit ist damit der Träger der Information, die erhoben werden soll.[7] Zur Identifikation statistischer Einheiten, die zur Grundgesamtheit gehören sollen, auf die sich unsere Fragestellung richtet, können Kriterien angegeben werden, die objektiv und genau sein müssen. Die Kriterien sollen die Grundgesamtheit eindeutig bestimmen und gegenüber anderen statistischen Einheiten abgrenzen. Dabei wird in der Regel mindestens ein Kriterium zeitlicher, räumlicher oder sachlicher Art verwendet.

Beispiel:

> Die Bestimmung des Bruttoinlandsprodukts folgt dem so genannten Inlandskonzept, die Bestimmung des Bruttosozialprodukts hingegen dem so genannten Inländerkonzept.[8] Um beide Größen der Volkswirtschaftlichen Gesamtrechnung (VGR) bestimmen zu können, müssen die statistischen Einheiten „Inländer“ beziehungsweise „inländischen Wirtschaftseinheiten“ also exakt bestimmt werden. Dabei ist in der VGR für die Abgrenzung der Inländer die Staatsangehörigkeit ohne Bedeutung. Ständig im Inland befindliche Produktionsstätten zählen zu den inländischen Wirtschaftseinheiten, unabhängig von den Eigentumsverhältnissen; umgekehrt gehören ständig im Ausland gelegene Produktionsstätten im Eigentum von Inländern nicht zu den inländischen Wirtschaftseinheiten. Ausnahmen von dieser Regel bilden diplomatische und konsularische Vertretungen sowie Streitkräfte. Studenten zählen stets als Gebietsansässige ihres Herkunftslandes, auch wenn sie länger als ein Jahr im Ausland studieren.[9]

Die Anzahl ihrer Elemente ist der Umfang der Grundgesamtheit. Unterschieden werden kann hier noch in reale Grundgesamtheiten, die immer endlich sind – etwa die Einwohner eines Landes – und in fiktive Grundgesamtheiten, die theoretisch unendlich viele Elemente haben können – etwa die Menge der Würfe, die mit einem Würfel getätigt werden können.[10] Eine statistische Grundgesamtheit ist demnach die vollständige Menge der statistischen Einheiten mit gleichen Identifikationsmerkmalen.

Oft ist die Fragestellung, die mit Hilfe der Statistik beantwortet werden soll, nicht direkt auf die statistischen Einheiten gerichtet; das wäre etwa bei der Frage der Fall, wie viele Nutztiere in der Bundesrepublik Deutschland – Hühner, Schweine, Gänse etc. – in einem bestimmten Jahr leben. Meist richtet sich das Interesse auf eine oder einige der Eigenschaften der statistischen Einheiten, die als Prädikatsmerkmale oder statistische Variablen bezeichnet werden. Als (Prädikats-)Merkmalsausprägungen werden unterscheidbare Erscheinungsformen eines Merkmals bezeichnet.

Beispiel:

> Statistische Grundgesamtheit können die Einwohner einer bestimmten Kommune sein; statistische Einheiten sind damit die Menschen, die das dafür relevante Identifikationsmerkmal „Haupt- oder Nebenwohnsitz in der Kommune X“ erfüllen.
> Merkmale oder statistische Variable, die je nach Fragestellung untersucht werden können, wären zum Beispiel Geschlecht, Alter, Familienstand, Einkommen pro Jahr.
> Das Merkmal „Geschlecht“ hat zwei Merkmalsausprägungen: männlich und weiblich.
> Das Merkmal „Einkommen pro Jahr“ kann als Merkmalsausprägung alle Beträge (Euro/Cent) annehmen, hypothetisch nach oben unbegrenzt, der niedrigste Wert wäre die Null (ohne eigenes Einkommen). Zur statistischen Weiterverarbeitung bietet sich hier ein sinnvolles Gruppierungsverfahren in Größenklassen an.

Identifikationsmerkmale sind also Bestandteil der Definition der statistischen Einheit. Je zahlreicher die Identifikationsmerkmale sind, desto geringer ist in der Regel die Zahl der statistischen Einheiten in der jeweiligen Grundgesamtheit. Identifikationsmerkmale sind nicht Gegenstand der statistischen Erhebung, sondern ihre Voraussetzung. Prädikatsmerkmale bilden hingegen die Menge der statistisch zu analysierenden Eigenschaften der Grundgesamtheit. Je zahlreicher die Prädikatsmerkmale sind, desto umfassender und differenzierter gerät die statistische Untersuchung.

Eine letzte Unterscheidung soll noch hinsichtlich der statistischen Einheiten getroffen werden: Bestandseinheiten sind über zumindest einen gewissen Zeitraum – wie der Name schon sagt – beständig: Bodenflächen, Menschen, Aktiengesellschaften, Ehen. Ereigniseinheiten haben keine Erstreckung in der Zeit, sondern stellen „punkthafte“ Ereignisse dar: Geburten und Sterbefälle, rechtskräftige Verurteilungen, Zahl der Abiturexamen in einem bestimmten Jahr. Vergleichbar ist dies am ehesten mit Bestands- und Stromgrößen in der VGR oder mit der Unterscheidung in Bilanz und Gewinn- und Verlustrechnung bei Unternehmen.

2.2.2 Daten erheben und messen

Die Begriffe erheben und messen werden umgangssprachlich oft ohne Bedeutungsunterschiede verwendet. In der Wissenschaftssprache sind die Begriffe eng miteinander verbunden, in ihrem Gehalt aber doch klar voneinander geschieden. „Erhebung“ von Daten bezeichnet den „realen Vorgang der Kenntnisnahme und Sammlung der Elemente der Erhebungsmenge“, Messung hingegen die „abstrakt-mathematische Isomorphie zwischen den empirischen Ausprägungen des Phänomens einerseits und reellen Zahlen andererseits.“[11]

Der Begriff der Erhebung umfasst damit auch die technische und administrative Organisation der Sammlung von Daten. Wissenschaftliche Erhebungen müssen objektiv sein, das heißt die Messvorschriften müssen nachprüfbar und damit offen gelegt werden. Jede Erhebung muss aber in der Praxis nicht nur wissenschaftlichen, sondern auch ökonomischen Kriterien genügen. Jede Datensammlung verursacht Kosten und braucht Arbeitszeit. Kosten und Nutzen von Erhebungen sollten daher – zumindest näherungsweise – bestimmt und gegeneinander abgewogen werden.[12]

Statistische Daten werden in der Regel durch zwei verschiedene Erhebungsformen gewonnen: durch Experimente und durch Beobachtungen.

- Bei Experimenten werden die Bedingungen für Versuchsanordnungen ex ante festgesetzt und gegebenenfalls im Verlauf einer Untersuchung planmäßig variiert. Die Merkmalsausprägungen, die im Verlauf derartiger Experimente dann entstehen, werden gemessen und als Daten festgehalten. Beispiele für derartige Experimente sind etwa die Ziehung von Lottozahlen oder eine vorher festgelegte Anzahl von Würfen eines Würfelbechers.

- Aus Beobachtungen werden ebenfalls Daten gewonnen. Dabei misst ein Beobachter Merkmalsausprägungen so, wie sie in der Realität vorkommen. Beispiele für Beobachtungen wären etwa die Ergebnisse der Schuleingangsuntersuchungen von Kindern oder die Wirtschaftsdaten, die zur Volkswirtschaftlichen Gesamtrechnung aggregiert werden.

2.2.3 Primär- und Sekundärdaten

Werden Daten für bestimmte Fragestellungen neu erhoben oder gemessen und anschließend ausgewertet, präsentiert und analysiert, so spricht man von Primärstatistiken. Ihr Vorteil ist, dass sie exakt auf die jeweilige Fragestellung zugeschnitten werden können, mit anderen Worten: die Identifikationsmerkmale der Grundgesamtheit und die Prädikatsmerkmale oder statistischen Variablen werden eigens für die jeweilige Erhebung festgelegt. Dass hier der Kostenfaktor sehr schnell eine ganz erhebliche Rolle spielen kann, ist offensichtlich. Ein Beispiel: Die Erstellung einer Statistik über die spezifische Lärmbelastung der Bewohner einer Kommune erfordert eben entweder flächendeckende Lärmmessungen oder Stichproben, die dann mit genauen Einwohnerstatistiken und Verkehrszählungen in einem Rechenmodell kombiniert werden müssen.

Eine unabsehbare Datenfülle wird jedoch von den unterschiedlichsten Behörden, Verwaltungen, statistischen Ämtern, aber auch von Unternehmen und Forschungseinrichtungen bereits regelmäßig erhoben. Solche Statistiken können oft auch für andere Fragestellungen nutzbar gemacht werden. Statistiken, die aus derartigen Messungen oder Erhebungen abgeleitet werden, bezeichnet man als Sekundärstatistiken. Vor- und Nachteile dieser Art des statistischen Arbeitens verhalten sich umgekehrt zu den Primärstatistiken: Die Kosten sind in der Regel erheblich niedriger, dagegen passen die Statistiken oft nicht exakt zu der neuen Fragestellung. Probleme können sich hier sowohl bei der Definition der Grundgesamtheit als auch bei den erhobenen statistischen Variablen ergeben.

In den letzten Jahrzehnten wird sowohl von Seiten der amtlichen Statistik als auch von den Institutionen, die regelmäßig wiederkehrende große Datenerhebungen durchführen, häufig versucht, durch bestimmten Arten der Standardisierung die sekundärstatistische Arbeit zu erleichtern. Erhebungen, die für so genannte Mehrzweckstatistiken[13] durchgeführt werden, sind so angelegt, dass sie die Datenbedürfnisse verschiedenster Interessenten befriedigen können. So werden beim Mikrozensus[14] oder beim sozioökonomischen Panel[15] eine Vielzahl von Fragen an eine Stichprobe der Grundgesamtheit gestellt, um eine möglichst große Spannbreite statistischer Auswertungen zu ermöglichen.

Neben Erhebung mit Hilfe von Stichprobenverfahren gibt es die Vollerhebung als die „klassische“ Erhebungsart, die bei demographischen Fragen wie der Ermittlung der Zahl der Geburten und Sterbefälle und bei wirtschaftlichen Fragen wie der Ermittlung des Volkseinkommens nach wie vor von der amtlichen Statistik als Regelfall angesehen wird.

Der Vollständigkeit halber sollen auch noch so genannte „Tertiäranalysen“ erwähnt werden – das sind statistische Analysen, die nicht mit Originaldaten arbeiten, sondern verschiedene primär- oder sekundärstatistische Auswertungen miteinander in Vergleich setzen und analysieren.

2.3 Ordnen von Daten

2.3.1 Merkmalstypen

Wie wir im letzten Abschnitt ausgeführt haben, sagen Prädikatsmerkmale etwas über Eigenschaften der von uns untersuchten statistischen Einheiten aus. Wenn nun diese Aussagen weiter analysiert werden sollen, etwa indem wir sie in sinnvolle Gruppen zusammenfassen oder aus diesen Daten Berechnungen wie Mittelwerte oder Standardabweichungen vornehmen, dann ist es zunächst sinnvoll, sich über die verschiedenen Kategorien der Prädikatsmerkmale klar zu werden.

Zunächst können Prädikatsmerkmale in kategoriale, komparative und metrische Merkmale unterteilt werden:[16]

- Kategoriale Merkmale sind qualitativer Natur, sie können auch als Attribute der statistischen Einheiten in der Grundgesamtheit aufgefasst werden, die sie besitzen. Beispiele sind Geschlecht oder Religionszugehörigkeit, Herkunftsland von Importwaren oder Antriebsart neu zugelassener Kraftfahrzeuge. Die Menge der möglichen Ausprägungen des Merkmals ist endlich, die Zahl der Ausprägungen kann sich zum Teil aber verändern: Beim Merkmal „Geschlecht" ist das nicht zu erwarten, jedoch können durchaus neue Religionen oder Länder im Zeitverlauf entstehen.
- Komparative Merkmale sind durch bestimmte Kategorien und zusätzlich durch die Angabe einer Vergleichsrelation zwischen den Kategorien gekennzeichnet. Typische Merkmalsausprägungen sind „hoch – mittel – niedrig" oder „größer als", „schwerer als". Beispiele sind typische Fragen aus Erhebungen von subjektiven, so genannten „weichen" Daten zu Einstellungen und Werthaltungen befragter Personen, etwa „Wie hat sich Ihre Arbeitszufriedenheit im Vergleich zum Vorjahr entwickelt" mit den Antwortvorgaben: „viel schlechter, schlechter, gleich geblieben, besser, viel besser". Komparative Merkmale können oftmals quantitativ ausgedrückt werden.
- Metrische Merkmale erlauben nicht nur Vergleiche wie komparative Merkmale, sondern die Abstände ihrer Ausprägungen – also ihre Differenzen und Verhältnisse – lassen sich sinnvoll interpretieren. Beispiel sind Angaben über das Alter von Personen (in Jahren), den Primärenergieverbrauch (in Steinkohleeinheiten) oder den Stundenlohn (in der jeweiligen Währungseinheit des Landes). Quantitative Merkmale führen von vornherein zu Zahlen – mit oder ohne Maßeinheit – als Merkmalsausprägung.

Der Unterscheidung in qualitativ kategoriale, komparative und metrische Merkmale entspricht einer bestimmten Differenzierung der Mess-Skalen:

- Ein Merkmal ist nominal skaliert, wenn nur festgestellt werden kann, ob eine bestimmte Merkmalsausprägung vorhanden ist oder nicht: weiblich oder männlich, vorbestraft oder nicht vorbestraft und so weiter.
- Eine statistische Variable ist ordinal skaliert, wenn die Merkmalsausprägungen in eine natürliche oder sinnvoll festzulegende Reihenfolge zu bringen ist. Die Klausurnote 2,7 ist schlechter als 2,3, aber man kann nicht sagen, um wie viel Prozent schlechter die mit 2,7 bewertete Arbeit ist. Ähnliches gilt für den Abstand des Intelligenzquotienten zweier Personen oder für zeitlich auseinander liegende Selbsteinschätzungen des Gesundheitszustandes von ein und derselben Person.

- Kardinal messbare Skalen entsprechen quantitativ metrischen Merkmalsausprägungen. Ein Ernte-Ertrag von 100 Doppelzentnern pro Hektar ist das Vierfache eines Ertrages von 25 dz/ha. Solche Aussagen lassen sich immer treffen, wenn die entsprechende Skala einen sinnvoll definierten Nullpunkt hat; sie heißt dann „verhältnisskaliert". Nicht alle kardinal messbare Skalen erfüllen diese Eigenschaft. Ist kein sachlogisch begründeter „absoluter Nullpunkt" vorhanden, wie etwa bei der Temperatur in Celsius-Graden", spricht man von einer lediglich intervallskalierten Merkmalsausprägung. Es ergibt eben keinen Sinn zu sagen, 80 Grad Celsius warmes Wasser sei doppelt so warm wie 40 Grad Celsius warmes Wasser oder 4 Grad Celsius sei das Doppelte von 2 Grad Celsius.

2.3.2 Klassenbildung

Die erhobenen Daten werden zunächst einmal als so genannte „Rohdaten" notiert und in heutiger Zeit dann in geeigneter Form einer elektronischen Datenverarbeitung zugänglich gemacht. Für fortgeschrittene Analysen gibt es viele spezialisierte Datenverarbeitungsprogramme; der ganz überwiegende Teil der Standardauswertungen lässt sich mit dem Programm MS Excel[17] oder der neuesten Version von SPSS[18] komfortabel bewerkstelligen.

In einem nächsten Schritt müssen nun die beobachteten oder gemessenen Rohdaten nach den Ausprägungen von einem oder von mehreren Merkmalen geordnet werden. Die Häufigkeit des Auftretens bestimmter Merkmalsausprägungen kann dann den Rohdaten hinzugefügt werden. Die Auszählung und der Vergleich einer dieser Häufigkeiten führen schließlich zu einer ersten „einfachen" Häufigkeitsverteilung, die – wie der Begriff andeutet – dadurch gekennzeichnet ist, dass nur ein Prädikatsmerkmal der statistischen Einheiten betrachtet wird. Kombinierte Häufigkeitsverteilungen sind dagegen dadurch charakterisiert, dass mehrere Merkmale gleichzeitig, das heißt, miteinander verknüpft, betrachtet werden.

Die Rohdaten werden nun zweckmäßigerweise so geordnet, indem sie nach den verschiedenen Ausprägungen eines Merkmals gruppiert werden, sofern es sich um eine qualitative, also nominal skalierte Variable handelt, die als primäres Ordnungskriterium des Ausgangsmaterials verwendet wird. Handelt es sich dagegen um ein metrisches Merkmal, also um eine ordinal oder kardinal skalierte Variable, wird man die Rohdaten zunächst nach der Größe, also der Ausprägung dieser Variable ordnen. Die Spannweite ist damit als die Differenz zwischen dem Minimal- und dem Maximalwert der Variable definiert, die die Ausgangsdaten der Erhebung oder Messung aufweisen. Nun kann es sich auch bei kardinal skalierten Variablen ergeben, dass die eine oder andere Merkmalsausprägung in den Rohdaten mehrfach vorkommt; so mag das Körpergewicht der 50 Hörer und Hörerinnen einer bestimmten Vorlesung zwischen 47 kg und 110 kg liegen, was eine Spannbreite von 63 kg ergeben würde; zufällig könnten aber 3 Studierende 56 kg und 5 andere 72 kg wiegen. Fasst man diese zusammen, so erhält man die so genannte ungruppierte Häufigkeitsverteilung – siehe Schaubild 2.1. Diese Darstellung ist an sich noch nicht besonders aussagekräftig.

Die wichtigste „Kunst" der statistischen Analyse in diesem Stadium ist eine angemessene Wahl der Klassenintervalle, die wiederum bei gegebenen Rohdaten die Anzahl der Klassen bestimmt. Wie man das Material gruppiert, wie groß die Klassenintervalle sein sollen und ob die Intervalle die gleichen oder verschieden große Spannweiten haben sollen, liegt weit gehend im Ermessen des Statistikers – der jedoch bei dieser Festlegung immer die Ausgangsfragestellung bedenken muss, derentwegen die Untersuchung überhaupt angestellt wird.[19]

Körpergewicht	Anzahl	Körpergewicht	Anzahl
47	1	70	1
48	1	71	1
50	1	72	5
51	1	75	1
53	1	77	1
54	1	78	1
55	1	79	1
56	3	80	1
57	1	82	1
58	1	83	1
59	2	84	2
61	1	86	1
62	1	88	1
64	1	90	1
65	1	92	1
66	2	93	1
67	1	95	1
68	1	102	1
69	1	110	1

Schaubild 2.1: Körpergewicht von Studierenden einer Vorlesung Angaben in kg (ungruppierte Häufigkeitsverteilung)

Gewicht von – bis	Anzahl	Gewicht von – bis	Anzahl
41–50 kg	3	81–90 kg	7
51–60 kg	11	91–100 kg	3
61–70 kg	10	101–110 kg	2
71–80 kg	11	110–120 kg	–

Gewicht von – bis	Anzahl	Gewicht von – bis	Anzahl
41–60 kg	14	81–100 kg	10
61–80 kg	21	101–120 kg	2

Schaubild 2.2: Körpergewicht von Studierenden einer Vorlesung Angaben in kg (gruppierte Häufigkeitsverteilungen)

Die unterschiedlichen Gruppierungen in Schaubild 2.2 zeigen, wie verschieden die Interpretationen je nach Art der Gruppierung ausfallen werden. Die Rohdaten sollen durch die Gruppierung wirklichkeitsgetreu dargestellt werden. Die Gruppierung führt auf der einen Seite zu einem Informationsverlust, auf der anderen Seite aber zu einem Mehr an Aussagekraft durch diese Art der Aufbereitung der Rohdaten. Das „Mehr" sollte den Verlust qualitativ überwiegen – die Entscheidung ist, so schreibt Wagenführ,[20] „im einzelnen Fall weitgehend dem Fingerspitzengefühl des Statistikers überlassen".

2.4 Darstellung

2.4.1 Tabellen

Die Ergebnisse statistischer Analysen können sehr oft in Tabellen präsentiert werden. Dabei muss darauf geachtet werden, dass Tabellen alle zum Verständnis der dargestellten Zahlen notwendigen Informationen enthalten müssen. Dies sind:

- Eine Haupt-Überschrift, etwa: „Die Hauptposten der deutschen Zahlungsbilanz";
- eine Unter-Überschrift mit weiteren Präzisierungen, etwa: „1998–2004 in jeweiligen Preisen";
- die Angabe der Einheit, etwa: „in Mrd. Euro"; es ist immer sinnvoll, sehr große oder sehr kleine Zahlen durch die Wahl einer geeigneten Maßeinheit „lesbarer" zu gestalten;

- eine Vorspalte, in der die Prädikatsmerkmale benannt werden, in unserem Beispiel etwa: „Leistungsbilanz, Vermögensübertragungen, Kapitalbilanz, Veränderung der Währungsreserven, Restposten";
- diese Funktion erfüllt auch die Kopfzeile, in der – in unserem Beispiel – die verschiedenen Jahre aufgeführt werden;
- die Hauptspalten mit den jeweiligen Zahlenwerten;
- häufig enthält eine letzte Tabellenzeile und/oder -spalte eine Aufsummierung der Zahlenwerte, und schließlich enthalten
- Schlusszeilen die genaue Quellenangabe der Zahlenwerte.

Für die Darstellung von Zahlenwerten in Tabellen haben sich außerdem bestimmte Symbole eingebürgert, die in größeren statistischen Werken wie etwa dem Statistischen Jahrbuch für die Bundesrepublik Deutschland normalerweise in einer Zeichenerklärung aufgeführt werden:[21]

-	nicht vorhanden (Ergebnis ist gleich Null);
0	mehr als nichts vorhanden, aber weniger als die Hälfte der kleinsten Einheit, die in der Tabelle zur Darstellung gebracht werden kann;
·	kein Nachweis vorhanden; Zahlenwert ist unbekannt oder wird geheim gehalten;
…	Nachweis ist zum Zeitpunkt der Veröffentlichung der Tabelle noch nicht möglich, Daten können erst später erhoben werden;
X	Nachweis ist nicht sinnvoll, weil die Fragestellung hier nicht zutrifft (Angabe kann aus sachlichen Gründen nicht gemacht werden);
▌	ein Balken am Tabellenrand markiert eine grundsätzliche Änderung im Erhebungsverfahren der Daten, sodass ein zeitlicher Vergleich der Daten beeinträchtigt ist;
/	bei Stichprobenerhebungen: kein Nachweis, da die Genauigkeit der Ergebnisse nicht ausreicht;
( )	Nachweis unter dem Vorbehalt, dass das Ergebnis beträchtliche Fehler aufweisen kann, weil der Zahlenwert relativ unsicher ist;
*	vorläufige Ergebnisse, in ausländischen Statistiken oft auch mit „p" gekennzeichnet;
r	revidierte Zahlen; Angaben, die in einer vorhergehenden Veröffentlichung mit einem anderen Zahlenwert versehen waren; diese Angabe entfällt in vielen, auch amtlichen Veröffentlichungen leider ganz.

Eine Beispieltabelle enthält Schaubild 2.3 auf der folgenden Seite.

2.4.2 Schaubilder

Zur besseren Darstellung des statistischen Zahlenmaterials wird sehr häufig der Weg beschritten, die aufbereiteten Daten in Graphiken zu veröffentlichen. Bei jeder Statistik muss sich der Bearbeiter daher für eine Form der Darstellung entscheiden. Je nach Art des Datenmaterials wird man der einen oder anderen graphischen Variante den Vorzug geben. Auch hier bieten die Datenverarbeitungsprogramme seit einigen Jahren eine Vielfalt von Möglichkeiten an, die überdies – ein Minimum an Grundkenntnissen vorausgesetzt – sehr leicht zu handhaben sind. Die mit Abstand größte Verbreitung hat hierbei der „Diagramm-Assistent" von MS Excel gefunden, der eine große Vielfalt verschiedener Punkt-, Linien-, Stab- oder „Kuchen"-Diagramme anbietet.

Besonders leicht verständlich ist zunächst das Stabdiagramm. Durch die unterschiedliche Höhe oder Länge von Stäben oder Linien soll der Eindruck bestimmter Größenordnungen vermittelt werden.

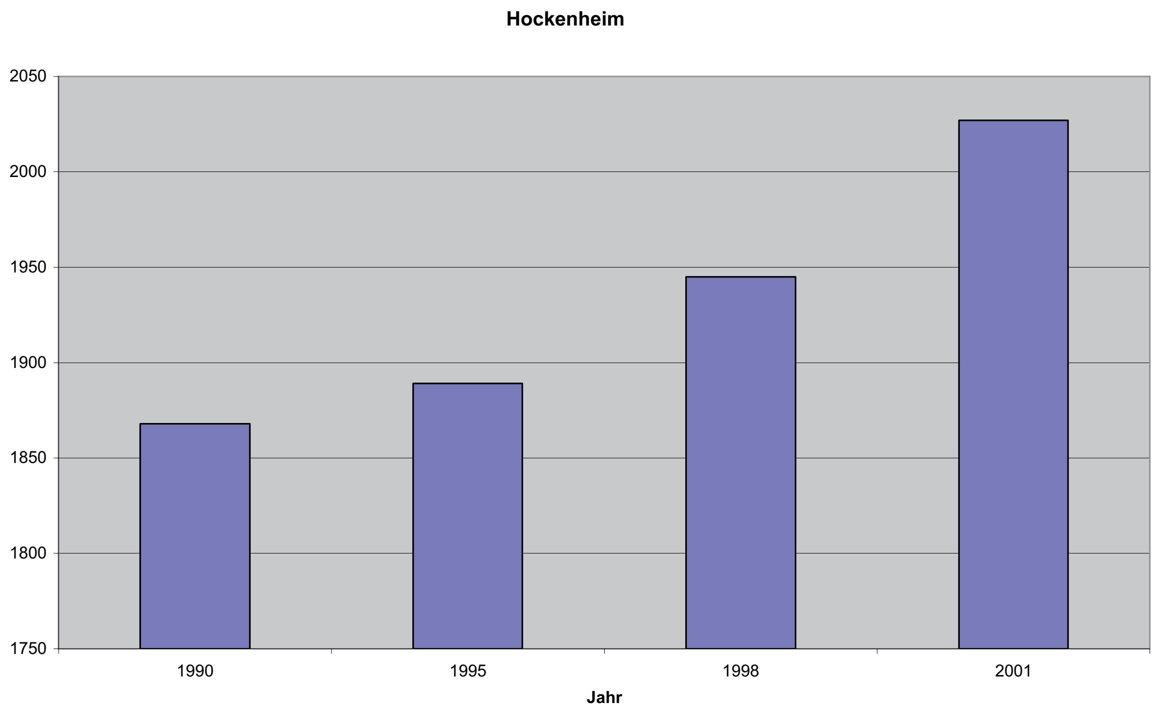


Schaubild 2.4: Stabdiagramm

Man sollte sich dabei aber vor Augen führen, dass hier viele optische Täuschungen möglich sind – die vom Urheber der Graphik beabsichtigt oder unbeabsichtigt sein können, denn die Wirkung der Graphik hängt von der Breite der Stäbe ab, ihrer Schraffur oder Farbe, dem Abstand voneinander und vielem mehr. Besonders beliebt ist dabei, die Aussagen durch eine geschickte Wahl der Achsenskalierung zu beeinflussen: Während in Schaubild 2.4 die y-Achse von 1.750 bis 2.050 kWh reicht, setzt Schaubild 2.5 die y-Achse mit einer Spannweite von 0 bis 2.050 kWh an – die dargestellten Daten sind exakt die gleichen!

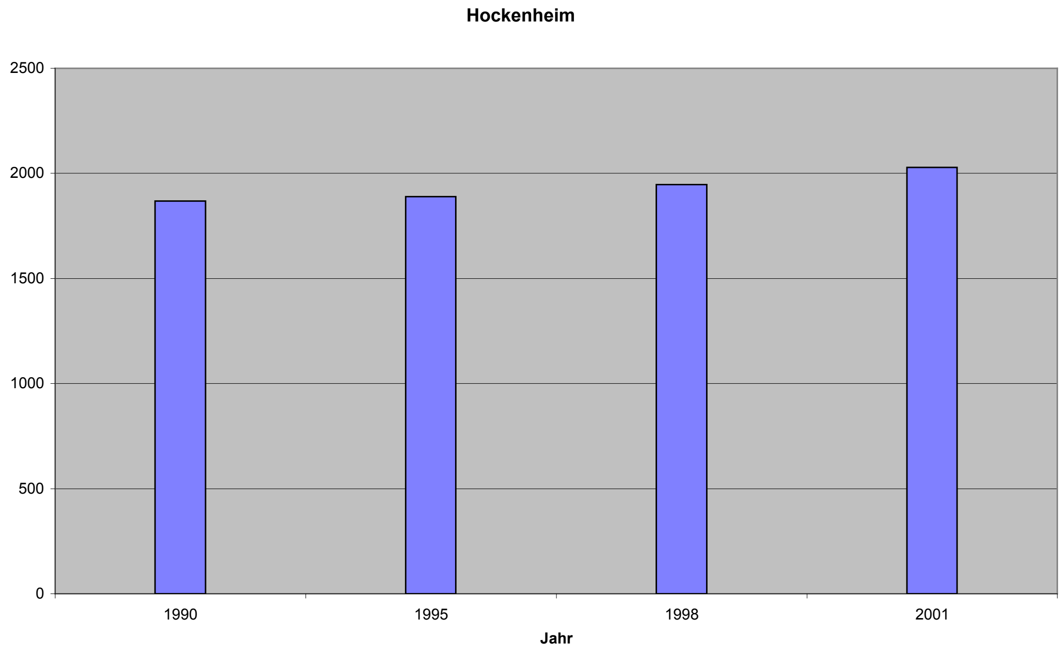


Schaubild 2.5: Stabdiagramm mit anderer Skalierung

Noch stärker treten die Veränderungen zurück, wenn die Daten in den heute so beliebten dreidimensionalen Graphiken dargestellt werden; das folgende Schaubild 2.6 beruht wiederum auf den exakt gleichen Werten wie die Schaubilder 2.4 und 2.5.

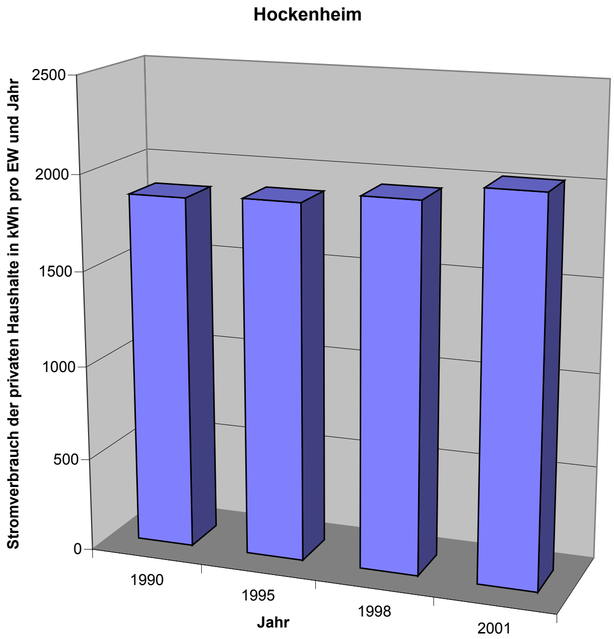


Schaubild 2.6: Dreidimensionales Stabdiagramm

Liniendiagramme erfüllen im Prinzip die gleiche Aufgabe wie Stabdiagramme, bringen in der optischen Darstellung jedoch viel stärker eine kontinuierliche Entwicklung zum Ausdruck.

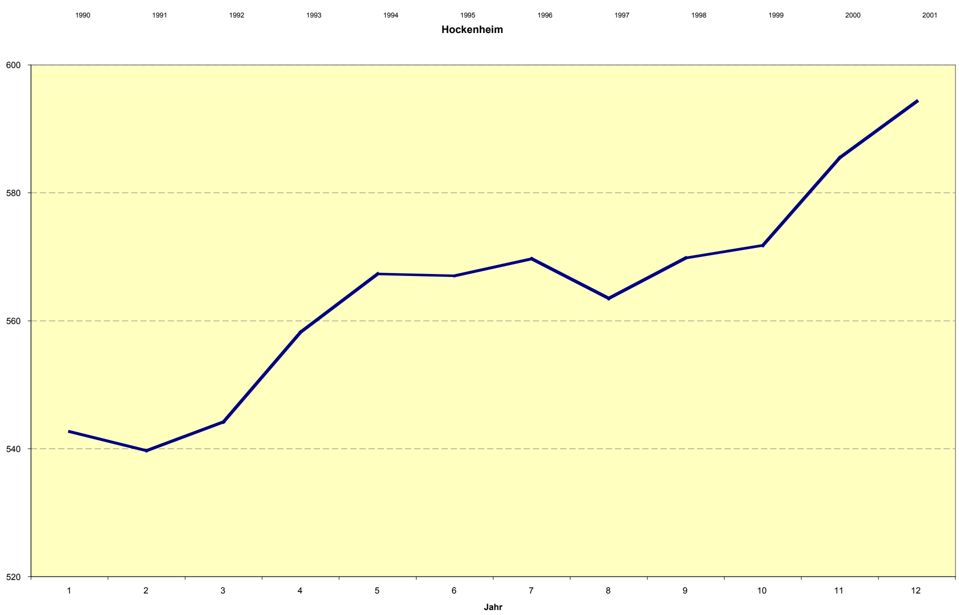


Schaubild 2.7: Liniendiagramm

Den Zusammenhang zweier Merkmale kann man in der Form eines Streudiagramms darstellen; diese Präsentationsart ist den Rohdaten am nächsten.

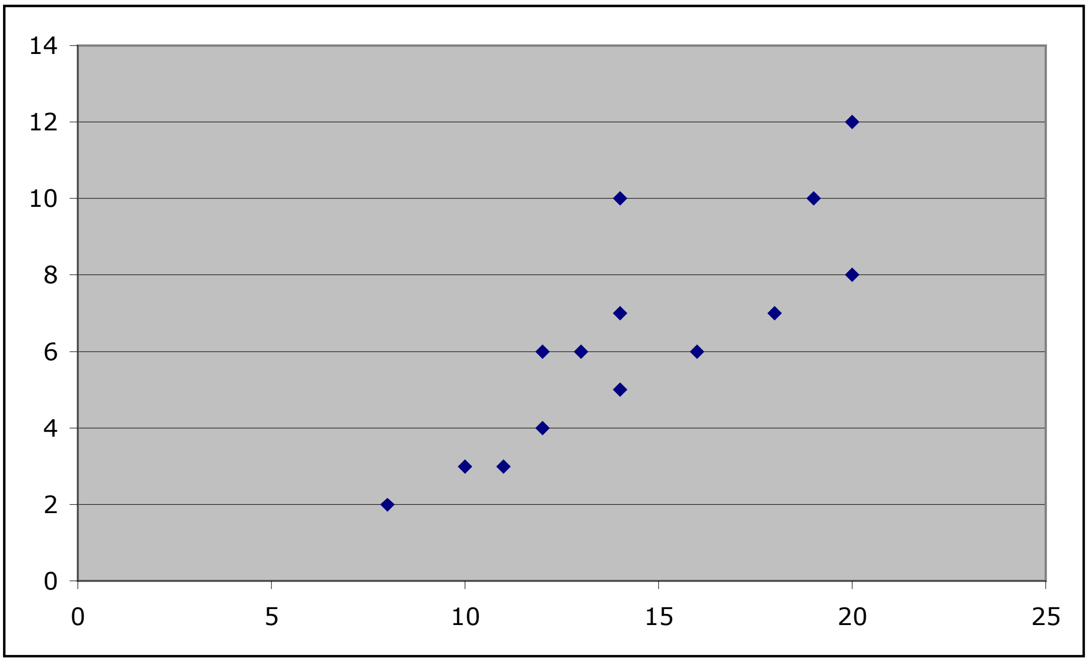


Schaubild 2.8: Streudiagramm

In diesem einfachen Beispiel lässt sich eine bestimmte Interpretation des Diagramms noch einmal näher lesen, wenn man eine Trendlinie einzeichnet.

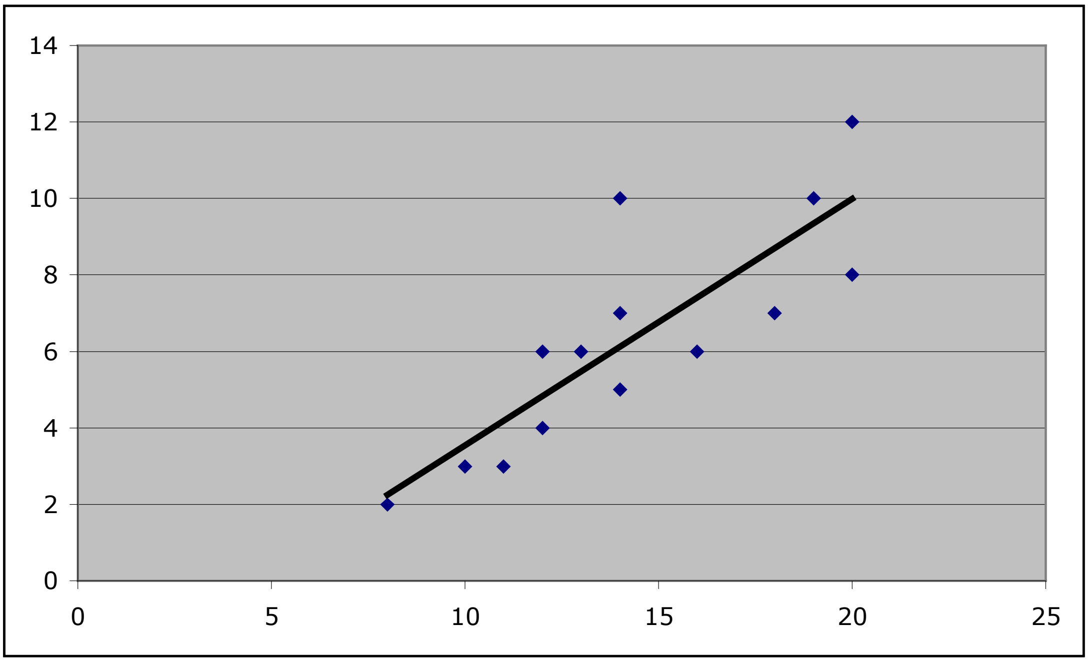


Schaubild 2.9: Streudiagramm mit linearer Trendlinie

Bei den Trendlinien wiederum kann man sich für die lineare Variante im letzten Schaubild oder für eine mathematisch anspruchsvollere Lösung entscheiden, die dann wiederum eine leicht andere Aussage transportiert, was an der hier eingezeichneten roten Exponentiallinie deutlich wird.

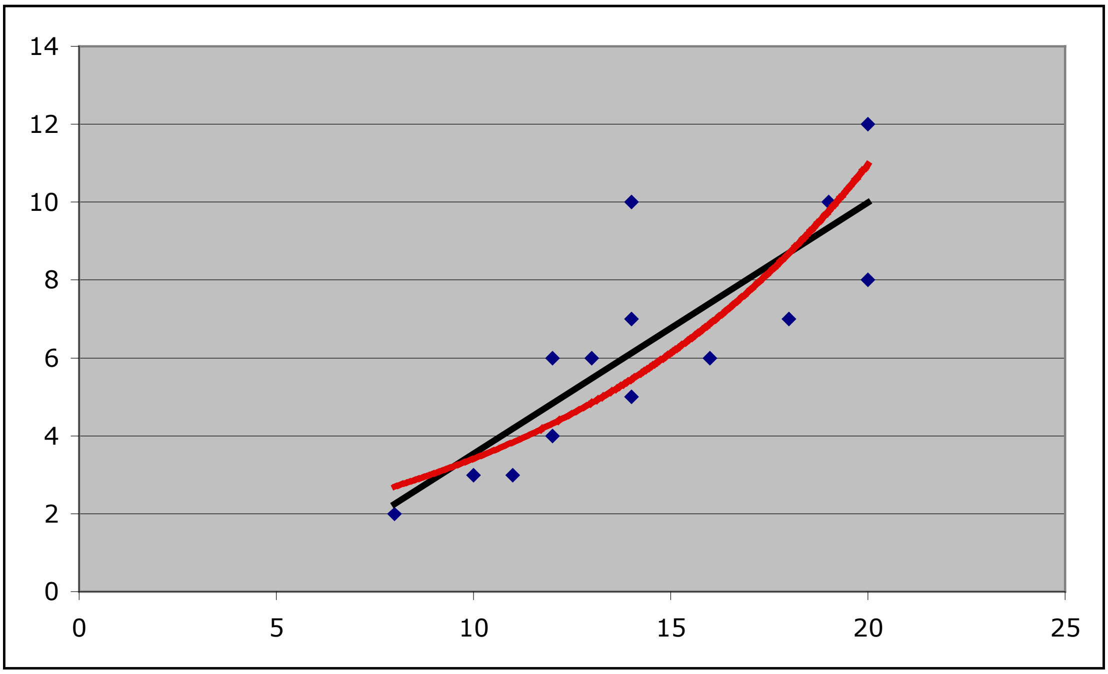
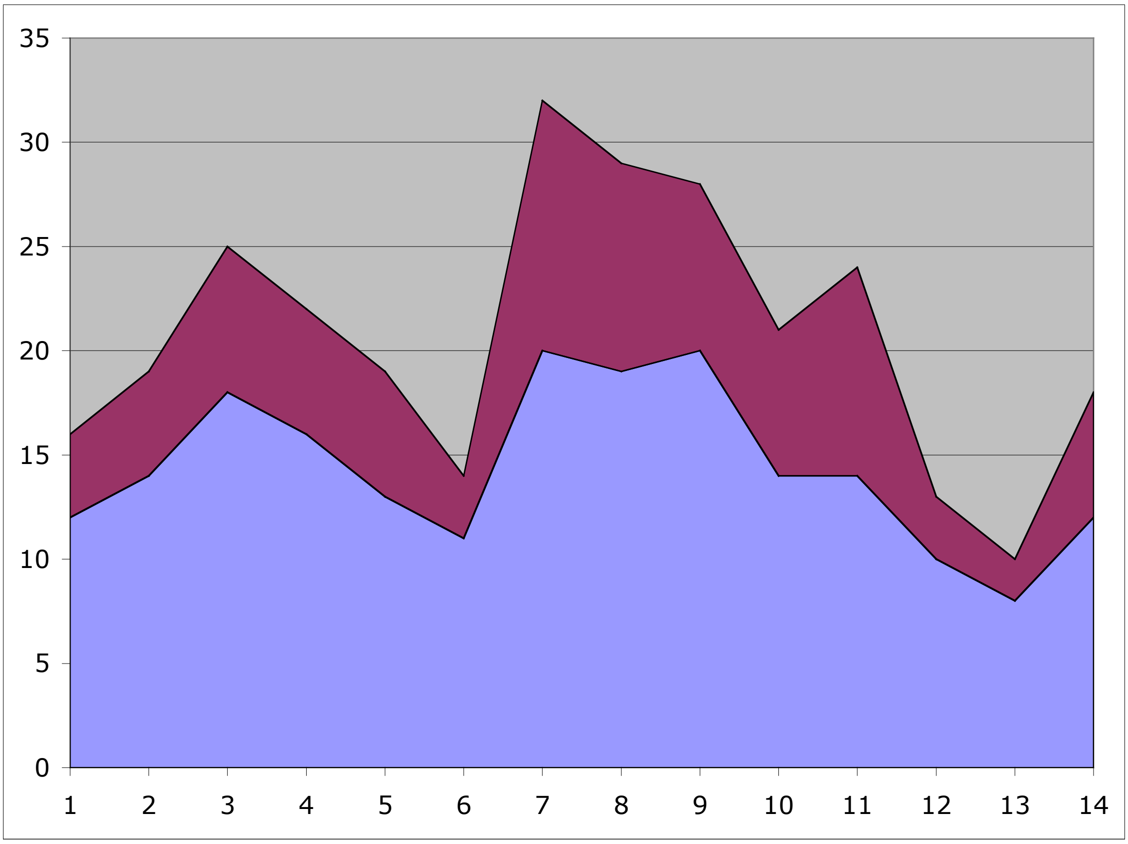


Schaubild 2.10: Streudiagramm mit Trendlinie und Exponential-Trend

Flächendiagramme werden oftmals genutzt, um die Eigenschaften von Balken- und Liniendiagrammen zu kombinieren.



3 Deskriptive Statistik

In der Statistik geht es zunächst darum, „Massenerscheinungen zu quantifizieren, um diese dann selbst zu interpretieren oder als Grundlage für Interpretationen anderen zur Verfügung zu stellen“ (Wagenführ (1971), S. 14). Diese Aufgabe der Statistik – die Qualifizierung von Massendaten – ist der Bereich der deskriptiven, also der beschreibenden Statistik. Hier werden Tatbestände quantifiziert, statistische Massen gebildet und beschrieben – zum Beispiel in Form von Häufigkeitsverteilungen. Statistische Massen bergen eine Fülle von Informationen, doch wenn sie nicht aufbereitet werden, bringen sie aufgrund der Unübersichtlichkeit wenig Nutzen. Die sinnvolle Erhebung, Aufbereitung, Zusammenfassung und Strukturierung der Information statistischer Massen ist Gegenstand des folgenden Kapitels. Es gibt verschiedene Methoden, die Informationen, die statistische Massen in sich tragen, darzustellen. Die wesentlichen sollen im Folgenden dargestellt werden.

3.1 Mittelwerte

Eine Form der Charakterisierung von Häufigkeitsverteilungen ist die Bildung eines Mittelwerts. Diese Maßzahl ist eine Art der Durchschnittsbildung. Sie gibt also an, welchen Wert eine Verteilung tendenziell annehmen wird. Dabei lassen sich, je nach Verteilung, eine Reihe verschiedener Mittelwerte unterscheiden.

3.1.1 Median

Definition: Der Median, oder auch Zentralwert, gibt in einer der Größe nach geordneten Reihe von Ausprägungen der Werte einer statistischen Variable den Merkmalswert an, bei welchem rechts und links gleich viele Ausprägungen liegen. Er ist eines der lagetypischen Mittel, die nicht durch Berechnung, sondern durch Gruppierung gewonnen werden.

Formel: $x_{med} = x_{\frac{n+1}{2}}$ für n ungerade; $x_{med} = 0{,}5 \cdot (x_{\frac{n}{2}} + x_{\frac{n}{2}+1})$ für n gerade

Beispiel: Das Einkommen verschiedener Personen beträgt 1.000, 1.500, 1.600, 2.000, 3.000, 4.500, 7.000 bzw. 10.000 Euro im Monat. Da es sich um acht Personen handelt, verwenden wir die untere der beiden Formeln und erhalten als Median den Wert 2500. Rechts und links dieses Wertes liegen jeweils vier verschiedene Einkommen.

Nehmen wir noch als neunte Einkommensgröße 12.000 Euro hinzu, verschiebt sich der Median, nun berechnet mit der oberen Formel, auf 3000. Auch hier liegen jeweils links und rechts vier Werte. Der Unterschied ist, dass bei geradem n der Median keiner der untersuchten Werte ist.

Anwendung: Der Vorteil des Medians liegt darin, dass extreme Werte nicht so einen großen Einfluss ausüben, er ist somit sehr robust. Er wird angewendet vor allem für ordinale Merkmale, aber auch für kardinale Merkmale. Allerdings kann der Median auch ein verzerrtes Bild liefern, wenn die Daten sich nicht schwerpunktmäßig in der Mitte der Reihe konzentrieren. Auch bei gruppierten Daten ist die Berechnung schwierig.

3.1.2 Modus

Definition: Der Modus ist der häufigste Merkmalswert einer Stichprobe.

Formel: $x_{\text{mod}} = x_i$, mit $\text{h}(x_i) > \text{h}(x_k)$, für alle $k \neq i$

Beispiel: In der statistischen Reihe 3, 4, 4, 6, 6, 6, 7, 8, 8 ist der Modus 6, da diese Ausprägung am häufigsten vorkommt.

Anwendung: Der Modus kann auf alle Arten statistischer Variablen angewendet werden. Er macht nur Sinn bei Vorliegen einer Häufigkeitsverteilung. Es kann auch vorkommen, dass mehrere Werte in einer Verteilung gefunden werden, die häufiger als alle anderen Werte und gleich häufig sind.

3.1.3 Das arithmetische Mittel

Definition: Das arithmetische Mittel gibt an, welchen Wert eine Ausprägung durchschnittlich annimmt.

Formel: $\hat{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$

Beispiel: Eine Schülerin hat folgende Noten in ihrem Zeugnis: 2, 2, 3, 3, 1, 2, 1, 4, 3, 2. Ihre Durchschnittsnote, also das arithmetische Mittel, ist demnach 23 geteilt durch 10, also 2,3.

Anwendung: Beim arithmetischen Mittel gehen alle Glieder der Reihe mit dem gleichen Gewicht ein; deren Reihenfolge spielt keine Rolle. Die Dimension darf allerdings, wie bei allen Mittelwerten, nie vergessen werden. Der Vorteil dieses Mittelwertes liegt in seiner einfachen Berechnung; außerdem ist er eindeutig bestimmt. Nachteilig ist, dass extremen Werten ein zu hohes Gewicht beigemessen wird. Das arithmetische Mittel kann einen Wert liefern, der in den Reihen gar nicht vorkommt. Deshalb ist er ein theoretischer Wert. Für qualitative Merkmale ist er nicht geeignet.

3.1.4 Das geometrische Mittel

Definition: Das geometrische Mittel gibt die n-te Wurzel der multiplikativ verknüpften n Ausprägungen an.

Formel: $x_g = \sqrt[n]{x_1 \cdot x_2 \cdot \ldots \cdot x_n}$, für $x_i > 0$

Beispiel: Die Wachstumsraten des Energieverbrauchs in Land A lagen in den vergangenen Jahren bei 10%, 5%, 8%, 2%, 4%. Das geometrische Mittel, also die durchschnittliche Wachstumsrate eines Jahres beträgt dann $(1{,}1 \times 1{,}05 \times 1{,}08 \times 1{,}02 \times 1{,}04)^{0{,}2}$, also rund 1,0578 oder 5,78%.

Anwendung: Das geometrische Mittel liefert niedrigere Werte als das arithmetische, die extremen Werte kommen nicht so stark zur Geltung. Außerdem kann es für Reihen mit negativen Zahlen oder der Null nicht berechnet werden. Wie das arithmetische ist auch das geometrische Mittel ein theoretischer Wert. Angewendet wird das geometrische Mittel bei multiplikativen Verknüpfungen, beispielsweise bei Verhältnissen.

3.1.5 Das harmonische Mittel

Definition: Das harmonische Mittel errechnet sich aus dem arithmetischen Mittel des Kehrwerts der Ausprägungen, von dem wiederum der Kehrwert berechnet wird.

Formel: $x_i = \dfrac{n}{\sum_{i=1}^{n} (1/x_i)}$

Beispiel: Herr Meier fährt von Mannheim nach Köln. Auf dem Hinweg ist die Straße frei und er erreicht eine Geschwindigkeit von 100 km/h. Auf dem Rückweg kommt er in einen Stau und schafft nur noch 50 km/h. Die Durchschnittsgeschwindigkeit beträgt nun 2/ (1/100 + 1/50) = 66,7 km/h. Die Gesamtstrecke beträgt damit 200 km, die dafür benötigte Zeit 3 Stunden.

Anwendung: Das harmonische Mittel wird nur selten verwendet. Es liefert, anders ausgedrückt, die Inverse des gewogenen arithmetischen Mittels der Inversen der Reihenwerte. Die beiden betrachteten Größen entwickeln sich also gegenläufig. Wenn die eine steigt (z.B. die Geschwindigkeit), dann sinkt die andere (z.B. die Fahrzeit) und umgekehrt. Wie aus dem Beispiel ersichtlich wird, geht es nicht nur um die Geschwindigkeit des Fahrzeuges, sondern auch darum, wie lange die Person unterwegs war. Bei Verwendung des arithmetischen Mittels würde nur die Geschwindigkeit, bzw. nur die Zeit berücksichtigt werden. Die Beziehung zwischen Geschwindigkeit und Zeit, die ja eine wesentliche Rolle spielt, würde vernachlässigt werden.

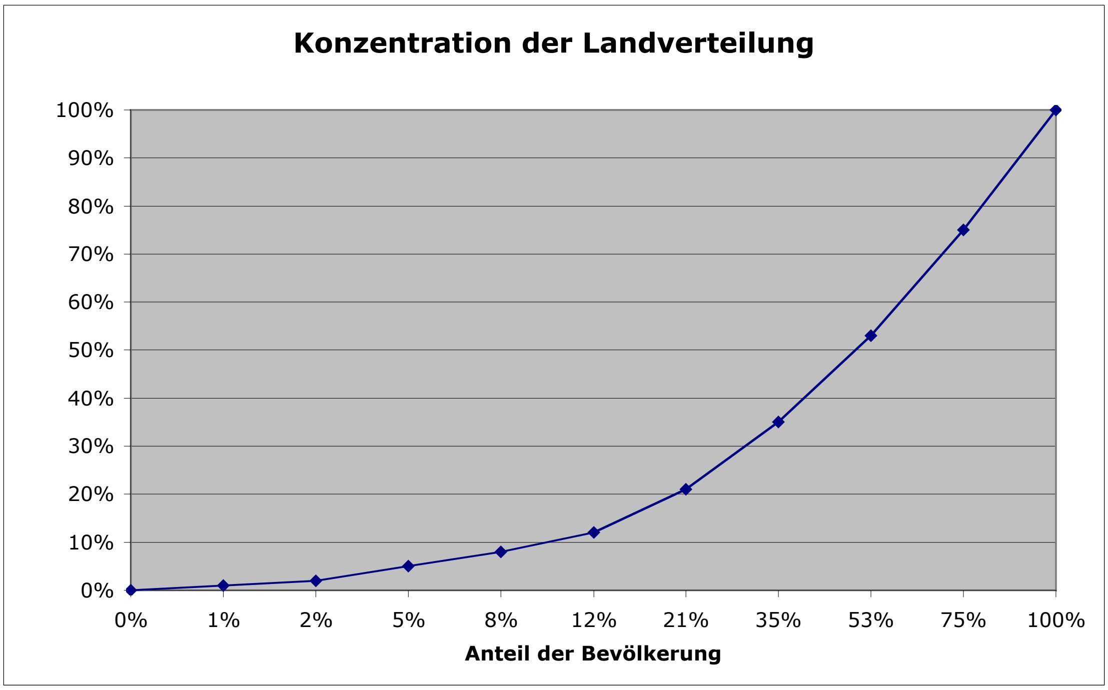
3.2 Konzentrationsmaße

Bei Konzentrationsmaßen geht es um die Frage, wie stark ein bestimmtes Merkmal in Bezug auf eine Gruppe von Merkmalsträgern geballt ist. Unter absoluter Konzentration versteht man das Ausmaß der Vereinigung eines erheblichen Anteils am gesamten Merkmalsbetrag einer geringen Zahl von Merkmalsträgern. Demgegenüber stehen die Disparitätsmaße, die die Abweichung eines Merkmals über die Merkmalsträger von der Gleichverteilung charakterisieren (Wagenführ, (1971), S. 127). Die Wahl des Konzentrationsmaßes hängt wiederum von der Fragestellung ab.

3.2.1 Lorenzkurve

Die Lorenzkurve ist eine grafische Darstellung der Konzentration. Sie wird zu den Disparitätsmaßen gezählt. Dabei werden in einem Diagramm auf der vertikalen Achse die prozentuale Verteilung der kumulierten Merkmale (v_i) und auf der horizontalen die prozentuale Verteilung der Merkmalsträger dargestellt. Daraus lässt sich dann ein Ungleichheitskoeffizient (Ginikoeffizeint) errechnen. Dieser ergibt sich aus der Fläche oberhalb der Lorenzkurve und unterhalb der Diagonalen, geteilt durch die Fläche des unteren Dreiecks (vgl. Schaubild 3.1).

Beispiel:



Bevölkerung	0 %	10 %	20 %	30 %	40 %	50 %	60 %	70 %	80 %	90 %	100 %
Landverteilung	0 %	1 %	2 %	5 %	8 %	12 %	21 %	35 %	53 %	75 %	100 %

Das Schaubild gibt die Daten der Tabelle wider. Die Daten in der zweiten Reihe (Landverteilung) sind kumulierte Werte. Das Schaubild zeigt, dass 10% der Bevölkerung 1% des Landes besitzen, 20 % besitzen 2% usw.

Anwendung: Um die Lorenzkurve zeichnen zu können, müssen einige Bedingungen erfüllt sein. Die Werte dürfen nicht negativ sein. Dann müssen sie der Größe nach geordnet werden. Schließlich sollte ein sinnvoller Zusammenhang zwischen den beiden betrachteten Größen bestehen.

3.2.2 Ginikoeffizient

Definition: Der Ginikoeffizeint ist das Verhältnis der Fläche zwischen der Diagonalen und der Lorenzkurve und der Fläche des Dreiecks unterhalb der Diagonalen.

Formel: $D_2 = \frac{\sum_{i=1}^{n} \sum_{j=1}^{n} (x_i - x_j)}{n^2}$; Ginikoeffizient: $D = \frac{D_2}{2 \cdot \hat{x}}$. Es gilt $D_2^2 = 2 \times s^2$ (Standardabweichung).

Alternativ kann auch die folgende Formel verwendet werden: $D = \frac{n + 1 - 2v}{n - 1}$, mit $v = \sum_{i=1}^{n} v_i$.

Beispiel: Wir verwenden das Beispiel von oben. Der Wert v als Summe der kumulierten Merkmale ergibt v = 3,12. Die Anzahl der Merkmale n beträgt zehn. Einsetzen in die Formel ergibt D = 0,876.

Anwendung: Der Ginikoeffizient nimmt den Wert 0 bei Gleichverteilung (die Lorenzkurve ist die Diagonale) an und geht gegen den Wert 1 bei völliger Ungleichverteilung (der Wert 1 wird nicht erreicht, da mindestens ein Merkmalsträger alles auf sich vereinigen muss).

3.2.3 Maße der absoluten Konzentration

Es gibt eine ganze Reihe von Maßen der absoluten Konzentration. Das einfachste Beispiel sind die Gliederungszahlen, die in Kapitel 5.2. besprochen werden. Doch sind diese in Bezug auf die Konzentration wenig aussagekräftig, wenn es darum geht, verschiedene Untersuchungen zu vergleichen. Zwei weitere Maße werden hier noch vorgestellt.

3.2.3.1 Hirschmann-Index

Definition: Der Hirschmann-Index ist die Summe der quadrierten Anteile der Merkmalsträger am gesamten Merkmalsbetrag.

Formel: $H(d_1, d_2, ..., d_n) = c \sum_{i=1}^{n} d_i^2$ mit c als wählbarer Konstante.

Beispiel: Wir bleiben bei dem Zahlenbeispiel der Landverteilung.

Dezile	Anteil am gesamten Land	Quadrat der Anteile
0–10%	0,01	0,0001
10–20%	0,01	0,0001
20–30%	0,03	0,0009
30–40%	0,03	0,0009
40–50%	0,04	0,0016
50–60%	0,09	0,0081
60–70%	0,14	0,0196
70–80%	0,18	0,0324
80–90%	0,22	0,0484
90–100%	0,25	0,0625
Summe	1	0,17

Wenn die Konstante c den Wert 1000 annimmt, dann ist der Hirschmann-Index 170.

Anwendung: Zu beachten ist, dass durch die Quadrierung die relativen Anteile der Merkmale mit hohem Anteil stärker gewichtet werden.

3.2.3.2 Rosenbluth-Konzentrationsmaß

Definition: Die Anteile der einzelnen Merkmalsträger werden der Größe nach geordnet und mit Rangziffern versehen. Daraus wird das Produkt gebildet und anschließend aufsummiert. Das Maß ergibt sich, wenn davon 0,5 subtrahiert und das Ergebnis verdoppelt wird. Der Kehrwert ist dann das Rosenbluth-Konzentrationsmaß.

Formel: $R(\alpha_i,\ldots,\alpha_n)=\dfrac{1}{2\cdot\left[\sum\limits_{i=1}^{n}(i\cdot\alpha_i)-0{,}5\right]}$ mit a als Anteilswert.

Beispiel: Auch hier verwenden wir die Zahlen zur Landverteilung.

Dezile und Rangziffer	Anteil am gesamten Land	Anteil mal Rangziffer
0–10%; 10	0,01	0,1
10–20%; 9	0,01	0,09
20–30%; 8	0,03	0,24
30–40%; 7	0,03	0,21
40–50%; 6	0,04	0,24
50–60%; 5	0,09	0,45
60–70%; 4	0,14	0,56
70–80%; 3	0,18	0,54
80–90%; 2	0,22	0,44
90–100%; 1	0,25	0,25
Summe	1	3,12

Einsetzen in die Formel liefert R = 0,19

Anwendung: Im Vergleich zum Hirschmann-Index haben die niedrigeren Anteilswerte ein höheres Gewicht.

3.3 Streuung, Schiefe, Wölbung

3.3.1 Streuungsmaße

Streuungsmaße geben an, um wie viel einzelne Werte voneinander oder von einem Mittelwert abweichen. Man kann zwischen rechnerischen und lagetypischen Streuungsmaßen unterscheiden. In unserer kurzen Übersicht werden wir aus der ersten Gruppe die durchschnittliche Abweichung, die Varianz und die Standardabweichung betrachten und aus der letzteren die Spannweite.

3.3.1.1 Die durchschnittliche Abweichung

Definition: Die durchschnittliche Abweichung d ist das arithmetische Mittel der absoluten Abweichungen von ihrem arithmetischen Mittel.

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

3.3.1.2 Die Varianz

Definition: Die Varianz ist die mittlere quadratische Abweichung vom arithmetischen Mittel.

Formel: $\sigma^2 = \frac{1}{n}\sum_{i=1}^{n}\left(x_i - \hat{x}\right)^2$ (Definitionsformel); $\sigma^2 = \frac{1}{n}\sum_{i=1}^{n}x_i^2 - \hat{x}^2$ (Rechenformel);

Wenn man die positive Wurzel der Varianz zieht, so erhält man die Standardabweichung s. Die Standardabweichung wiederum kann man durch das arithmetische Mittel teilen und erhält den Variationskoeffizienten v.

Beispiel:

Merkmalsträger	Gewicht (x_i) in kg	$\lvert x_i - \hat{x}\rvert$	$(x_i - \hat{x})^2$
1	52	19	361
2	65	6	36
3	65	6	36
4	58	13	169
5	82	11	121
6	76	5	25
7	79	8	64
8	91	20	400
∑	568	88	1212
1/n×∑	71	11	151,5

Für die durchschnittliche Abweichung ergibt sich ein Wert von d = 11 kg; für die Varianz σ^2 = 151,5 kg, für die Standardabweichung σ = 12,31 kg und für den Variationskoeffizienten v = 0,173.

Anwendung: Die durchschnittliche Abweichung ist ungenau im Vergleich zur Standardabweichung und wird deshalb kaum verwendet. Am häufigsten wird die Standardabweichung verwendet. Sie gilt als die zuverlässigste Art der Streuungsmessung. Alle Maße sind benannte Zahlen, d.h. die Dimension muss immer mit angegeben werden. Bei der Berechnung der Varianz können beide Formeln verwendet werden, die Rechenformel ist in der Regel einfacher anwendbar. Der Variationskoeffizient hat keine Dimension und eignet sich deshalb gut für Vergleiche.

3.3.1.3 Die Spannweite

Definition: Die Spannweite ist die Differenz der größten und der kleinsten Ausprägung der Merkmale der statistischen Reihe.

Formel: $Spannweite = x_{max} - x_{min}$

Beispiel: In unserem Zahlenbeispiel von oben ergibt sich für die Spannweite zwischen dem größten Wert (91) und dem kleinsten Wert (52) ein Wert von 39.

Anwendung: Die Spannweite ist leicht zu berechnen, aber Extremwerte haben ein sehr großes Gewicht.

3.3.2 Schiefemaße

Die Schiefe einer Verteilung gibt eine Auskunft über die Symmetrie einer Verteilung beziehungsweise über die Abweichung von der Symmetrie. Eine grobe Information über die Schiefe gibt die Anwendung der Fechner'schen Regel: Fallen die Mittelwerte Modus, Median und arithmetisches Mittel zusammen, liegt eine symmetrische Verteilung vor. Liegt das arithmetische Mittel vor dem Median und dem Modus, so ist die Verteilung rechtssteil; ist die Reihenfolge Modus, Median und dann arithmetisches Mittel, sprechen wir von einer linkssteilen Verteilung. Im Folgenden werden wir einige genauere Schiefemaße betrachten.

3.3.2.1 Schiefe nach Pearson

Definition: Das Pearson'sche Maß der Schiefe ergibt sich aus der Differenz des arithmetischen Mittels und des Modus geteilt durch die Standardabweichung.

Formel: $$sk = \frac{(\hat{x} - x_{\text{mod}})}{\sigma}$$

Beispiel: In Anwendung des Zahlenbeispiels zur Gewichtsverteilung aus 3.1. ergibt sich für
sk = (71 -65)/ 12,31 = 0,49

Anwendung: Der Wert für sk liegt zwischen -1 und 1. Die Verteilung ist symmetrisch bei sk = 0, linkssteil bei positiven Werten für sk und rechtssteil im anderen Falle.

3.3.2.2 Schiefekoeffizient nach Pearson

Definition: Der Koeffizient der Schiefe nach Pearson ist die Potenz der Summe der Abweichungen der einzelnen Werte vom Mittelwert in der dritten Potenz geteilt durch die Standardabweichung in der sechsten Potenz (die Varianz in der dritten Potenz).

Formel: $$\beta_1 = \frac{\left[\frac{1}{N}\sum_{i=1}^{k} f_i (x_i - \hat{x})^3\right]^2}{\sigma^6}$$

Beispiel: Wir verwenden wieder das bekannte Beispiel: $\beta_1 = (1/\ 8 \times 480)^2\ /\ 12{,}31^6 = 0{,}001$

3.3.2.3 Schiefekoeffizient nach Fisher

Definition: Der Koeffizient der Schiefe nach Fisher ist die Summe der Abweichung der einzelnen Werte vom Mittelwert in der dritten Potenz geteilt durch die Standardabweichung in der dritten Potenz.

Formel: $$\chi_1 = \frac{\frac{1}{N}\sum_{i=1}^{k} f_i (x_i - \hat{x})^3}{\sigma^3}$$

Beispiel: Hier ergibt sich für $\chi_1 = 0{,}032$

Anwendung: Der Koeffizient von Pearson ist wesentlich schärfer als das Maß von Fisher, gibt jedoch nicht an, ob eine Verteilung links- oder rechtssteil ist. Das leistet das Maß von Fisher, da es auch negative Werte erlaubt. Bei negativen Werten ist die Verteilung rechtssteil. Beide Koeffizienten werden 0 bei symmetrischen Verteilungen.

3.3.3 Wölbungsmaße

Die Wölbung lässt sich am leichtesten in einem Diagramm sehen, und zwar daran, wie stark die Spannweite auf der vertikalen Achse ist. Berechnen kann man sie mit den folgenden beiden Maßen.

3.3.3.1 Wölbungskoeffizient nach Pearson

Definition: Der Wölbungskoeffizient nach Pearson ist die Summe der Abweichung der einzelnen Werte vom Mittelwert in der vierten Potenz geteilt durch die Standardabweichung in der vierten Potenz.

Formel: $$\beta_2 = \frac{\frac{1}{N}\sum_{i=1}^{k} f_i (x_i - \hat{x})^4}{\sigma^4}$$

Beispiel: Für β_2 ergibt sich $(1/8 \times 340836)/\ 12{,}31^4 = 1{,}855$

3.3.3.2 Wölbungskoeffizient nach Fisher

Definition: Der Wölbungskoeffizient nach Fisher ergibt sich durch die Subtraktion der Zahl drei vom Wölbungskoeffizienten nach Pearson.

Formel: $\gamma_2 = \beta_2 - 3$

Beispiel: $\gamma_2 = 1{,}855 - 3 = -1{,}145$

Anwendung: Wenn $\beta_2 < 3$, dann ist die Verteilung flach gestreckt, bei $\beta_2 > 3$ scharf zugespitzt und bei $\beta_2 = 3$ ist sie normalverteilt. Analog gilt das für γ_2.

3.4 Regression und Korrelation

In den bisherigen Kapiteln wurden nur eindimensionale Häufigkeitsverteilungen betrachtet. Elemente der Grundgesamtheit wurden also nur nach einem Merkmal untersucht. Es ist aber auch möglich, mehrere Merkmale zu untersuchen. Dies wollen wir im Folgenden tun und betrachten den Fall der zweidimensionalen Häufigkeitsverteilung. Oft ist es so, dass zwischen zwei betrachteten Merkmalen ein Zusammenhag besteht. Diese Zusammenhänge können mittels der Regressions- und Korrelationsrechnung untersucht werden.

3.4.1 Lineare Einfachregression

Bei der Regressionsrechnung geht es darum zu untersuchen, ob zwischen zwei Variablen ein funktionaler Zusammenhang besteht. Dabei geht man davon aus, dass es eine endogene oder auch erklärte Variable Y sowie eine exogene beziehungsweise erklärende Variable X gibt. Es wirkt also X auf Y. Außer X gibt es auf Y auch noch weitere Einflüsse, die mittels eines Störterms dargestellt werden. Natürlich kann man auch noch weitere erklärende Variablen hinzufügen, doch werden wir uns auf eine erklärende Variable beschränken. Der einfachste Fall, den wir gleich betrachten, ist der des linearen Zusammenhangs.

Formel: $\chi_1 = \alpha + \beta \cdot x_1 + e_1$

Die empirischen Werte der beiden Merkmale X und Y kann man in einem Diagramm abtragen, wobei eine Punktwolke entsteht. Wenn ein linearer Zusammenhang aufgrund der Punktwolke nahe liegt, kann versucht werden, diese Punktwolke durch eine so genannte Regressionsgerade möglichst genau abzubilden. Die wahren Werte von X und Y liegen meist nicht auf der Geraden, sondern etwas davon entfernt. Der Abstand zur Geraden wird durch den Störterm e_t dargestellt. Um eine möglichst genaue Schätzung der Geraden zu erreichen, wird das Verfahren von Gauß, die Methode der kleinsten Quadrate, verwendet. Die Abweichungen e_t der empirischen Werte von der Gerade, also dem errechneten Wert von Y (= Y_r), werden quadriert und minimiert. Diese Methode liefert dann die beiden Parameterschätzungen a und b als Schätzwerte für den Ordinatenabschnitt und die Steigung der Regressionsgeraden $Y_r = a + b\,X_t$:

$$a = \hat{Y} - b \cdot \hat{x}\,;\; b = \frac{\operatorname{cov}(x,y)}{\sigma^2};$$

$$\text{mit } \hat{Y} = \frac{1}{T}\sum_{t=1}^{T} Y_t\;;\; \hat{x} = \frac{1}{T}\sum_{t=1}^{T} x_t \text{ und } \operatorname{cov}(x,y) = \frac{1}{T}\sum_{t=1}^{T}(x_i - \hat{x})(y_t - \hat{y})\;;\; \sigma^2 = \frac{1}{T}\sum_{t=1}^{T}(x_t - \hat{x})^2$$

Beispiel: Die folgende Tabelle gibt einen Überblick über Einkommen und Konsumausgaben einer Familie (in Tausend Euro):

Jahr	1998	1999	2000	2001	2002	2003	2004
Einkommen	32,3	33,1	33,4	33,8	34,5	34,7	35
Konsumausgaben	25,6	26	26,1	26,7	27,4	27,9	28,3

Die Angaben lassen einen linearen Zusammenhang zwischen dem Einkommen X_t und den Konsumausgaben Y_t vermuten. Deshalb wird die Regressionsgerade durch die Berechnung von a und b mittels der oben angegebenen Formeln ermittelt. Dazu bedienen wir uns einer Arbeitstabelle. Bei aufwendigeren Rechnungen wäre die Verwendung eines Computerprogramms hilfreich.

Jahr	Y_t	x_t	$Y_t - \hat{Y}$	$(Y_t - \hat{Y})^2$	$x_t - \hat{x}$	$(x_t - \hat{x})^2$	$(Y_t - \hat{Y})(x_t - \hat{x})$
1998	25,6	32,3	-1,26	1,59	-1,53	2,34	1,93
1999	26	33,1	-0,86	0,64	-0,73	0,53	0,63
2000	26,1	33,4	-0,76	0,58	-0,43	0,18	0,33
2001	26,7	33,8	-0,16	0,03	-0,03	0	0
2002	27,4	34,5	0,54	0,29	0,67	0,45	0,36
2003	27,9	34,7	1,04	1,08	0,87	0,76	0,9
2004	28,3	35	1,44	2,07	1,17	1,37	1,68
Σ	188	236,8	0	6,28	0	5,63	5,83
1/7 × Σ	26,86	33,83	0	0,9	0	0,8	0,83

Anhand der Formeln können wir nun b berechnen mit b = 0,83/ 0,8 = 1,042 und
a = 26,86 – 1,042 × 33,83 = – 8,39
Durch Einsetzen erhalten wir die Regressionsgerade mit $Y_r = -8{,}39 + 1{,}042 \times X_t$

Anwendung: Die Regressionsrechnung hat vor allem durch die Verwendung von Computerprogrammen eine weite Verbreitung gefunden und bildet mittlerweile einen wichtigen Zweig der empirischen Forschung, der Ökonometrie.

3.4.2 Korrelationsrechnung

Bei der Korrelationsrechnung wird versucht, einen Zusammenhang zwischen zwei (oder mehreren) Variablen in einer möglichst dimensionslosen Zahl auszudrücken. Hier beschränken wir uns auf den Fall von zwei Variablen. Die Korrelationsrechnung ermöglicht nicht die Feststellung eines kausalen Zusammenhangs. Welche Veränderliche die Verursachende ist, muss über die Sachlogik festgestellt werden. Möglicherweise gibt es auch mehrere Verursachende, während wir nur eine betrachtet haben. Auch zu unterscheiden ist zwischen einseitiger und zweiseitiger Beziehung von Variablen. Im letzteren Fall gibt es keine eindeutige Verursachende. Als Beispiel kann man die Noten in den Fächern Mathematik und Physik nennen.

Auch bei der Interpretation von Zusammenhängen muss man vorsichtig sein. Es können schon innerhalb einer Reihe Beziehungen zwischen den einzelnen Reihengliedern vorliegen. Typische Fälle wären die Autokorrelation, bei der die Veränderung einer Variablen von deren absoluter Höhe abhängt oder die Trendkorrelation. Hierbei liegen Reihen mit gleichem Trend vor, die eher über die Zeit korreliert sind, als über einen inneren Zusammenhang.

3.4.2.1 Kovarianz

Definition: Die Kovarianz der Werte zweier Merkmalsträger wird berechnet durch die Summe des Produkts der Abweichungen der Werte der beiden Merkmalsträger von ihrem Mittelwert.

Formel: $\operatorname{cov}(x,y) = \frac{1}{T}\sum_{t=1}^{T}(x_t - \hat{x})(y_t - \hat{y})$

Beispiel: Die folgende Tabelle gibt einen Überblick über Schadstoffemissionen und Atemwegserkrankungen.

Jahr	2000	2001	2002	2003	2004
Emissionen in t	56	62	63	67	70
Krankheitsfälle	1025	1087	1101	1150	1188

Zur Berechnung der Kovarianz errechnen wir zunächst die Mittelwerte der Emissionen und der Krankheitsfälle. Diese sind $\hat{x}$ = 63 und $\hat{y}$ = 1110. Damit lässt sich die Kovarianz ermitteln: Cov(x,y) = 317,8.

Anwendung: Diese Formel ist in der oben angegebenen Form geeignet für die Längsschnittanalyse. Wenn wir eine Querschnittsanalyse durchführen, es also um einen bestimmten Zeitpunkt geht, dann ersetzen wir das t durch ein i, das für verschiedene Merkmalsträger zu einem bestimmten Zeitpunkt steht. Beispielsweise könnte man auf diese Weise den Zusammenhang zwischen Körpergröße und Gewicht berechnen. Der Nachteil der Kovarianz ist ihre Dimensionsabhängigkeit. Wenn die Einheit des Merkmalsträgers sich ändert, dann ändert sich auch der Wert der Kovarianz. Um dem zu entgehen, wurde die Stichprobenkorrelation entwickelt, ein dimensionsloses Maß.

3.4.2.2 Der Korrelationskoeffizient nach Bravais-Pearson

Definition: Der Korrelationskoeffizient nach Bravais-Pearson ergibt sich durch die Division der Kovarianz durch die Standardabweichungen der jeweiligen Merkmalsträger.

$$\textit{Formel:}\ r_{xy} = \frac{\operatorname{cov}(x,y)}{\sigma_x \cdot \sigma_y}$$

Beispiel: Wir verwenden die Arbeitstabelle aus 4.2.1. Zur Berechnung des Bravias-Pearson'schen Korrelationskoeffizienten brauchen wir noch die Standardabweichungen der beiden Merkmalsträger: σ_x = 6,44 und σ_y = 62,63. r_{xy} = 0,788.

Anwendung: Mit dieser Berechnung wird eine Normierung des Maßes für die Strenge des linearen Zusammenhangs erreicht. Denn eine große Kovarianz kann auch durch die große Streuung einer der beiden Merkmalsträger verursacht sein. Der Koeffizient kann Werte zwischen -1 und 1 annehmen. Bei r_{xy} = 1 liegt eine vollkommene positive Korrelation vor. Das heißt, dass kleinen X-Werten kleine Y-Werte entsprechen, bzw. großen X-Werten entsprechen große Y-Werte. Bei r_{xy} = -1 liegt eine vollkommene negative Korrelation vor. Dabei entsprechen dann kleinen X-Werten große Y-Werte und umgekehrt.

3.4.2.3 Der Spearman'sche Rangkorrelationskoeffizient

Definition: Der Spearman'sche Rangkorrelationskoeffizient ergibt sich aus der Division der sechsfachen Summe der quadrierten Abweichungen der Ränge der Merkmale geteilt durch das Produkt des Quadrats der Merkmale minus eins mit der Anzahl von eins.

Formel: $\rho = 1 - \frac{6 \cdot \sum_{i=1}^{N} D_i^2}{N(N^2-1)}$ wobei $D_i = R_{xi} - R_{yi}$ (R = Rang des Merkmals x_i bzw. y_i)

Die betrachteten Merkmale werden der Größe nach sortiert, damit wird eine Rangfolge gebildet. Danach wird die Differenz der Ränge an der Stelle i gebildet. Aus den Differenzen kann dann der Koeffizient berechnet werden.

Beispiel: Die folgende Tabelle zeigt die Ränge von Sportlern bei einen 100m-Lauf und einem 200m-Lauf.

Sportler	1	2	3	4	5	6	7	8
Rang 100m-Lauf (R_{xi})	2	5	4	3	1	6	8	7
Rang 200m-Lauf (R_{yi})	3	5	2	6	4	8	7	1
D_i^2	1	0	4	9	9	4	1	36

In der weiteren Berechnung nimmt ρ einen Wert von 0,238 an.

Anwendung: ρ nimmt Werte zwischen -1 und 1 an. Bei großen Abweichungen der R wird D_i^2 groß und ρ geht gegen -1, bei großer Gleichheit der Ränge wird D_i^2 klein und ρ geht gegen 1. Hier wird nur der monotone Anteil des statistischen Zusammenhangs gemessen, streng monotone Veränderungen verändern ρ nicht. Der Spearman'sche Rangkorrelationskoeffizient wird zu den lagetypischen Korrelationskoeffizienten gezählt.

3.4.2.4 Der Kontingenzkoeffizient

Definition: Der Kontingenzkoeffizient ist die Wurzel aus dem Verhältnis von Chi-Quadrat mit der Summe aus Chi-Quadrat und der Anzahl der Merkmalsträger.

Formel: $K = \sqrt{\frac{\chi^2}{\chi^2 + n}}$ mit $\chi^2 = \sum_{i=1}^{r}\sum_{j=1}^{s} \frac{\left(n_{ij} - n_{ij}^*\right)^2}{n_{ij}^*}$ und $n_{ij}^* = \frac{n_{i\bullet} \cdot n_{\bullet j}}{n}$

Beispiel: Gegeben sei folgende Tabelle mit 100 Personen, die unterteilt sind nach dem Geschlecht und ihrem Rauchverhalten.

	Raucher	Nichtraucher	Summe
Männlich	15	30	45
Weiblich	30	25	55
Summe	45	55	100

Der Wert n_{ij}^{*} ist der Wert, der sich bei Unabhängigkeit einstellen würde. Er ergibt sich für das Kästchen oben links, wenn man die Summe der Raucher mit der Summe der Männer multipliziert und dann durch 100 teilt. Im Beispiel ist das der Wert 20,25. Für Chi-Quadrat ergibt sich der Wert 4,5. Der Kontingenzkoeffizient beträgt dann K = 0,21.

Anwendung: K nimmt Werte zwischen 0 und K_{max} an. Bei 0 sind die beiden Merkmale unabhängig, bei K_{max} liegt vollständige Abhängigkeit vor. Um verschiedene K-Werte besser vergleichen zu können, wurde der korrigierte Kontingenz-Koeffizient entwickelt.

Formel: $$K_{korr} = \sqrt{\frac{\chi^2}{\chi^2 + n} \cdot \frac{m}{m-1}}$$

mit m = min (r,s); r, s = Anzahl der Reihen und Spalten in der Tabelle.

Dadurch liegen die Werte zwischen 0 und 1. Bei 0 ist Unabhängigkeit, bei 1 Abhängigkeit gegeben.

3.5 Indikatoren und Indices

In diesem Kapitel geht es um die Darstellung statistischer Massen im Verhältnis zu anderen statistischen Massen. Das Verhältnis statistischer Massen zueinander liefert in der Regel mehr oder weniger übersichtliche Informationen über Sachverhalte als absolute Größen statistischer Massen. Für die weitere statistische Verarbeitung, so z.B. für die Anwendung multivariater Verfahren, bilden diese noch näher zu beschreibenden Verhältnisse einen Orientierungsrahmen.

3.5.1 Indikatoren

Definition: Ein Indikator setzt den Umfang N einer statistischen Masse M_1 ins Verhältnis zum Umfang einer anderen statistischen Masse M_2.

Formel: $$I = \frac{U_{M_1}}{U_{M_2}}$$

Indikatoren werden auch Verhältniszahlen genannt. Das Aneinanderreihen von Indikatoren gleicher Art und unterschiedlicher Periode bzw. Zeitpunkts nennt man eine statistische Reihe.

Indikatoren lassen sich in verschiedene Untergruppen unterteilen:

- Gliederungszahlen
- Beziehungszahlen
- Mess- und Indexzahlen

3.5.2 Gliederungszahlen

Definition: Bei Gliederungszahlen wird der Umfang einer statistischen Teilmasse M_x ins Verhältnis zu der dazugehörigen Gesamtmasse M gesetzt.

Formel: $$G=\frac{U_{M_x}}{U_M}$$

Beispiel: Die Staatsquote bezeichnet das Verhältnis der Staatsausgaben zum Bruttoinlandsprodukt. Im Jahr 2004 betrugen die Staatsausgaben in Mrd. Euro 1.036,6 und das Bruttoinlandsprodukt 2.207,2. Die Staatsquote betrug somit 46,95% (Quelle: Statistisches Bundesamt, www.destatis.de).

Anwendung: Neben der schon erwähnten Verwendungsmöglichkeit bei der Darstellung der Konzentration finden Gliederungszahlen eine große Verbreitung in der angewandten Wirtschafts- und Sozialstatistik.

3.5.3 Beziehungszahlen

Beziehungszahlen setzten im Gegensatz zu Gliederungszahlen zwei ungleiche statistische Massen ins Verhältnis zueinander. Allerdings besteht zwischen diesen beiden statistischen Massen ein sinnvoller Zusammenhang. Es werden drei verschiedene Beziehungszahlen unterschieden:

- Häufigkeitsziffern: das Verhältnis einer Bewegungsmasse zu der korrespondierenden Bestandsmasse.

  Beispiel: Die Geburtenziffer, die sich aus der Anzahl der Lebendgeborenen und der Wohnbevölkerung zusammensetzt.

- Dichteziffern: das Verhältnis von zwei unterschiedlichen statistischen Massen, die aus einem ähnlichen Milieu stammen.

  Beispiel: Bevölkerungsdichte. Im Jahr 2004 betrug die Bevölkerung in Deutschland 82.498.000 Personen, die Fläche betrug 357.030 km^2. Daraus ergibt sich eine Bevölkerungsdichte von 231 (Quelle: Statistisches Bundesamt, www.destatis.de).

- Verursachungsziffern: das Verhältnis einer statistischen Masse zu einer anderen statistischen Masse, die die erstere verursacht.

  Beispiel: Umsatz je Beschäftigten in 1.000 Euro in der deutschen Industrie: Der Umsatz des Jahres 2004 in der deutschen Industrie belief sich auf 1.423,4 Mrd. Euro, die Anzahl der dort Beschäftigten auf 6.015.300. Daraus ergibt sich eine Verursachungsziffer von 236,6 (Quelle: Statistisches Bundesamt, www.destatis.de).

Anwendung: Beziehungszahlen finden eine sehr häufige Anwendung in den unterschiedlichsten Bereichen der Volkswirtschaftslehre. Beispiele sind Veröffentlichungen zur konjunkturellen Lage oder auch Nachhaltigkeitsberichte von Kommunen, die einen Überblick über die lokale Situation mittels Indikatoren geben.

3.5.4 Mess- und Indexzahlen

3.5.4.1 Messzahlen

Definition: Eine Messzahl ergibt sich aus dem Verhältnis einer Beobachtung y_t mit t = 1,…,T und der Beobachtung y_1 aus derselben Zeitreihe.

Formel: $M = \frac{y_t}{y_1} * 100$

Die Werte einer Zeitreihe kann man auf eine Periode dieser Reihe beziehen; diese Reihe wird dann Basisperiode genannt.

Beispiel: Die Kohlenmonoxidbelastung entwickelte sich in Deutschland wie folgt:

Jahr	1996	1998	2000	2001	2002
CO in 1000t	6167	5557	4925	4578	4316

Die Messzahlenreihe zur Basis 1996 lautet dann: 100; 90,1; 79,9; 74,2; 70,0

Anwendung: Auch Messzahlen sind in der Anwendung sehr stark verbreitet. Man findet fast immer Messzahlen(reihen), wenn die Situation eines Landes oder einer Region beschrieben wird.

3.5.4.2 Indexzahlen

Indexzahlen werden verwendet, wenn man zwei verschiedene Merkmale miteinander verbinden will. Eines der beiden Merkmale wird dabei rechnerisch konstant gehalten. Dies macht Sinn z.B. bei Umsatzzahlen, wo nicht direkt ersichtlich ist, worauf eine Änderung des Umsatzes zurückzuführen ist, auf eine Änderung des Preises oder der abgesetzten Menge. Die am häufigsten verwendeten Indices sind Preis- und Mengenindices. Bei Preisindices wird die Mengenkomponente konstant gehalten und bei Mengenindices die Preiskomponente. Die bekanntesten Indices sind die Preis- und Mengenindices nach Laspeyres bzw. Paasche.

3.5.4.2.1 Laspeyres-Index

Der Preisindex nach Laspeyres:

Definition: Der Preisindex nach Laspeyres setzt sich zusammen aus der Summe des Produkts der Preise der Berichtsperiode mit den Mengen der Basisperiode geteilt durch die Summe des Produkts der Preise und der Mengen der Basisperiode.

Formel: $$P_L = \frac{\sum_{i=1}^{n}(p_t \cdot q_{i0})}{\sum_{i=1}^{n}(p_{i0} \cdot q_{i0})}$$

wobei p den Preis des i-ten Gutes in Periode 0 (Basisperiode) bzw. in Periode t (Berichtsperiode) angibt und q die entsprechende Menge.

Beispiel: In der Basisperiode bezahlte ein Haushalt für Nahrung (Menge = 100) den Preis 4, für Kleidung (Menge = 150) den Preis 6 und für Wohnen (Menge = 300) den Preis 8. In der Berichtsperiode änderten sich Preise und Mengen wie folgt: Nahrung: Menge 120, Preis 5; Kleidung: Menge 180, Preis 5; Wohnen: Menge 300, Preis 9.

Der Laspeyres-Preisindex ergibt demnach (5 ×100 + 5 ×150 + 9 ×300) / (4 ×100 + 6 ×150 + 8 ×300) ×100 = 106,76.

Der Mengenindex nach Laspeyres:

Definition: Der Mengenindex nach Laspeyres setzt sich zusammen aus der Summe des Produkts der Mengen der Berichtsperiode mit den Preisen der Basisperiode geteilt durch die Summe des Produkts der Preise und der Mengen der Basisperiode.

Formel: $$M_L = \frac{\sum_{i=1}^{n} (q_i \cdot p_{i0})}{\sum_{i=1}^{n} (q_{i0} \cdot p_{i0})}$$

Beispiel: Der Mengenindex kann ebenfalls aus den oben genannten Daten berechnet werden: (4 ×120 + 6 ×180 + 8 ×300) / (4 ×100 + 6 ×150 + 8 ×300) ×100 = 107,03.

3.5.4.2.2 Paasche-Index

Der Preisindex nach Paasche:

Definition: Der Preisindex nach Paasche setzt sich zusammen aus der Summe des Produkts der Preise mit den Mengen der Berichtsperiode geteilt durch die Summe des Produkts der Preise der Basisperiode und der Mengen der Berichtsperiode.

Formel: $P_P = \frac{\sum_{i=1}^{n}(p_t \cdot q_t)}{\sum_{i=1}^{n}(p_{i0} \cdot q_t)}$

wobei p den Preis des i-ten Gutes in Periode 0 (Basisperiode) bzw. in Periode t (Berichtsperiode) angibt und q die entsprechende Menge.

Beispiel: Wenn wir die Zahlen des Beispiels oben (Laspeyres-Index) verwenden, ergibt sich der Paasche-Index aus (5 · 120 + 5 · 180 + 9 · 300) / (4 · 120 + 6 · 180 + 8 · 300) · 100 = 106,06.

Der Mengenindex nach Paasche:

Definition: Der Mengenindex nach Paasche setzt sich zusammen aus der Summe des Produkts der Preise mit den Mengen der Berichtsperiode geteilt durch die Summe des Produkts der Preise der Berichtsperiode und der Mengen der Basisperiode.

Formel: $M_P = \frac{\sum_{i=1}^{n}(q_{it} \cdot p_{it})}{\sum_{i=1}^{n}(q_{i0} \cdot p_{it})}$

Beispiel: Wir verwenden wieder die bekannten Zahlen und erhalten als Mengenindex nach Paasche: (5 · 120 + 5 · 180 + 9 · 300) / (5 · 100 + 5 × 150 + 9 · 300) · 100 = 106,33.

Anwendung: In der Praxis wird in der Regel der Laspeyres-Index verwendet. Das hat folgende Gründe: Der Erhebungsaufwand ist beim Laspeyres-Index wesentlich geringer als beim Paasche-Index, da nur eine der beiden Größen aus der Berichtsperiode bekannt sein muss. Beim Paasche-Index müssen in jeder Periode die aktuellen Größen erfasst werden. Dies hat auch zur Folge, dass sich beim Paasche-Index keine zusammenhängende Reihe ergibt, weil nämlich immer Gegenwartswerte als Bezugsgröße verwendet werden, und diese logischerweise in jeder Periode anders sind.

3.5.4.3 Wert-Indices

Ein Wert ist immer das Produkt aus einem Preis und einer Menge. Daher ist der Wertindex auch ein Produkt aus einem Preis- und einem Mengenindex.

Definition: Der Wertindex gibt das Verhältnis der Wertsumme der Berichtsperiode t zur Wertsumme der Basisperiode 0 an.

Formel: $$M_P = \frac{\sum_{i=1}^{n}(q_{it} \cdot p_{it})}{\sum_{i=1}^{n}(q_{i0} \cdot p_{it})}$$

W kann auch berechnet werden durch die Multiplikation von $P_l \cdot M_p$, bzw. aus $P_p \cdot M_l$. Dabei muss man natürlich noch beide Ergebnisse durch 100 teilen.

Beispiel: Verwendet man die oben angegebenen Zahlen, so ergibt sich für W:
$P_l \cdot M_p$ = 106,76 · 106,33 / 100 = 113,52. $P_p \cdot M_l$ = 107,03 × 106,06 / 100 = 113,52.

Anwendung: Wertindices finden häufige Anwendung vor allem im industriestatistischen Bereich.

3.6 Zeitreihen

Zeitreihen haben die Funktion, die Ausprägung einer Veränderlichen im Zeitablauf darzustellen. Hierbei werden Bestandsmassen einem Zeitpunkt und Bewegungsmassen einem Zeitraum zugeordnet. Zusätzlich zu der zeitlichen Festlegung muss auch eine räumliche Festlegung erfolgen. Die übliche Darstellungsform ist innerhalb eines Diagramms, wobei x (Wert) auf der Ordinate und t (Zeit) auf der Abszisse abgetragen wird. Das Zeitintervall hängt von der Fragestellung ab. Häufig verwendete Intervalle sind Tage, Monate, Quartale oder Jahre. Zeitreihen ermöglichen, Aussagen über Entwicklungen von Variablen im Zeitablauf zu treffen.

3.6.1 Begriffsbestimmung

Definition: Eine empirische Zeitreihe ist eine Sequenz oder zeitlich geordnete Folge von T Beobachtungen

$$x_1, x_2, \ldots, x_t, \ldots, x_T$$

einer statistischen Variablen X im Zeitpunkt, bzw. Zeitintervall t = 1, …, T

Bei Bestandsmassen ist der Abstand der einzelnen Zeitpunkte stets gleich, bei Bewegungsmassen haben die Intervalle die gleiche Länge. Weiterhin definieren wir noch Saisonabschnitte von j = 1, …, p.

Zeitreihen werden in vier Komponenten unterteilt:

- Trend Tr_{tj} oder säkulare Komponente, der eine langfristige Bewegungsrichtung angibt.
- Zyklus Cy_{tj}, eine durch konjunkturelle Ursachen bedingte Komponente, die eine zyklisch-freie Bewegung darstellt.
- Saison S_j, eine zyklisch-gebundene Bewegung, die also periodisch wiederkehrt und deshalb unabhängig ist von t.
- Rest U_{tj} oder irreguläre Komponente, bei der es sich um stochastische Einflüsse und Störungen handelt.

Trend und Zyklus lassen sich zu der sog. Glatten Komponente zusammenfassen:
$G_{tj} = Tr_{tj} + Cy_{tj}$.
Die vier Komponenten lassen sich entweder additiv oder multiplikativ verknüpfen:

- $X = Tr_{tj} + Cy_{tj} + S_j + U_{tj}$
- $X = Tr_{tj} \cdot Cy_{tj} \cdot S_j \cdot U_{tj}$

3.6.2 Trend und glatte Komponente

Der Trend kann als Grundzug einer Bewegung bezeichnet werden. Die Bewegung erstreckt sich über einen längeren Zeitraum, wobei die Länge unbestimmt ist. Trends sollten in Bezug auf historische Perioden betrachtet werden, sie stellen frei nach Schumpeter ein Stück Wirtschaftsgeschichte dar. Es ist zu beachten, dass sich die gebildeten Zeitabschnitte in eine Gesamtheit einfügen sollten. Das heißt, dass gravierende gesellschaftliche Umbrüche berücksichtigt werden müssen.

Wenn sich durch einen Krieg oder fundamentalen Systemwechsel die gesellschaftlichen Rahmenbedingungen ändern, dann ist eine Zeitreihe über solch einen Bruch hinaus kritisch zu betrachten. Im Folgenden werden verschiedene Verfahren zur Trendbestimmung vorgestellt. Häufig werden die glatte Komponente und der Trend gleichgesetzt. Deshalb werden wir nur noch den Trend betrachten.

3.6.2.1 Verfahren der halben Durchschnitte

Bei diesem sehr einfachen Verfahren wird die Zeitreihe in zwei Hälften geteilt. Von jeder Hälfte wird dann das arithmetische Mittel berechnet. Diese können dann in das Diagramm eingezeichnet und verbunden werden. Die Verbindungslinie gibt den Trend an. Der Vorteil dieses Verfahrens ist die einfache Handhabung. Extreme Werte werden ausgeschaltet und es gibt keine systematischen Fehler. Bei sehr beweglichen Zeitreihen sind diese Verfahren allerdings zu ungenau.

3.6.2.2 Verfahren der gleitenden Durchschnitte

Beim Verfahren der gleitenden Durchschnitte werden aus den jeweiligen aufeinander folgenden Werten Mittelwerte gebildet. Dabei ist es wichtig, die Anzahl k der verwendeten Werte festzulegen, die dann die ganze Zeit über gleich bleibt. Es wird unterschieden in gerade und ungerade k.

Fall 1: k ist ungerade.

Definition: Es wird das arithmetische Mittel von k aufeinander folgenden X-Werten gebildet. Dieses ergibt den Trendwert zur Mitte des Teilbereichs k.

Beispiel: k = 3

t_i	1	2	3	4	5	6	7	8	9	10	11	12
X_i	6	8	9	12	7	9	10	13	11	10	12	14
T_i		7,67	9,67	9,34	9,34	8,67	10,67	11,34	11,34	11	12	

Fall 2: k ist gerade

Definition: Es wird die Summe von k aufeinander folgenden X-Werten gebildet. Von diesen Summen werden jeweils zwei aufeinander folgende wiederum zusammengefasst und daraus das arithmetische Mittel gebildet, indem durch 2k geteilt wird. Alternativ können auch k-1 Werte ganz und der vorhergehende und der nachfolgende Wert je zur Hälfte zur Bildung des arithmetischen Mittels verwendet werden. Dieser Wert wird dann durch k geteilt und ergibt den Trendwert.

Beispiel: k = 4

t_i	1	2	3	4	5	6	7	8	9	10	11	12
X_i	6	8	9	12	7	9	10	13	11	10	12	14
T_i			8,875	9,125	9,375	9,625	10,25	10,875	11,25	11,625		

Der erste Wert ergibt sich aus 6 + 8 + 9 + 12 = 35. Der zweite Wert ergibt sich aus 8 + 9 + 12 + 7 = 36. 35 + 36 = 71. Diesen Wert teilen wir durch 2k, also durch 8 und erhalten als T-Wert T_2 = 8,875.

Alternativ:

t_i	1	2	3	4	5	6	7	8	9	10	11	12
X_i	6	8	9	12	7	9	10	13	11	10	12	14
T_i			8,875	9,125	9,375	9,625	10,25	10,875	11,25	11,625		

Der erste Wert ergibt sich aus $0{,}25 \cdot (0{,}5 \times 6 + 8 + 9 + 12 + 0{,}5 \cdot 7) = 8{,}875$.

Anwendung: Bei starker Trendkrümmung wird ein kleiner Wert für k gewählt, bei schwacher Trendkrümmung ein großer Wert. Das Verfahren sollte nur angewendet werden, wenn die Schwankungen regelmäßig und von gleicher Intensität sind. Es kommt sonst zu übersteigerten (niedrigen) Durchschnitten, wenn der betreffende Abschnitt konkav nach oben (unten) gerichtet ist.

3.6.2.3 Methode der kleinsten Quadrate

Die bei der Methode der gleitenden Durchschnitte auftretenden Probleme lassen sich vermeiden, wenn die Methode der kleinsten Quadrate angewendet wird. Dieses Verfahren haben wir schon weiter oben bei der Regressionsrechnung kennen gelernt. Die Regressionsformel ist dabei die Trendgleichung mit der Zeit als unabhängiger Variable. Unterscheiden muss man, ob ein linearer oder nicht-linearer Trend vorliegt. Dies ist aus dem Schaubild ersichtlich. Bei linearen Trends verwenden wir eine Trendgleichung ersten Grades: $Tr = a + b \cdot X$. Bei nicht-linearen Trends werden Trendgleichungen zweiten oder höheren Grades verwendet: $Tr = a + b \cdot X + c \cdot X^2$. Dies ist bei längerfristigen Betrachtungen häufiger der Fall. Es kann auch ein Exponentialtrend vorliegen, so dass wir folgende Gleichung verwenden: $Tr = a \cdot b^x$. Zu lösen ist diese Gleichung mittels Logarithmierung.

Beispiel: Zur Berechnung eines linearen Trends verwenden wir den Index der Produktivität der Erwerbstätigen. Um die beiden Parameter a und b der Regressionsgleichung zu berechnen, nehmen wir den Index als abhängige Variable y und die Zeitwerte als unabhängige Variable X.

Jahr	Index	X (Zeitwerte)	a	b×X	T
1991	86,5	1	87,509	1,191	88,7
1992	89,8	2	87,509	2,382	89,891
1993	90,2	3	87,509	3,573	91,082
1994	92,7	4	87,509	4,764	92,273
1995	94,3	5	87,509	5,955	93,464
1996	95,5	6	87,509	7,146	94,655
1997	97,3	7	87,509	8,337	95,846
1998	98,1	8	87,509	9,528	97,037
1999	98,7	9	87,509	10,719	98,228
2000	100	10	87,509	11,91	99,419
2001	100,8	11	87,509	13,101	100,61
2002	101,4	12	87,509	14,292	101,801
2003	102,2	13	87,509	15,483	102,992
2004	103,5	14	87,509	16,674	104,183
2005	104,6	15	87,509	17,865	105,374

Quelle: Statistisches Bundesamt, zitiert nach Deutschland in Zahlen 2006; S. 46; (Institut der deutschen Wirtschaft, Köln).

Die Parameter a und b berechnen wir nach dem bekannten Verfahren aus Kapitel 4.1.1., a = 87,509 und b = 1,191. $Tr = 87{,}509 + 1{,}191 \cdot X$

3.6.3 Saisonale Schwankungen

Saisonale Schwankungen lassen sich beschreiben als rhythmisch gebundene Bewegungen im Laufe eines Jahres. Kennzeichnend für die Bewegung ist die Amplitude, also die Differenz zwischen dem höchsten und dem niedrigsten Wert, sowie die Periode, die die Zeitspanne umfasst, die nötig ist, um eine Gruppe von Werten wieder zu ihrem Ausgangspunkt zurückkehren zu lassen. In der Regel werden 12 Monate als Periode verwendet. Als auslösende Faktoren für saisonale Schwankungen gibt es drei verschiedene Möglichkeiten:

- natürliche Ursachen (z.B. das Wetter)
- institutionelle Ursachen (z.B. feste Termine, Feiertage)
- Ungleichheiten des Kalenders ((Arbeits-)Tage im Monat)

Zur Bestimmung der saisonalen Schwankungen gibt es verschiedene Methoden.

3.6.3.1 Konstante additive Saisonfiguren

Voraussetzung für die Anwendung ist das Vorliegen von einer stabilen Saisonfigur, die Schwankung also einen konstanten Wert hat. Wir gehen von Quartalsdaten aus und dann ergeben sich vier saisonale Schwankungswerte (S_p): S_1, S_2, S_3, S_4. Aus der Ausgangsgleichung (unter Vernachlässigung von Cy) $X_{tj} = Tr_{tj} + S_j + U_{tj}$ errechnen wir zunächst den Trend und stellen die Gleichung um: $X_{tj} - Tr_{tj} = S_j + U_{tj}$. Nun werden die jeweiligen Mittelwerte bestimmt: $\frac{1}{T}\sum_{t=1}^{T}(X_{tj} - Tr_{tj}) = \frac{1}{T}\sum_{t=1}^{T} S_j + \frac{1}{T}\sum_{t=1}^{T} U_{tj}$. Da S_j unabhängig ist von t, schreiben wir einfach S_j. Der letzte Term der Gleichung ergibt 0, so dass $S_j = \frac{1}{T}\sum_{t=1}^{T}(X_{tj} - Tr_{tj})$. Die Ausgangswerte kann man dann noch um die Saisonziffer bereinigen. Das wird erreicht, indem die Saisonziffer von dem Ausgangswert abgezogen wird.

Beispiel:

Tag	Schicht	X_{tj}	Tr_{tj} (k=3)	X_{tI}-Tr_{tI}	X_{tII}-Tr_{tII}	X_{tIII}-Tr_{tIII}	X_{tj}^*
	I	6	-	-			5,66
1	II	8	7,67		0,33		8,08
	III	9	9,67			0,33	8,92
	I	12	9,34	2,66			11,66
2	II	7	9,34		-2,34		7,08
	III	9	8,67			0,33	8,92
	I	10	10,67	-0,67			9,66
3	II	13	11,34		1,66		13,08
	III	11	11,34			-0,34	10,92
	I	10	11	-1			9,66
4	II	12	12		0		12,08
	III	14	-			-	13,92
Summe				1	-0,34	0,34	

Die S_j ergeben sich, wenn man über die Summe das arithmetische Mittel bildet, also $S_I = 1 / 3 = 0{,}34$, $S_{II} = -0{,}08$, $S_{III} = 0{,}08$.
Als saisonbereinigter Wert ergibt sich für $X_{1II} = 8 -(- 0{,}08) = 8{,}08$.

3.6.3.2 Konstante multiplikative Saisonfiguren

Ist die saisonale Schwankung nicht konstant, sondern verändert sich proportional, dann wenden wir das multiplikative Modell an. Die Vorgehensweise ist analog zu 6.3.1. Zunächst berechen wir den Trend und stellen die Zeitreihengleichung um:

$X_{tj} / Tr_{tj} = S_j \times U_{tj}$. Bildung der Mittelwerte ergibt: $\frac{1}{T}\sum_{t=1}^{T}\frac{X_{tj}}{Tr_{tj}} = \frac{1}{T}\sum_{t=1}^{T} S_j \cdot U_{tj}$. Da S_j unabhängig ist und sich U_{tj} zu eins aufaddiert, ergibt sich S_j = $S_j = \frac{1}{T}\sum_{t=1}^{T}\frac{X_{tj}}{Tr_{tj}}$. Die Saisonbereinigung erfolgt hier durch die Division der X-Werte durch die Saisonziffer.

Beispiel:

Tag	Schicht	X_{tj}	Tr_{tj} (k=3)	X_{tI} / Tr_{tI}	X_{tII} / Tr_{tII}	X_{tIII} / Tr_{tIII}	X_{tj}^*
	I	6	-	-			5,747
1	II	8	7,67		1,043		8,122
	III	9	9,67			0,931	9,184
	I	12	9,34	1,285			11,494
2	II	7	9,34		0,749		7,107
	III	9	8,67			1,038	9,184
	I	10	10,67	0,937			9,579
3	II	13	11,34		1,146		13,198
	III	11	11,34			0,97	11,224
	I	10	11	0,909			9,579
4	II	12	12		1		12,183
	III	14	-			-	14,286
Summe				3,131	3,939	2,939	

Die S_j ergeben sich, wenn man über die Summe das arithmetische Mittel bildet, also $S_I = 3{,}131 / 3 = 1{,}044$, $S_{II} = 0{,}985$, $S_{III} = 0{,}98$. Die Saisonbereinigung ergibt für $X_{1I} = 6 / 1{,}044 = 5{,}747$.

3.6.4 Restschwankungen

Formel: im additiven Modell: $U_{tj} = X_{tj} - Tr_{tj} - S_j$;

im multiplikativen Modell: $U_{tj} = \dfrac{X_{tj}}{Tr_{tj} \cdot S_j}$.

Beispiel: Wir verwenden die Zahlen aus 6.3.1. bzw. 6.3.2. Als I_{2I} ergibt sich im additiven Modell $I_{2I} = 8 - 7{,}67 -(- 0{,}08) = 0{,}41$; im multiplikativen Modell ergibt sich $I_{2I} = 8 / (7{,}67 \times 1{,}044) = 1{,}039$.

4 Wahrscheinlichkeitsrechnungen

4.1 Grundbegriffe

In den ersten Kapiteln haben wir statistische Methoden dargestellt, die zur Beschreibung und Analyse von Grundgesamtheiten geeignet sind. Oft ist aber die Untersuchung von allen Elementen der Grundgesamtheit gar nicht möglich, vielleicht erscheint sie in der Praxis zu schwierig oder zu umständlich – oder sie ist einfach zu teuer. In diesen Fällen wird versucht, einen Ausweg zu finden, in dem nur Teile der Grundgesamtheit analysiert werden. Diese Teile müssen methodisch sorgfältig ausgesucht werden, und ebenso methodisch sorgfältig müssen dann die Rückschlüsse von der Analyse der Teile auf die Grundgesamtheit im Ganzen gezogen werden. Dieser Bereich der Statistik wird „schließende" oder induktive Statistik genannt, die wichtigsten Methoden dieses Gebietes werden im nachfolgenden Kapitel 5 vorgestellt.

Die schließende Statistik benötigt jedoch die Argumentation der Wahrscheinlichkeitsrechnung, um die Methoden, die hier zur Anwendung gebracht werden, ihrerseits zu begründen. Dies wiederum ist der Grund, weshalb wir in diesem Kapitel nun einige wenige Grundlagen der Wahrscheinlichkeitsrechnung vorstellen. Die Wahrscheinlichkeitsrechnung hat sich zu einem eigenen, sehr umfangreichen Forschungsgebiet der theoretischen Statistik entwickelt. Wir beschränken uns hier strikt darauf, die Aussagen der Wahrscheinlichkeitstheorie zu erläutern, die für die grundlegenden Arbeitssschritte der induktiven Statistik unbedingt erforderlich sind. Dazu werden in diesem Abschnitt zunächst einige wichtige Begriffe erläutert.

Zufallsexperiment:

ein Geschehen, das beliebig oft und unter den immer gleichen Versuchsbedingungen wiederholt werden kann, und dessen Ausgang unterschiedlich sein kann – etwa das Werfen eines Würfels oder das Ziehen von drei Karten aus einem gut gemischten Kartenspiel von 32 Karten.

Ereignis beziehungsweise *Zufallsereignis:*

bezeichnet das mögliche Ergebnis eines Zufallsexperiments – bei den obigen Beispielen also das Ergebnis eines Wurfes mit dem Würfel oder eine bestimmte Kombination von drei Karten. Der Begriff Zufallsereignis bezeichnet noch expliziter, dass es sich um das Ergebnis eines Zufallsexperimentes handelt, der Eintritt oder Nicht-Eintritt des Ereignisses also zufällig erfolgt.

Ereignisfeld oder *Ereignisraum:*

bezeichnet die Gesamtmenge **aller** möglichen Ergebnisse eines Zufallsexperiments – bei den obigen Beispielen also die Zahlen von eins bis sechs oder jede beliebige Kombination von drei aus 32 Karten.

Elementarereignisse und zusammengesetzte Ereignisse:

Ein Elementarereignis lässt sich nicht weiter in Teile zerlegen; der Wurf eines Würfels ist zum Beispiel ein solches Elementarereignis. Ein zusammengesetztes Ereignis besteht dagegen aus mehreren Elementarereignissen; so lässt sich die Ziehung von drei Karten in drei einzelne Ziehungen unterteilen; wird als Ereignis das Ergebnis „gerade Zahl" beim Würfeln definiert, so kann dies durch die Elementarereignisse 2, 4 oder 6 eintreten.

Unmögliche und *sichere Ereignisse:*

Wie die Begriffe schon aussagen: unmögliche Ereignisse können aufgrund der Versuchsanordnung des Zufallsexperimentes niemals eintreten, etwa die Ziehung einer grünen Kugel aus einer Urne, die mit roten, gelben und blauen Kugeln gefüllt ist; sichere Ereignisse sind so definiert, dass sie das gesamte Ereignisfeld als (zusammengesetztes) Ereignis definieren, im vorliegenden Beispiel: das Ziehen einer roten, oder gelben oder blauen Kugel.

Absolute und relative Häufigkeit:

Wird ein Zufallsexperiment n mal hintereinander ausgeführt, so bezeichnet die absolute Häufigkeit die Anzahl der einzelnen (elementaren oder zusammengesetzten) Ereignisse, die relative Häufigkeit die Anzahl der Ereignisse bezogen auf die Zahl der durchgeführten Zufallsexprimente.

Ein Beispiel: Aus einer Urne, in der sich jeweils 250 gelbe, rote, grüne und blaue Kugeln befinden, werden 100 Kugeln gezogen.

Das Resultat sei: 20 mal gelb, 26 mal rot, 31 mal grün, 23 mal blau.

Die absoluten Häufigkeiten sind: 20 + 26 + 31 + 23 = 100;

die relativen Häufigkeiten sind 20/100 = 0,2 – 26/100 = 0,26 – 31/100 = 0,31 – 23/100 = 0,23 mit dem Gesamtergebnis: 0,20 + 0,26 + 0,31 + 0,23 = 1,00.

Wahrscheinlichkeit und Gegenwahrscheinlichkeit eines Ereignisses:

Die Wahrscheinlichkeit eines Ereignisse p (e) ist eine Zahl zwischen 0 und 1, wobei p = 0 ein unmögliches Ereignis und p = 1 ein sicheres Ereignis beschreibt:

$0 \leq p\ (e) \leq 1$

Als Gegenwahrscheinlichkeit eines Ereignisses wird die Wahrscheinlich bezeichnet, dass dieses Ereignis nicht eintritt, also

$p'\ (e) = 1 - p\ (e)$

4.2 Rechnen mit Wahrscheinlichkeiten

4.2.1 Abzählregel

Definition:

Ist die Wahrscheinlichkeit, dass bei einem Zufallsexperiment ein bestimmtes Elementarereignis aus dem Ereignisfeld eintritt, gleich groß, dann errechnet sich die Wahrscheinlichkeit für das Eintreten dieses Elementarereignisses als Quotient aus:

1 / Anzahl aller Elementarereignisse

Bei zusammengesetzten Ereignissen errechnet sich die Wahrscheinlichkeit für das Eintreten dieses zusammengesetzten Ereignisses als Quotient aus:

Anzahl der günstigen Fälle / Anzahl der möglichen Fälle

Beispiel:

Die Wahrscheinlichkeit, bei einem Wurf eines Würfels eine sechs zu würfeln, beträgt 1/6, da es sechs möglich Ereignisse gibt.

Die Wahrscheinlichkeit, bei einem Wurf eines Würfels eine gerade Zahl zu würfeln, beträgt 1/2, da es drei günstige Ereignisse – 2, 4 und 6 – und 6 mögliche Ereignisse – 1, 2, 3, 4, 5 und 6 – gibt.

4.2.2 Multiplikationsregel

Definition:

Hat bei einem Experiment aus mehreren Teilexperimenten ein zusammengesetztes Ereignis A insgesamt x Elementarereignisse und ein zusammengesetztes Ereignis B insgesamt y Elementarereignisse, sind bei der Durchführung des Experiments insgesamt x mal y Ereignisse möglich, wenn sowohl A als auch B eintreten sollen.

Beispiel:

Die Wahrscheinlichkeit, bei einem Wurf zweier Würfel als Gesamtsumme 11 zu würfeln, beträgt 1/18 – es gibt zwei „günstige Fälle" – 5+6 und 6+5 – sowie 36 mögliche Ereignisse: jedes mögliche Ergebnis des ersten Würfels – 1 bis 6 – kombiniert mit jedem möglichen Ergebnis des zweiten Würfels – ebenfalls 1 – 6. Die Einzelwahrscheinlichkeit eines Kombinationswurfes bestimmt sich aus der Multiplikation der Einzelwahrscheinlichkeiten: 1/6 mal 1/6 = 1/36.

4.2.3 Die Additionsregel

Definition:

Bei einer Anzahl von x sich nicht überschneidenden Ereignissen ergibt sich die Wahrscheinlichkeit für das Eintreten eines dieser Ereignisse aus der Summe der Einzelwahrscheinlichkeiten der x Ereignisse.

Beispiel:

Die Wahrscheinlichkeit der Windstärken an einem Ort W ergibt sich durch folgende Tabelle:

Windstärken	*Wahrscheinlichkeit*
0–1	0.22
2–3	0.40
4–5	0.18
6–7	0.15
8–9	0.04
≥ 10	0.01

Das Wahrscheinlichkeit, dass die Windstärke 4 oder höher beträgt, liegt in diesem Fall dann bei 0.38.

4.3 Kombinatorik

4.3.1 Unterscheidungskriterien

Mit Hilfe der Kombinatorik wird untersucht, auf welche und wie viel verschiedene Arten eine gegebene Anzahl von Elementen angeordnet und zu Gruppen zusammengefasst werden kann.[22] Methoden der Kombinatorik werden genutzt, um relative Häufigkeiten zu berechnen. Außerdem sind sie wichtig, um Verteilungsmodelle der theoretischen Statistik abzuleiten.

An einem Beispiel sollen wesentliche Unterscheidungskriterien eines Experiments dargestellt werden: In einer Urne seien vier verschiedenfarbige Kugeln – rot, gelb, grün, blau. Wieviele Möglichkeiten gibt es, von den Kugeln zwei herauszugreifen?

1) *Findet das Experiment „mit oder ohne Zurücklegen" statt?*
   Zunächst muss geklärt werden, ob die erste Kugel, nachdem sie gezogen wurde, auf die Seite gelegt wird oder ob sie zurück in die Urne kommt. Wird die Kugel zurückgelegt, ist es möglich, zwei mal die gleiche Kugel zu ziehen: Nach der Multiplikationsregel gibt es hier 4 mal 4 = 16 Ereignisse, die Wahrscheinlichkeit, zwei mal die rote Kugel zu ziehen, beträgt 1/16.

   Bleiben wir bei dem Beispiel, legen nun aber die erste Kugel zur Seite. Dann verbleiben für die zweite Ziehung nur noch drei Kugeln, insgesamt also 4 x 3 = 12 Ereignisse. Eine bestimmte Kombination zweier Kugeln hat hier also die Wahrscheinlichkeit von 1/12.

2) *Spielt die Reihenfolge eine Rolle oder nicht?*
   Es muss nun geklärt werden, ob das Ergebnis „erst rot, dann blau" bei zwei Ziehungen als gleiches oder als verschiedenes Ereignis zur Ziehung „erst blau, dann rot" betrachtet wird; die kombinatorischen Ergebnisse (1) und (2) werten die beiden Ereignisse als verschiedene Ereignisse. Spielt die Reihenfolge jedoch keine Rolle, dann fallen 6 solche Paarkombination aus, und die Gesamtmenge der möglichen Ereignisse reduziert sich auf 10 (4 mal 4 minus 6).

Wird der zuletzt betrachtete Fall noch einmal dergestalt geändert, dass die bei der ersten Ziehung gezogenen Kugeln nicht wieder zurückgelegt werden, dann reduziert sich die Gesamtzahl der Möglichkeiten noch einmal um 4 – nämlich um die Fälle, in denen eine bestimmte Kugel zweimal hintereinander gezogen wird. Die Gesamtzahl der Ereignisse beträgt also 6.

Zählt die Reihenfolge der Ziehungen, so spricht man von „Permutation“; zählt die Reihenfolge dagegen nicht, verwendet man den Begriff der „Kombination“.

3) *Werden alle Elemente oder nur eine Teilmenge betrachtet?*
   Bei einer Kombinationsrechnung können alle Elemente in die Rechnung miteinbezogen werden, etwa bei der Frage: wie viele verschiedene Reihenfolgen gibt es, in der 20 nummerierte Kugeln aus einer Urne gezogen werden können? In den Abschnitten (1) und (2) dieses Kapitels haben wir den Fall betrachtet, in dem nur eine Teilmenge m aus der Gesamtheit von n Elementen analysiert wird (m = 2 Kugeln, die gezogen werden, aus einer Zahl n = 4 Kugeln insgesamt).

4) *Sind alle Elemente verschieden oder nicht?*
   Als weiteres Unterscheidungskriterium muss noch betrachtet werden, ob alle Elemente verschieden sind (in unserem Beispiel: verschiedenfarbige Kugeln oder Kugeln mit unterschiedlichen Nummern) oder ob bestimmte Elemente sich nicht voneinander unterscheiden lassen. So könnten in einer Urne mit 1000 Kugeln etwa 500 rote, 480 blaue und 19 silberne Kugeln und eine goldene Kugel sein.

Diese vier Kriterien können nun beliebig untereinander kombiniert werden, sodass sich 16 verschiedene Grundformen kombinatorischer Experimente ergeben. Einige Lösungsansätze sollen im Folgenden kurz beschrieben werden.

4.3.2 Lösungsansätze

Versuchsanordnung:

Permutation ohne Wiederholung, alle Elemente werden betrachtet, alle sind voneinander verschieden. Beispiel: In einer Urne sind 7 Kugeln mit den Zahlen von eins bis sieben. Wie viele verschiedene Möglichkeiten gibt es, sie aus der Urne zu ziehen?

Lösung:

Bei der ersten Ziehung gibt es sieben Möglichkeiten; da die Kugel nicht zurückgelegt wird, verbleiben bei der zweiten Ziehung noch sechs Möglichkeiten und so fort. Dies führt zur Gesamtzahl von 7! = 5040 Möglichkeiten, allgemein:

$$P_n = n \cdot (n-1) \cdot (n-2) \cdot \ldots \cdot 4 \cdot 3 \cdot 2 \cdot 1 = n!$$

Versuchsanordnung:

Permutation ohne Wiederholung, alle Elemente werden betrachtet, einige Elemente sind gleich. Beispiel: 10 Landarbeiter werden morgens zur Arbeit eingeteilt: Einer soll Bäume schneiden, einer jäten, zwei sollen Beete umgraben, sechs sollen Kirschen ernten. Wie viele Kombinationsmöglichkeiten gibt es?

Lösung:

Nun sind die zwei beziehungsweise sechs „Plätze" beim Umgraben und beim Kirschen ernten ununterscheidbar. Sind bei n Elementen jeweils K_1, K_2, …, K_n Elemente gleich, ergibt sich:

$$P_{n(k_i)} = \frac{n!}{k_1! \cdot k_2! \cdot \ldots \cdot k_i!}, \quad wobei \sum_i k_i = n$$

Versuchsanordnung:

Permutation ohne Wiederholung, alle Elemente werden betrachtet, alle sind voneinander verschieden. Beispiel: Acht verschiedene Gewinne werden nacheinander an acht Personen verlost, jede Person bekommt bei jeder Ziehung ein Los, von deren jeweils ein Gewinnlos und sieben Nieten sind. Wie viele Möglichkeiten gibt es für die Gewinnverteilung?

Lösung:

Theoretisch kann jede Person jeden Gewinn erhalten, im Extremfall können alle Gewinne an eine Person gehen. Allgemein gilt:

$$P_n = n^n$$

Versuchsanordnung:

Kombination ohne Wiederholung, alle Elemente werden berücksichtigt, alle Elemente sind verschieden. Beispiel: Wie viele Möglichkeiten gibt es, n Elemente ohne Wiederholung, das heißt, so anzuordnen, dass die Reihenfolge innerhalb der Anordnung kein Unterscheidungskriterium ist.

Lösung:

Das ist eine Fangfrage! Denn wenn alle Elemente ohne Wiederholung gezogen werden, dann ist ja die Reihenfolge der Ziehung das einzige Unterscheidungskriterium von Belang – und dies ist hier ausgeschlossen. Die Lösung ist also 1 – es gibt genau eine Möglichkeit, die diese Kriterienkombination erfüllt.

Versuchsanordnung:

Kombination ohne Wiederholung, nur die Teilmenge k von n Elementen werden berücksichtigt, alle Elemente sind verschieden. Beispiel: Zwei von acht Kollegen sollen ein Team bilden, um eine bestimmte Aufgabe zu lösen. Wie viele mögliche Teams gibt es?

Lösung:

In wie viel verschiedenen Anordnungen kann die Gruppe der Teilmenge k vorliegen? Offenkundig in k! verschiedenen Möglichkeiten. Aus der Multiplikationsregel kann man daher aus dem Fall der Permutation ableiten:

$$c_{nk} = \frac{n!}{(n-k)! \cdot k!}$$

Für den Ausdruck auf der rechten Seite der Formel kann man auch schreiben: $\binom{n}{k}$

(lies: n über k)

Versuchsanordnung:

Kombination mit Wiederholung, alle Elemente werden berücksichtigt, alle Elemente sind verschieden.

Lösung:

Die mathematische Ableitung wird hier aus Platzgreünden nicht angegeben; im allgemeinen Fall bestimmt sich die Lösungszahl durch:

$$c_n = \binom{2n-1}{n}$$

Versuchsanordnung:

Kombination mit Wiederholung, nur k von n verschiedenen Elementen werden berücksichtigt, alle Elemente sind verschieden.

Lösung:

k aus n verschiedenen Elementen werden so ausgewählt, dass Wiederholungen möglich sind, die Reihenfolge innerhalb der Anordnungen jedoch kein Unterscheidungskriterium bildet:

$$c_{nk} = \binom{n+k-1}{k}$$

4.4 Wahrscheinlichkeitsverteilungen

4.4.1 Wahrscheinlichkeitsbegriff und Begriff der Wahrscheinlichkeits-verteilung

„All diejenigen Ereignisse sind mit Wahrscheinlichkeiten behaftet, deren Eintreten ungewiss ist.“[23] Der klassische Wahrscheinlichkeitsbegriff bezeichnet dann die so genannten a-priori-Wahrscheinlichkeiten: Tritt in n voneinander unabhängigen Versuchen, die unter den exakt gleichen Bedingungen erfolgen, m mal das Ereignis E auf, so ist die Wahrscheinlichkeit, dass bei einem Versuch das Ereignis E auftritt, definiert als

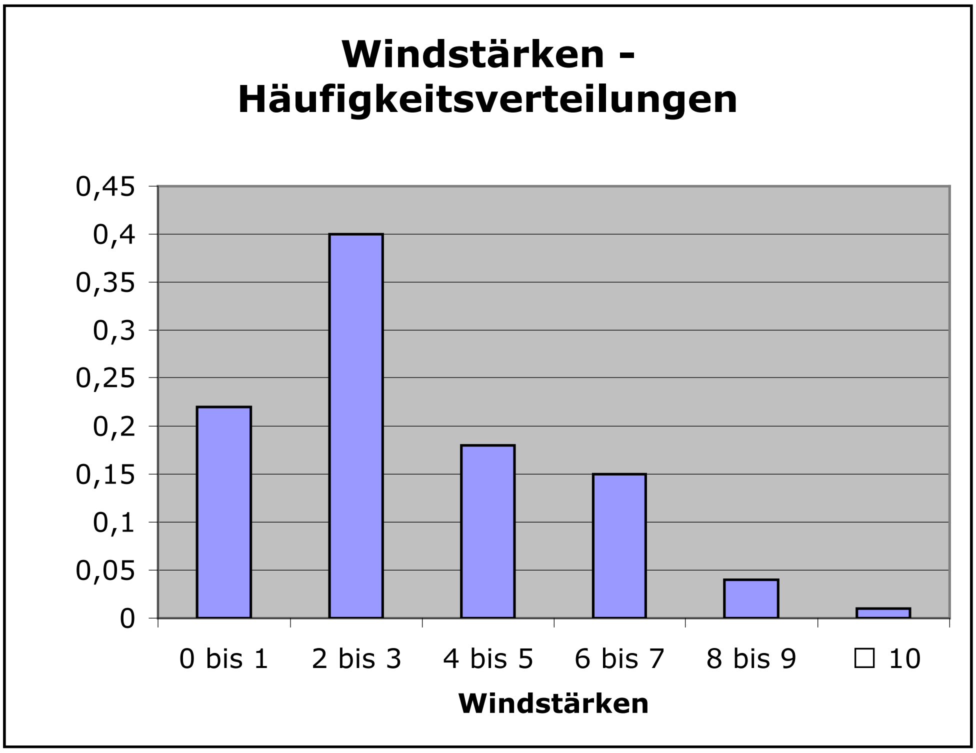
$$p(E) = \frac{m}{n}$$

Für den weiteren Verlauf der Betrachtung müssen wir uns immer gegenwärtig sein, dass in der Realität der angewandten Statistik die Bedingung, dass alle Versuche unter den exakt gleichen Bedingungen durchzuführen sind, nur in ganz seltenen Fällen realistisch ist. Mit anderen Worten: In der Praxis haben wir es mit Problemen zu tun, bei denen die Gleichwahrscheinlichkeit der einzelnen Ereignisse nicht gegeben ist, bei denen wir die Zahl der „Versuche“ – also das n in der Formel – nicht kennen oder bei der n unendlich sein kann und bei denen wir Wahrscheinlichkeiten auch für Ereignisse bestimmen wollen, die eben nicht voneinander unabhängig sind.

In einem Schaubild können wir nun alle Merkmalsausprägungen x einer beliebigen Zufallsvariablen auf der x-Achse und die Wahrscheinlichkeit, mit der diese Merkmalsausprägung auftritt, auf der y-Achse abtragen. Dabei muss gelten:

$$\sum p(x_i) = 1$$

Diese Zuordnung von Merkmalsausprägungen zu Eintrittswahrscheinlichkeiten ist die Wahrscheinlichkeitsverteilung; bei der Windstärkenverteilung in Kapitel 4.2 ergäbe sich folgendes Bild:



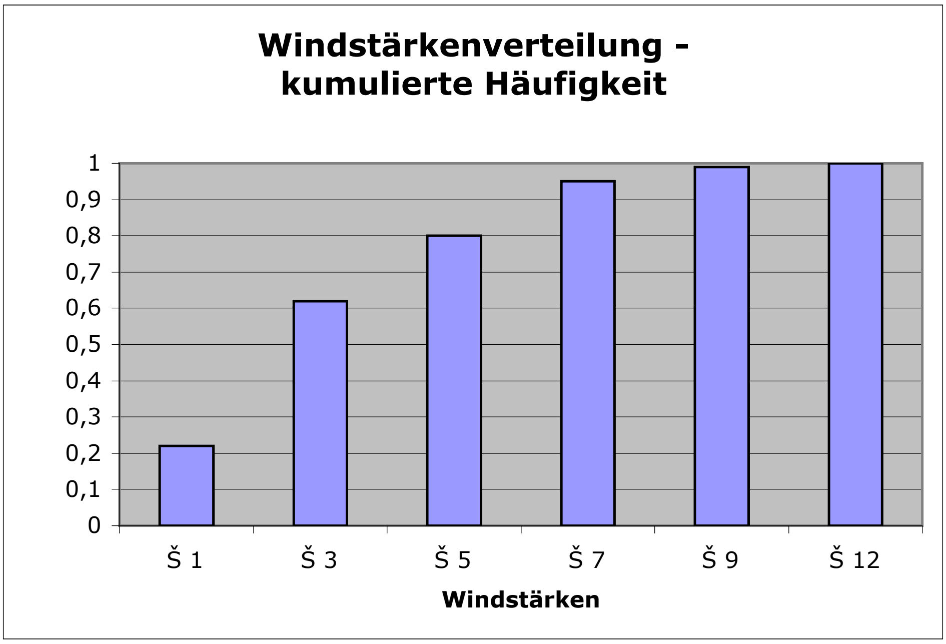
4.4.2 Wahrscheinlichkeitsfunktion

Die Wahrscheinlichkeitsfunktion beschreibt dabei die funktionale Beziehung zwischen den Ereignissen eines Experiments und den Wahrscheinlichkeiten, die diesen Ereignissen zugeordnet werden. Mit Hilfe dieser Funktion soll die Grundgesamtheit beschrieben werden, die Häufigkeitsverteilung selbst beschreibt „nur" die Stichprobe.

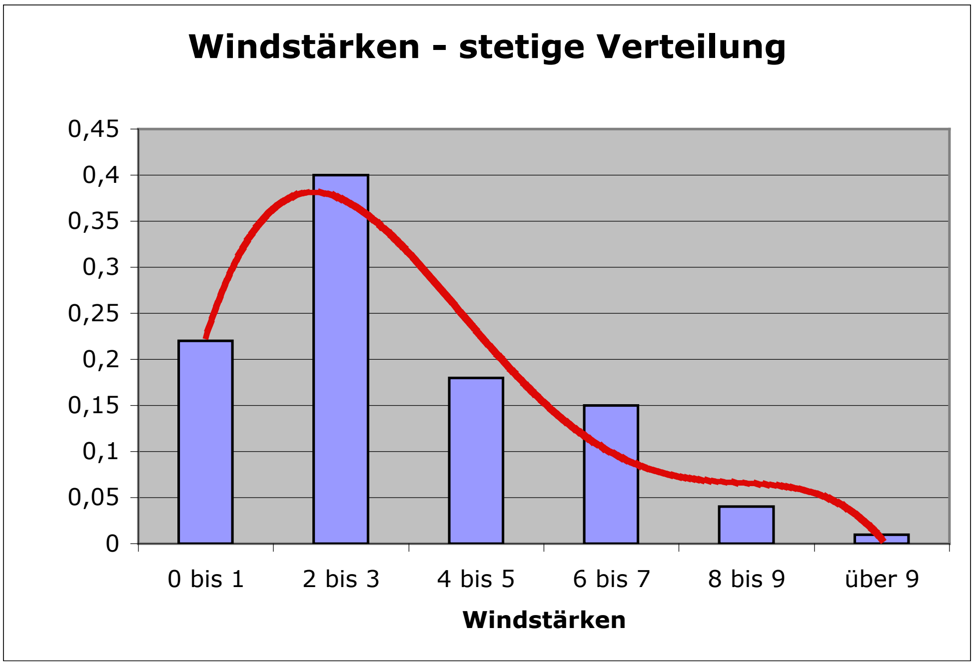
Die Funktion, die sich nun durch die Aufwärtskumulierung der Wahrscheinlichkeiten der einzelnen Merkmalsausprägungen ergibt, ist die Verteilungsfunktion, also eine Funktion der Wahrscheinlichkeitsverteilung.

$$P(X) = \sum_{i=1}^{t} p(x_i)$$

Im Falle unserer Windstärkenverteilung ergibt sich dazu folgendes Schaubild:



Bei diskreten Merkmalsausprägungen hat die Verteilungsfunktion die Gestalt einer Treppe mit unterschiedlicher Stufenhöhe. Bei einer stetigen Funktion kann die Zufallsvariable innerhalb des relevanten Bereiches des definierten Ereignisraums jeden beliebigen Wert annehmen, die Häufigkeit würde sich dann in der roten Linie zeigen:



4.4.3 Parameter von Wahrscheinlichkeitsfunktionen bei diskreten Verteilungen

4.4.3.1 Mittelwert

Wahrscheinlichkeitsverteilungen können zunächst durch ihren Mittelwert charakterisiert werden, der sich ergibt als:

$$\mu = \sum_{i=1}^{n} x_i \cdot p(x_i)$$

Der Mittelwert einer Verteilung wird auch der Erwartungswert der Zufallsvariablen x genannt.

4.4.3.2 Varianz

Die Varianz einer diskreten Wahrscheinlichkeitsverteilung[24] ergibt sich als:

$$\sigma^2 = \sum_{i=1}^{n} (x_i - \mu)^2 \cdot p(x_i)$$

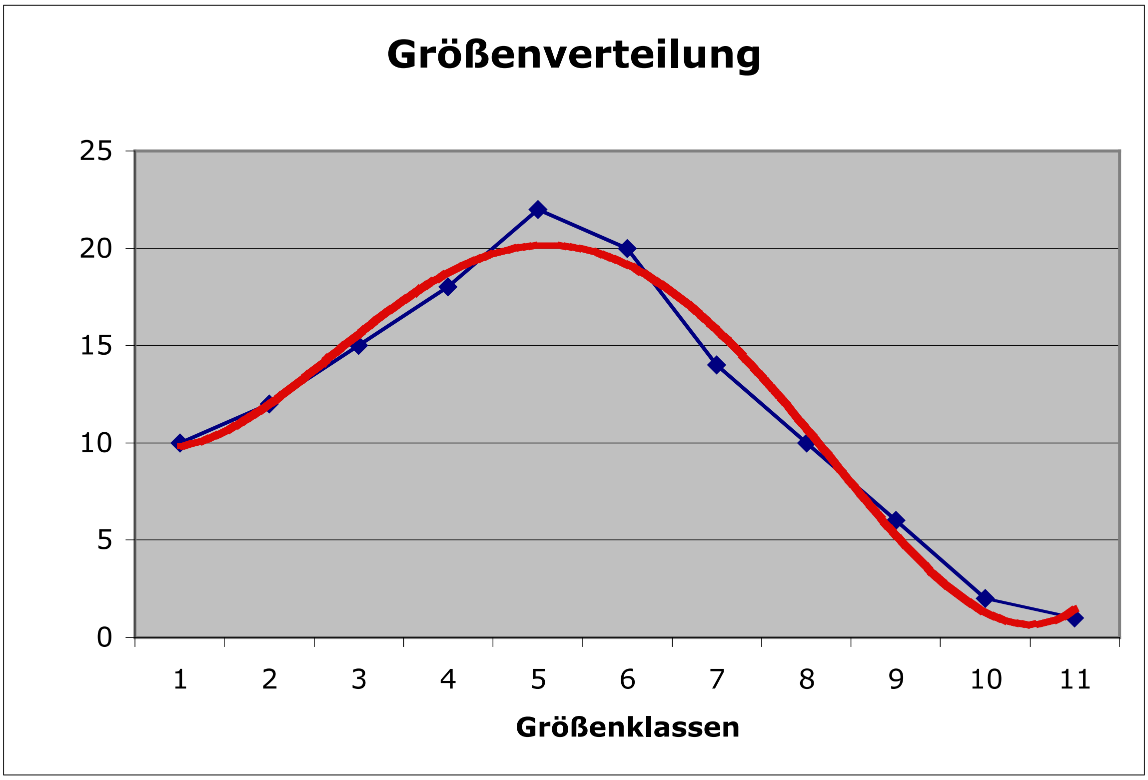
4.4.4 Parameter von Wahrscheinlichkeitsfunktionen bei stetigen Verteilungen

4.4.4.1 Dichtefunktion

Bei stetigen Funktionen ist es kaum sinnvoll, von der Eintrittswahrscheinlichkeit eines bestimmten Wertes zu sprechen, da die Wahrscheinlichkeit dann gegen Null geht oder zumindest sehr klein ist, wenn die Werte eng beieinander liegen. Ein Beispiel: Man kann die Größe von Menschen in Zentimetern messen, könnte die Messergebnisse aber auch auf Millimeter genau angeben. Misst man nun die Größe von 50 Menschen in mm, ist die Wahrscheinlichkeit sehr hoch, dass die jeweiligen „Größenklassen" immer nur mit einer Person „besetzt" sind. Sinnvoller ist es daher, die Wahrscheinlichkeit anzugeben, dass ein Wert x in einem bestimmten Intervall des Ereignisraums auftreten soll.

Dieses Intervall – in dem Beispiel etwa: alle Personen mit einer Größe zwischen 1,75 cm und 1,80 cm – bezeichnet man als Klassenbreite, die Wahrscheinlichkeitsdichte als Fläche unter der Wahrscheinlichkeitsfunktion. Reale Messergebnisse könnten in dem nachfolgenden Schaubild in der blauen Linie repräsentiert werden, mit der dazugehörigen Wahrscheinlichkeitsfunktion der roten Linie. Für bestimmte Intervalle (etwa zwischen Größenklasse 2 und 4, angedeutet durch die senkrechten schwarzen Markierungen) lässt sich dann die Wahrscheinlichkeitsdichte wie folgt angeben:

$$P\,(x_1 \le X \le x_2) = F\,(x_2) - F\,(x_1) = \int_{x_1}^{x_2} f\,(x)\,dx$$



4.4.4.2 Arithmetisches Mittel

Bei einer stetigen Verteilung ist das arithmetische Mittel definiert als

$$\mu = \int_{-\infty}^{+\infty} x \cdot f\,(x)\; dx$$

4.4.4.3 Varianz

Analog zum bereits betrachteten Fall der diskreten Wahrscheinlichkeitsverteilung ergibt sich hier:

$$\sigma^2 = \int_{-\infty}^{+\infty} (x-\mu)^2 \cdot f\,(x)\; dx$$

4.4.5 Spezielle Verteilungen

4.4.5.1 Bernoulli-Verteilung

Eine Bernoulli-Verteilung ist eine der diskreten Verteilungen; sie ist das Ergebnis eines Zufallsexperimentes, bei dem bei jedem einzelnen Versuch das Ereignis A eintreten kann oder nicht. Tritt A nicht ein, werden alle anderen potentiell möglichen Ergebnisse als „nicht-A“ zusammengefasst. Die einzelnen Versuche müssen voneinander unabhängig sein – in der Nomenklatur der im letzten Kapitel geschilderten Kombinatorik sind es also Versuche „mit Zurücklegen“. Die Eintrittswahrscheinlichkeit von A sei p(A) = π, die Eintrittswahrscheinlichkeint von „nicht-A“ ist demgemäß (1-π). Die Wahrscheinlichkeit, dass bei n Versuchen x_i mal A und (n-x_i) mal „nicht-A“ auftritt, ergibt sich als:

$$p\,(x_i) = \binom{n}{xi} \cdot \pi^{x_i} \cdot (1-\pi)^{n-x_i}$$

Der Mittelwert jedes einzelnen Versuches ist π, das arithmetische Mittel des Bernoulli-Experiments insgesamt ergibt sich damit als:

$$\mu = n \cdot \pi$$

Die Standardabweichung ergibt sich als:

$$\sigma = \sqrt{n \cdot \pi \cdot (1-\pi)}$$

Münzwürfe, Ergebnisse von Würfen mit Würfeln generell, aus dem „praktischen Leben“ auch Ergebnisse von Qualitätskontrollen bei Produktionsserien können mit Bernoulli-Verteilungen analysiert werden.

4.4.5.2 Hypergeometrische Verteilung

Bei der hypergeometrischen Verteilung wird das Bernoulli-Experiment dergestalt modifiziert, dass nun die Versuche nicht mehr voneinander unabhängig sind; das Experiment wird also – in der Sprache der Kombinatorik – „ohne Zurücklegen" ausgeführt. Es wird also untersucht, wie groß die Wahrscheinlichkeit p ist, in einer Stichprobe vom Umfang n, die ohne Zurücklegen aus einer Grundgesamtheit vom Umfang N gezogen wurde, x_i mal ein Element mit der Eigenschaft A anzutreffen, dabei kann x_i alle Werte zwischen 0 und n annehmen, und A tritt in einem Versuch mit der Wahrscheinlichkeit π auf. Diese Wahrscheinlichkeit bestimmt sich als

$$p = \frac{\binom{N \cdot \pi}{x_i} \cdot \binom{N \cdot (1-\pi)}{n - x_i}}{\binom{N}{n}}$$

Arithmetisches Mittel und Varianz ergeben sich exakt wie bei der Binomialverteilung!

4.4.5.3 Poisson-Verteilung

Die Poisson-Verteilung wird häufig in der Literatur als Grenzfall der Binomialverteilung beschrieben. Sie kommt zur Anwendung, wenn die Wahrscheinlichkeit zu berechnen ist, dass x_i Ereignisse auftreten, wenn durchschnittlich l mal dieses Ereignis auftritt und die Ereignisse voneinander unabhängig sind – etwa die Anzahl von Blitzen bei einem Gewitter in einem bestimmten Gebiet, die Anzahl von Schadensmeldungen bei einer Versicherung in einem bestimmten Zeitraum und Ähnliches. Diese Wahrscheinlichkeit ergibt sich als:

$$p\,(xi) = \frac{\lambda^{x_i}}{x_i!} \cdot e^{-\lambda}, \qquad wobei\ \lambda = n \cdot \pi$$

Der Mittelwert dieser Verteilung ist l, die Standardabweichung ergibt sich als:

$$\sigma = \sqrt{\lambda}$$

4.4.5.4 Normalverteilung

Die Normalverteilung ist die meist benutzte theoretische Wahrscheinlichkeitsverteilung stetiger Zufallsvariablen. Die Dichtefunktion der Normalverteilung ergibt sich als:

$$f\,(x) = \frac{1}{\sigma \cdot \sqrt{2\pi}} \cdot e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

dabei bezeichnet s die Standardabweichung der Normalverteilung, π die Kreiskonstante, e die Eulersche Zahl, x sind die Werte der stetigen Zufallsvariablen X und m ist das arithmetische Mittel der Normalverteilung.

Die Dichtekurve ist symmetrisch zum Lot im arithmetischen Mittel X = μ; dort liegt das einzige Maximum der Kurve. Ihre Wendepunkte hat die Kurve bei

$x_1 = \mu - \sigma$ und bei
$x_2 = \mu + \sigma$

Auf beiden Seiten nähert sich die Kurve asymptotisch der x-Achse, mit anderen Worten, die Variable X kann alle Werte zwischen -∞ und +∞ annehmen; die Fläche unter der Kurve ist gleich 1. Je größer die Standardabweichung ist, desto flacher und breiter wird die Dichtekurve. Die Kurve überdeckt 95,44% der Fläche im Bereich von ± 2 s und 99,9% der Fläche im Bereich ± 3,291s.

5 Induktive Statistik

Die Induktive Statistik findet Anwendung, wenn keine vollständigen Daten vorliegen, wenn man zum Beispiel nur über eine Teilerhebung verfügt, aus der man Rückschlüsse auf die Grundgesamtheit ziehen muss. Im Folgenden wird eine Reihe von Verfahren in Grundzügen dargestellt. Dabei greifen wir auf einige Begriffe und Verfahren aus der Wahrscheinlichkeitsrechnung zurück.

5.1 Grundlagen der Stichprobentheorie

In der Stichprobentheorie geht es um die Schätzung unbekannter Parameter. Das kann durch Punktschätzungen oder Intervallschätzungen erfolgen. In einem (Zufalls-)Auswahlverfahren werden zunächst einmal Zufallsexperimente durchgeführt, die dazu dienen, Elemente der Grundgesamtheit auszuwählen. Wenn der Zufall bei der Auswahl eine wesentliche Rolle spielt, dann spricht man auch von Stichproben.

Hierbei kann man zwischen zwei Varianten unterscheiden. Zum einen gibt es die repräsentative Stichprobe, bei der die Auswahl der Elemente nur zum Teil dem Zufall überlassen wird. Dies wird erreicht, indem ein oder mehrere Merkmale der Elemente von der Struktur her gleich oder zumindest ähnlich sind. Im Gegensatz dazu spricht man von einer reinen Zufallssauswahl, wenn jedes Element der Grundgesamtheit mit der gleichen Wahrscheinlichkeit in der Stichprobe vorkommen kann.

Das Urnenmodell entspricht der reinen Zufallsauswahl. Bei diesem Modell befindet sich eine bestimmte Anzahl von Elementen in einer Urne, die dann nacheinander gezogen werden, ohne die Elemente voneinander unterscheiden zu können. Hier wird unterschieden in Ziehen mit und Ziehen ohne Zurücklegen.

5.2 Punktschätzungen

Dabei werden unbekannte Parameter aus der Verteilung einer Zufallsvariable geschätzt, wie der Erwartungswert oder die Varianz oder aber Mittelwerte und Varianzen aus vorgegebenen Grundgesamtheiten.

Aus einer Stichprobe $x_1, x_2, \ldots, x_n$ soll ein fester, uns unbekannter Parameter q der Verteilung einer Grundgesamtheit geschätzt werden. Die Parameterschätzung U ist dabei eine Funktion der Stichprobenrealisationen:

$$U = u(x_1, x_2, \ldots, x_n)$$

Der Index des Parameters gibt die Abhängigkeit vom Stichprobenumfang an.

5.2.1 Anforderungen an Schätzfunktionen

Erwartungstreue

Definition: Eine Schätzung U eines Parameters θ heißt erwartungstreu, wenn für den Erwartungswert der Schätzung gilt E[U] = θ
Die Verzerrung oder der bias der Schätzung wird angegeben durch die Differenz
E[U] – q. Die Schätzung heißt asymptotisch erwartungstreu, wenn gilt $\lim_{n \to \infty} E[U] = \theta$

Konsistenz

Definition: Eine Schätzung U eines Parameters q heißt konsistent, wenn $\lim_{n \to \infty} p(|U - \theta| \geq \varepsilon) = 0$ für ε > 0.
Eine alternative Schreibweise wäre $\underset{n \to \infty}{p \lim} U = \theta$.

Effizienz

Definition: Innerhalb einer Klasse K von erwartungstreuen Schätzungen U_K heißt die erwartungstreue Schätzung U_1 effizient, wenn $V[U_1] \leq V[U_K]$ für alle U_K $\in$ K. Eine effiziente Schätzung zeichnet sich also durch eine minimale Varianz innerhalb der erwartungstreuen Schätzungen aus.

5.2.2 Die Schätzung des Mittelwerts

Wir gehen davon aus, dass der Mittelwert $\hat{x}$ des Merkmals X einer Grundgesamtheit unbekannt ist. Er wird mit Hilfe einer Stichprobe vom Umfang n geschätzt. Der Schätzwert $\hat{\mu}$ des Mittelwertes $\hat{x}$ lässt sich berechnen, indem man das arithmetische Mittel aus den Stichprobenwerten x_i der einzelnen Stichprobenelemente berechnet:

$$\hat{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \quad \rightarrow \qquad \hat{\mu} = \hat{x}$$

Diese Punktschätzung des unbekannten Mittelwertes macht keine Aussage darüber, ob die Schätzung richtig ist, bzw. wie nah der Schätzer dem wahren Mittelwert kommt. Anhand der oben angeführten Anforderungen an Schätzfunktionen lässt sich überprüfen, ob es sich um eine gute Schätzformel handelt.

Erwartungstreue: Eine Schätzung ist erwartungstreu, wenn der Erwartungswert des Schätzers mit dem gesuchten Parameter übereinstimmt, in unserem Fall also $E(\hat{\mu}) = \mu$.

Um das zu überprüfen, gilt es zunächst festzustellen, dass die Punktschätzung $\hat{\mu} = \hat{x}$ eine Realisation der Zufallsvariablen $\hat{X}$ darstellt, also des Mittelwertes der Zufallsvariablen. Diese sind identisch und unabhängig verteilt. Die Wahrscheinlichkeitsverteilung der X_i ist gegeben durch die Häufigkeitsverteilung der Grundgesamtheit: $E(X_i) = \mu$ und $V(X_i) = \sigma^2$.

Es gilt – (siehe Kap.**???**): $E(\hat{\mu}) = E(\hat{X}) = \frac{1}{n} E\left(\sum_{i=1}^{n} X_i\right) = \frac{1}{n}\sum_{i=1}^{n} E(X_i) = \frac{1}{n} n\mu = \mu$

Konsistenz: Hier überprüfen wir, ob die Varianz des Schätzers mit zunehmenden n gegen 0 konvergiert.

$$V(\hat{\mu}) = V(\hat{X}) = V\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n^2}\sum_{i=1}^{n} V(X_i) = \frac{1}{n^2} n\sigma^2 = \frac{\sigma^2}{n}$$

Für steigendes n geht die Varianz des Schätzers gegen 0 und der Schätzer ist somit konsistent.

Beispiel: Aus der Grundgesamtheit von Schülern eines Gymnasiums wurde eine Stichprobe vom Umfang n = 12 gezogen. Dabei wurde das Gewicht (X) in kg festgestellt. Die Daten sind in der folgenden Tabelle erfasst.

i	1	2	3	4	5	6	7	8	9	10	11	12
x_i	54	48	51	62	43	52	49	41	58	50	46	43

Die Werte der Stichprobe haben den Mittelwert 49,75 kg. Die Schätzung für den Mittelwert des Gewichts der Schüler des Gymnasiums lautet dann: $\hat{\mu} = \hat{x} = 49{,}75\,\text{kg.}$ kg.

5.2.3 Die Schätzung der Varianz

Wir gehen davon aus, dass der Erwartungswert der Grundgesamtheit mit identisch und unabhängig verteilten Zufallsvariablen m bekannt, die Varianz σ^2 aber unbekannt ist. Die Varianz kann in der Grundgesamtheit erwartungstreu geschätzt werden durch:

$\tilde{s}^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)^2$. Dies kann man zeigen, indem man den Erwartungswert von $\tilde{s}$ berechnet: $E(\tilde{s}^2) = E\left[\frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)^2\right]$.

Durch Umformung kommt man zu dem Ergebnis: $E(\tilde{s}^2) = \sigma^2$.

Schwieriger wird es, wenn m unbekannt ist. Dann verwendet man den Schätzer von m, nämlich $\hat{\mu}$, zur Berechnung der Stichprobenvarianz $s^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \hat{\mu})^2$. Die Stichprobenvarianz wird für die Schätzung der Varianz der Grundgesamtheit verwendet. Diese erweist sich allerdings für identisch und unabhängig verteilte Zufallsvariablen X_i mit i=1,…,n und $E(X_i) = \mu$, $E(X_i^2) = \mu_2$ und $V(X_i) = \sigma^2$ als nicht erwartungstreu, denn:

$E\left(s^2\right) = E\left[\frac{1}{n}\sum_{i=1}^{n}\left(X_i - \hat{X}\right)^2\right]$ ergibt $E\left(s^2\right) = \frac{n-1}{n}\cdot\sigma^2$. Die Stichprobenvarianz ist für n gegen unendlich nur asymptotisch erwartungstreu. Eine erwartungstreue Schätzung wird möglich durch eine Modifizierung der Stichprobenvarianz: $\tilde{s}^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(x_i - \hat{\mu}\right)^2$, bzw. $\tilde{s}^2 = \frac{n}{n-1}\cdot s^2$. Der Erwartungswert von $\tilde{s}^2$ ist $E\left(\tilde{s}^2\right) = \sigma^2$.

Beispiel: Wir verwenden das Zahlenbeispiel der Schüler bei der Schätzung des Mittelwertes. Zunächst berechnen wir die Varianz der Stichprobenwerte:

$s_x^2 = \frac{1}{12}(54^2 + 48^2 + \ldots + 43^2) - 49{,}75^2 = 35{,}6875;$ daraus lässt sich die erwartungstreue Punktschätzung berechnen: $\tilde{s}^2 = \frac{12}{11}\cdot 36{,}6875 = 38{,}9318$.

Anwendung: Die oben dargestellte Formel findet Anwendung beim Ziehen mit Zurücklegen. Beim Ziehen ohne Zurücklegen ändert sich der Korrekturfaktor. Die Schätzformel lautet dann: $\tilde{s}^2 = \frac{N-1}{N}\frac{n}{n-1}s^2$.

5.3 Verteilungen

Im Folgenden werden drei wichtige Verteilungen vorgestellt, die bei Intervallschätzungen und Hypothesentests von Bedeutung sind.

5.3.1 Chi-Quadrat-Verteilung

Definition: $\chi_n^2 := Y_1^2 + Y_2^2 + \ldots + Y_n^2$; Die $Y_1, Y_2, \ldots, Y_n$ seien unabhängige und standardnormalverteilte Zufallsvariablen. Das Ergebnis χ_n^2 ist ebenfalls eine Zufallsvariable und heißt chi-Quadrat-verteilt mit n Freiheitsgraden. Die Freiheitsgrade geben an, wie viele unabhängige Zufallsvariablen darin enthalten sind.

Formel:
- Erwartungswert: $E\left(\chi_n^2\right) = n$
- Varianz: $V\left(\chi_n^2\right) = 2n$

Anwendung: Die chi-Quadrat-Verteilungen sind stetig verteilt und die Wahrscheinlichkeitsdichte ist positiv im Intervall zwischen Null und unendlich. Für $\chi_n^2 \to \infty$ strebt die Dichtefunktion gegen Null.

5.3.2 Student-t-Verteilung

Definition: Die beiden Zufallsvariablen χ_n^2 und Y seien chi-Quadrat-, bzw. standardnormalverteilt und stochastisch voneinander unabhängig.

Die Zufallsvariable $T_n := \frac{Y}{\sqrt{\frac{1}{n}\cdot\chi_n^2}}$ heißt dann t-verteilt mit n Freiheitsgraden.

Formel: - Erwartungswert: $E(T_n) = 0$

- Varianz: $V(T_n) = \frac{n}{n-2} \mathrm{f}\, 1$

Anwendung: Die Student-t-Verteilung ist stetig und der Wertebereich liegt zwischen $-\infty$ und $+\infty$.

5.3.3 F-Verteilung

Definition: Eine Zufallsvariable, die aus dem Quotient zweier unabhängig und chi-Quadrat-verteilter Zufallsvariablen X und Y mit den Freiheitsgraden m beziehungsweise n besteht, heißt F-verteilt, wenn die jeweilige Zufallsvariable mit dem Kehrwert ihres Freiheitsgrades multipliziert wird.

Formel: $F_n^m := \dfrac{\frac{1}{m} X_m}{\frac{1}{n} Y_n}$

Anwendung: F-Verteilungen sind stetig und haben positive Wahrscheinlichkeitsdichten im Intervall zwischen 0 und ∞. Ihre Dichtefunktion ist linkssteil und strebt für $F \to \infty$ gegen Null. Der Erwartungswert beträgt: $E(F_n^m) = \frac{n}{n-2}$ für n > 2 und die Varianz: $V(F_n^m) = \frac{2n^2(m+n-2)}{m(n-2)^2(n-4)}$ für n > 4. Weiterhin gilt $F_n^m = \frac{1}{F_m^n}$.

5.4 Intervallschätzungen

Definition: Ein Intervall, dessen Grenzen aus Stichprobendaten berechnet werden kann und das den unbekannten, zu schätzenden Parameter θ mit einer vorgegebenen Wahrscheinlichkeit 1-α überdeckt, heißt Konfidenzintervall für θ. 1-α ist der Sicherheitsgrad.

Formel: $W(U_1 \leq \theta \leq U_2) = 1 - \alpha$ mit $U_1 = u_1(X_1, X_2, \ldots, X_n)$ und $U_2 = u_2(X_1, X_2, \ldots, X_n)$

Es gilt zu beachten, dass die Intervallgrenzen U_1 und U_2 vor der Stichprobennahme Zufallsvariablen sind. Die Wahrscheinlichkeit, ein Intervall zu bekommen, in dem der gesuchte Parameter θ liegt, ist 1-α. Nach der Stichprobennahme sind die Intervallgrenzen Realisationen von Zufallsvariablen, also reelle Zahlen. Der gesuchte Parameter θ liegt entweder im Intervall oder nicht. Dann ist es nicht mehr sinnvoll, von Wahrscheinlichkeiten zu sprechen.

5.4.1 Konfidenzintervall für Mittelwerte

Formel: $\left(\hat{x} - z_{\frac{\alpha}{2}} \cdot \sigma_{\hat{x}} \leq \mu \leq \hat{x} + z_{\frac{\alpha}{2}} \cdot \sigma_{\hat{x}} \right) = 1 - \alpha$

Anwendung: Die Konfidenzwahrscheinlichkeit $1-\alpha$ gibt an, wie sehr man darauf vertraut, dass der Wert m, der fest aber unbekannt ist, innerhalb des Intervalls $\left[\hat{x} - z_{\frac{\alpha}{2}} \cdot \sigma_{\hat{x}} ; \hat{x} + z_{\frac{\alpha}{2}} \cdot \sigma_{\hat{x}} \right]$ liegt. $\sigma_{\hat{x}} = \frac{\sigma}{\sqrt{n}}$ ist die Varianz des Mittelwertes, der Wert für z ergibt sich aus der Tabelle für die Normalverteilung. Der Wert für z ist der Wert, der in die Dichtefunktion der Wahrscheinlichkeitsverteilung eingesetzt wird. Die Fläche unterhalb der Dichtefunktion in den Grenzen von z ist 1 – a. a ist die Irrtumswahrscheinlichkeit, die angibt, wie oft man sich durchschnittlich beim Aufstellen der Intervalle irrt. Je kleiner a gewählt wird, desto verlässlicher ist die Schätzung.

Oft ist die Varianz – neben μ – ebenfalls unbekannt, so dass eine Schätzung der Varianz verwendet werden muss. Diese ergibt sich aus den im vorherigen Abschnitt behandelten Punktschätzungen. Das neue Konfidenzintervall lautet dann:

Formel: $\left(\hat{x} - z_{\frac{\alpha}{2}} \cdot \tilde{s}_{\hat{x}} \leq \mu \leq \hat{x} + z_{\frac{\alpha}{2}} \cdot \tilde{s}_{\hat{x}} \right) = 1 - \alpha$

Da durch die Schätzung der Varianz eine zusätzliche Ungenauigkeit entsteht, sollten nur große Stichprobenumfänge verwendet werden. In der Regel ist ein Stichprobenumfang ab n = 30 als ausreichend anzusehen.

Bei kleinen Stichprobenumfängen wird mit der Student-t-Verteilung gearbeitet. Die Zufallsvariable $\frac{\hat{x}-\mu}{\tilde{s}_{\hat{x}}}$ ist Student-t-verteilt mit n-1 Freiheitsgraden, wenn die Stichprobe aus einer (μ, σ)-Normalverteilung ist und $\hat{x}=\frac{1}{n}\sum_{i=1}^{n}x_i$; $\tilde{s}^2=\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\hat{x})^2$.

Das Konfidenzintervall für μ lautet dann:

Formel: $\left(\hat{x}-t_{\frac{\alpha}{2};(n-1)}\cdot\tilde{s}_{\hat{x}}\le\mu\le\hat{x}+t_{\frac{\alpha}{2};(n-1)}\cdot\tilde{s}_{\hat{x}}\right)$

Anwendung: Allgemein lässt sich zu Konfidenzintervallen sagen, dass das Konfidenzintervall umso breiter ist, je größer σ ist. Je größer der Umfang n, umso kleiner ist das Konfidenzintervall. Eine Vervierfachung des Stichprobenumfangs würde die Länge des Konfidenzintervalls halbieren. Schließlich ist das Konfidenzintervall umso breiter, je größer der Sicherheitsgrad 1-α ist.

Beispiel: Eine Befragung von 50 Arbeitnehmern zur Ermittlung der Lohnsituation hat einen Stichprobenmittelwert μ = 12,50 Euro pro Stunde und eine Standardabweichung von s = 2,56 ergeben. Wir berechnen zunächst die Verteilung in der Grundgesamtheit. In welchem Intervall liegen 50% der Stundenlöhne? Wir gehen von der Normalverteilung aus und die Punktschätzung des tatsächlichen Stundenlohnes ergibt sich als

$\hat{\mu}$ = 12,5. Die Punktschätzung der Standardabweichung beträgt $\tilde{s}=\sqrt{\frac{50}{49}}\cdot 2{,}56\approx 2{,}59$. Damit ist die $z=\frac{X-12{,}5}{2{,}59}$ Variable standardnormalverteilt. Für das Intervall für 50% der Löhne kann man die Werte aus der Tabelle der Standardnormalverteilung ablesen:

-0,675 $\le z \le$ 0,675. Einsetzen von z ergibt für die Grenzen des gesuchten Intervalls: $12{,}5-0{,}675\cdot 2{,}59\le X\le 12{,}5+0{,}675\cdot 2{,}59$ = 12,5 – 1,748 ≤ X ≤ 12,5 + 1,748.

Jetzt interessieren wir uns für die Größe des Konfidenzintervalls bei einer vorgegebenen Konfidenzwahrscheinlichkeit von 0,9 für den durchschnittlichen Lohn pro Stunde. Dazu benötigen wir noch die geschätzte Standardabweichung des Mittelwertes. $\tilde{s}_{\hat{x}}=\frac{\tilde{s}}{\sqrt{n}}=\frac{2{,}59}{\sqrt{50}}\approx 0{,}37$. Als z bei einer Wahrscheinlichkeit von 1-α= 0,9 kann man aus der Tabelle ablesen, dass z = 1,645. Das Konfidenzintervall hat damit die Grenzen $[12{,}5-1{,}645\cdot 0{,}37\le\mu\le 12{,}5+1{,}645\cdot 0{,}37]=0{,}9$ oder [11,9 €; 13,1 €].

5.4.2 Konfidenzintervall für Varianzen

Wir gehen von dem Fall aus, dass das Merkmal in der Grundgesamtheit annähernd normalverteilt ist mit dem Mittelwert μ und der Stμandardabweichung σ. Die Einzelstichproben werden mit Zurücklegen gezogen, sind also unabhängig.

Die Varianz s^2 der Zufallsstichprobe ist eine Zufallsstichprobe. Durch Umformung erhält man eine chi-Quadrat-verteilte Zufallsvariable mit n-1 Freiheitsgraden.

$$\textit{Formel}:\ n \cdot \frac{s^2}{\sigma^2} = \chi^2_{n-1}$$

Das Konfidenzintervall für die Varianz lautet dann:

$$KONF\left(\frac{n \cdot s^2}{\chi^2_{n-1}[1-\alpha/2]} \le \sigma^2 \le \frac{n \cdot s^2}{\chi^2_{n-1}[\alpha/2]} \right)$$

Die Werte für χ^2 sind unterschiedlich, beide müssen in der Tabelle der chi-Qudrat-Verteilung nachgeschlagen werden.

Beispiel: Eine Stichprobe einer normalverteilten Grundgesamtheit vom Umfang n = 25 liefert die Varianz $s^2 = 150$. Die Punktschätzung für die Grundgesamtheit ergibt $\tilde{s}^2 = \frac{25 \cdot 150}{24} = 156{,}25$ und für $\tilde{s} = 12{,}5$. Nun stellen wir das Konfidenzintervall mit einem Sicherheitsgrad von 95 % auf. Die Grenzen des Intervalls liegen bei $\chi^2_{24}[0{,}025] = 12{,}401$ und $\chi^2_{24}[0{,}975] = 39{,}364$.

Das Konfidenzintervall lautet dann:

$$KONF(\left(\frac{25 \cdot 150}{39{,}364}\right) \le \sigma^2 \le \left(\frac{25 \cdot 150}{12{,}401}\right)) = 0{,}95 \text{ bzw. } KONF(95{,}26 \le \sigma^2 \le 302{,}39).$$

5.5 Grundlagen des Testens von Hypothesen

Wenn Annahmen über die Verteilung einer Gesamtheit oder einer Zufallsvariablen getroffen werden, dann nennt man diese Annahmen statistische Hypothesen. Diese statistischen Hypothesen lassen sich mit ja oder nein beantworten, es gibt also nur zwei mögliche Ergebnisse. Allerdings wird nicht festgestellt, ob die Hypothesen richtig oder falsch sind. Es geht darum, die Hypothesen beizubehalten oder zu verwerfen. Dafür werden die statistischen Hypothesen auf ihre Verträglichkeit mit den Stichprobendaten überprüft. Die dazu nötigen Vorschriften werden statistische Tests genannt. Die zu überprüfende Hypothese nennt man auch Prüf- oder Null-Hypothese, ihre Bezeichnung ist in der Regel H_0.

Folgende Schritte sind für die Durchführung eines Tests der Hypothese notwendig:

- Definition der Null-Hypothese: Ziel ist es, diese zu verwerfen, also abzulehnen, so dass die Alternativhypothese nachgewiesen ist.
- Festlegung des Signifikanzniveaus α. Dieses gibt die Irrtumswahrscheinlichkeit an, in der Regel werden kleine Zahlen gewählt (0,01 oder 0,05).
- Festlegung einer Stichprobenfunktion $U = u(x_1, x_2, \ldots, x_n)$. Zur Durchführung des Tests muss die Verteilung von U bekannt sein, unter der Annahme, dass H_0 richtig ist. U ist die Prüfgröße des Tests. Berechnung des Ablehnbereichs A des Tests. A ist eine Teilmenge der reellen Zahlen, die von U und von α abhängt. Entscheidungsregel: Die Null-Hypothese wird verworfen (abgelehnt), wenn der aus der Stichprobe berechnete Wert der Prüfgröße U in den Ablehn-Bereich A fällt. Dann gilt es als erwiesen, dass H_0 falsch ist. Das Gegenteil von H_0, H_1, ist dann richtig. Fällt U nicht in den Ablehn-Bereich A, wird H_0 nicht abgelehnt. Das bedeutet allerdings nicht, dass die Null-Hypothese dann richtig ist.

Bei solchen Tests sind Fehler möglich. Zum einen kann es passieren, dass die Null-Hypothese abgelehnt wird, obwohl sie richtig ist (Fehler 1. Art). Zum anderen kann die Null-Hypothese nicht abgelehnt werden, obwohl sie falsch ist (Fehler 2. Art).

Die Wahrscheinlichkeit für einen Fehler erster Art sei α und die Wahrscheinlichkeit für einen Fehler zweiter Art sei β. Da es das Ziel ist, die Null-Hypothese zu verwerfen, wird großen Wert darauf gelegt, dass der Fehler erster Art mit geringer Wahrscheinlichkeit auftritt, α also möglichst klein ist. Üblicherweise mit α mit 1% oder 5% angenommen.

5.5.1 Test einer Hypothese über einen Mittelwert

Wir formulieren die Nullhypothese mit $H_0 : \mu = \mu_0$. Mittels einer Stichprobe soll überprüft werden, ob die Null-Hypothese abgelehnt wird. Dazu wird die Abweichung des Stichprobenmittelwertes $\hat{x}$ vom angenommenen Mittelwert μ_0 errechnet. Wenn $\left|\hat{x} - \mu_0\right|$f 0 signifikant von Null verschieden ist, dann wird die Null-Hypothese verworfen. Die signifikante Abweichung wird unter Verwendung der Verteilung des Stichprobenmittelwertes $\hat{X}$ berechnet. Die Wahrscheinlichkeit, dass $\hat{X}$ in den Ablehnbereich fällt, obwohl H_0 richtig ist, liegt bei α. Bei einer großen, annähernd normalverteilten Stichprobe gilt in standardisierter Form:

$$Formel:\ P\left(\frac{\left|\hat{X} - \mu_0\right|}{\sigma_{\hat{X}}} > z\left[1 - \alpha / 2\right] H_0 richtig\right) = \alpha\,.$$

α wird auch als Signifikanzniveau des Tests bezeichnet. z ist der kritische Wert, bei dessen Überschreiten die Null-Hypothese verworfen wird. Ist die Standardabweichung nicht bekannt, dann muss sie geschätzt und für $\sigma_{\hat{X}}$ eingesetzt werden.

Hypothesentests werden noch unterschieden in zweiseitige, rechtsseitige und linksseitige Tests. Zweiseitige Tests sind solche, bei denen die Nullhypothese eine Gleichheitsbeziehung darstellt, die Alternativ-Hypothese dementsprechend eine Ungleichheitsbeziehung.

Zweiseitige Hypothesentests:
$$\begin{aligned} H_0 &: \mu = \mu_0 \\ H_1 &: \mu \neq \mu_0 \end{aligned}$$

Die Entscheidungsregel lautet: $$\frac{\left|\hat{X} - \mu_0\right|}{\sigma_{\hat{X}}} > z\left[1 - \alpha / 2\right] \Rightarrow H_0 ablehnen\,.$$

Der Ablehnungsbereich liegt in diesem Fall auf beiden Seiten der dazugehörigen Dichtefunktion, der Flächeninhalt des Ablehnungsbereiches beträgt auf jeder Seite α/2.

Rechtsseitige Tests:
$$\begin{aligned} H_0 &: \mu \leq \mu_0 \\ H_1 &: \mu > \mu_0 \end{aligned}$$

Die Entscheidungsregel lautet: $$\frac{\left|\hat{X} - \mu_0\right|}{\sigma_{\hat{X}}} > z\left[1 - \alpha\right] \Rightarrow H_0 ablehnen\,\cdot$$

Der Ablehnungsbereich liegt am rechten Rand der Dichtefunktion, der Flächeninhalt beträgt α.

Linksseitige Tests:
$$\begin{aligned} H_0 &: \mu \geq \mu_0 \\ H_1 &: \mu < \mu_0 \end{aligned};$$

Die Entscheidungsregel lautet: $$\frac{\left|\hat{X} - \mu_0\right|}{\sigma_{\hat{X}}} < z\left[\alpha\right] \Rightarrow H_0 ablehnen\,.$$

Der Ablehnungsbereich liegt am linken Rand der Dichtefunktion, der Flächeninhalt beträgt α.

Falls der Stichprobenumfang klein ist, findet wieder die Student-t-Verteilung Anwendung.

Die Entscheidungsregel lautet dann bei zweiseitiger Fragestellung:

$$\frac{|\hat{x}-\mu_0|}{\tilde{s}_{\hat{x}}} > t_{n-1}[1-\alpha/2] \Rightarrow H_0 ablehen.$$

Rechts- und linksseitige Tests sind analog zu oben.

Beispiel: Das Einkommen von 1.000 Angestellten eines Betriebs steigt im Laufe der Betriebszugehörigkeit erfahrungsgemäß um 30% an. Eine Stichprobe von 15 Personen ergab folgende Steigerungen in Prozent: 32, 29, 30, 31, 34, 28, 29, 33, 27, 32, 31, 29, 33, 30, 29.
Der Mittelwert der Stichprobe lautet $\hat{x}$ = 30,47, die Standardabweichung s = 1,96.
Die Nullhypothese lautet: H_0 = μ = 30%; H_1 = μ ≠ 30%.

Die Schätzung der Standardabweichung des Mittelwerts ergibt (Achtung: Ziehen ohne Zurücklegen!): $\tilde{s}_{\hat{x}} = \sqrt{\frac{1 \cdot 15}{15 \cdot 14}} \cdot 1{,}96 = 0{,}52$. Daraus ergibt sich eine Prüfgröße von $\frac{|\hat{x}-\mu_0|}{\tilde{s}_{\hat{x}}} = \frac{|30{,}47-30|}{0{,}52} = 0{,}904$. 0,904. Da es sich um eine kleine Stichprobe handelt, verwenden wir den t-Test und der kritische Wert bei einem Sicherheitsgrad von 95 % und 14 Freiheitsgraden beträgt $|t_{14}\ (0{,}025)|$ = t_{14} (0.975) = 2,145. Da die Prüfgröße kleiner ist als der kritische Wert, wird die Nullhypothese nicht abgelehnt.

5.5.2 Test einer Hypothese über einer Varianz

Die Nullhypothese lautet: H_0: $\sigma^2 = \sigma_0^2$

Die Prüfgröße ist von den Konfidenzintervallen für Varianzen schon bekannt: $n \cdot \frac{s^2}{\sigma^2} = \chi^2_{n-1}$. Die Entscheidungsregel lautet dann:

$$n \cdot \frac{s^2}{\sigma^2} < \chi^2_{n-1}[\alpha/2] \text{ oder } n \cdot \frac{s^2}{\sigma^2} > \chi^2_{n-1}[1-\alpha/2] \Rightarrow H_0 ablehnen.$$

Beispiel: Die Varianz einer Stichprobe vom Umfang n = 30 ist s^2 = 0,5. Die Varianz der Grundgesamtheit liegt bei σ^2 = 0,45 und darf nicht überschritten werden. Die Nullhypothese lautet also: H_0 : $\sigma^2 \le 0{,}45$; H_1 : $\sigma^2 > 0{,}45$. Die Prüfgröße lässt sich berechnen mit $n \cdot \frac{s^2}{\sigma^2}$ = 30 × 0,5 / 0,45 = 33,34. Die kritischen Werte bei einem Sicherheitsgrad von 95% und 39 Freiheitsgraden lauten: $\chi^2_{29}[0{,}025] = 16{,}047$ und $\chi^2_{29}[0{,}975] = 45{,}722$. Die Prüfgröße befindet sich innerhalb der kritischen Werte und wird somit nicht abgelehnt.

5.6 Multivariate Regression und Korrelation

Die einfache lineare Regression[25] reicht oft nicht aus, da mehrere Größen Einfluss auf eine Variable nehmen. Deshalb wurden Modelle entwickelt, die den Einfluss mehrerer Größen berücksichtigen.

Formel: $Y = f(X_1, X_2, \ldots, X_n) + U$

oder in linearer Form: $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_n X_n + U$

Matrixschreibweise:

Zur Vereinfachung der Schreibweise wird in der multivariaten Regressionsrechnung häufig die Matrixschreibweise verwendet. Matrizen werden im Folgenden mit fetten Buchstaben geschrieben. Kleine Buchstaben stellen einspaltige Matrizen dar, große Buchstaben mehrspaltige. Die Matrizen für das Modell sehen folgendermaßen aus:

$$\mathbf{y}=\begin{pmatrix} y_1 \\ y_2 \\ y_3 \\ \vdots \\ y_n \end{pmatrix};\ \mathbf{X}=\begin{pmatrix} 1 & x_{11} & x_{21} & \ldots & x_{j1} \\ 1 & x_{12} & x_{22} & \ldots & x_{j2} \\ 1 & x_{13} & x_{23} & \ldots & x_{j3} \\ \vdots & & & & \\ 1 & x_{1n} & x_{2n} & \ldots & x_{jn} \end{pmatrix};\ \boldsymbol{\beta}=\begin{pmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \\ \vdots \\ \beta_j \end{pmatrix};\ \mathbf{u}=\begin{pmatrix} u_1 \\ u_2 \\ u_3 \\ \vdots \\ u_n \end{pmatrix}$$

Damit schreiben wir dann das Modell als

Formel: $$\begin{pmatrix} y_1 \\ y_2 \\ y_3 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} 1 & x_{11} & x_{21} & \dots & x_{j1} \\ 1 & x_{12} & x_{22} & \dots & x_{j2} \\ 1 & x_{13} & x_{23} & \dots & x_{j3} \\ \vdots & & & & \\ 1 & x_{1n} & x_{2n} & \dots & x_{jn} \end{pmatrix} \cdot \begin{pmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \\ \vdots \\ \beta_j \end{pmatrix} + \begin{pmatrix} u_1 \\ u_2 \\ u_3 \\ \vdots \\ u_n \end{pmatrix}$$ oder kürzer als $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{u}$

Für die einzelnen Teile der Formel werden bestimmte Annahmen getroffen. Für **u** gilt:

$E(\mathbf{u}) = \mathbf{0}$ und $E(\mathbf{uu}') = \sigma^2 \mathbf{I} = \begin{pmatrix} \sigma^2 & 0 & \cdots & 0 \\ 0 & \sigma^2 & & 0 \\ \vdots & & \ddots & \\ 0 & 0 & \cdots & \sigma^2 \end{pmatrix}$. Dabei wird Homoskedastizität[26] der latenten Variablen und Abwesenheit von Autokorrelation unterstellt. Weiterhin sind die Störvariablen unabhängig und normalverteilt.

Die Matrix **X** ist so beschaffen, dass **X**′**X** regulär ist, das heißt, dass ihre Inverse existiert. Voraussetzung dafür ist Folgendes:

- n > j, also die Anzahl der Beobachtungstupel ist größer als die Anzahl der zu schätzenden Parameter
- die Spaltenvektoren von **X** sind linear unabhängig

Die Modellparameter werden auf folgende Art geschätzt: Die endogene Variable y lässt sich schreiben als die Summe der theoretischen Werte und der latenten Variablen oder als Summe der Schätzwerte und der Residuen.

Formel: $$X\beta + u = X\hat{\beta} + e = y$$
$$\tilde{y} + u = \hat{y} + e = y$$

Die Berechnung von $\hat{\beta}$ wird wie bei der linearen Regression nach der Methode der kleinsten Quadrate durchgeführt. Die Quadrate der Abweichungen, also der Residuen e, werden minimiert. Die Berechnung geht über den Rahmen des Buches hinaus, aber das Ergebnis für den Regressionskoeffizienten lautet:

$$\hat{\beta} = (X'X)^{-1} X'y$$

Damit lassen sich die anderen gesuchten Größen bestimmen: Der Schätzwert der Endogenen ist $\hat{y} = X\hat{\beta}$ und für die Residuen ergibt sich $e = y - X\hat{\beta}$.

Beispiel: Wir erweitern das Beispiel aus Kap. 3.4.1. Dort wurden die Konsumausgaben in Abhängigkeit des Einkommens berechnet. Jetzt fügen wir das Vermögen als weitere erklärende Variable hinzu. Die Konsumausgaben einer Person ergeben sich dann in der linearen Schreibweise als Konsumausgaben = $\beta_0 + \beta_1 \times$ Einkommen + $\beta_2 \times$ Vermögen + U.

Anwendung: Die multivariate Regressionsrechnung hat mittlerweile eine große Verbreitung gefunden. Die Berechnung wird in der Regel mit entsprechenden Computerprogrammen (z.B. EViews) durchgeführt. Aus der multivariaten Regressionsrechnung heraus hat sich ein eigenständiger Forschungsbereich entwickelt, die Ökonometrie.

6 Anhang

6.1 Abkürzungsverzeichnis

X_{med}	=	Median
X_{mod}	=	Modus
$\hat{x}$	=	arithmetisches Mittel
X_g	=	geometrisches Mittel
X_h	=	harmonisches Mittel
D	=	Ginikoeffizient
d	=	durchschittliche Abweichung
σ^2	=	Varianz
σ	=	Standardabweichung
H	=	Hirschmann-Idex
R	=	Rosenbluth-Konzentrationsmaß
sk	=	Schiefemaß nach Pearson
β_1	=	Schiefekoeffizient nach Pearson
χ_1	=	Schiefekoeffizient nach Fisher
β_2	=	Wölbungskoeffizient nach Pearson
γ_2	=	Wölbungskoeffizient nach Fisher
cov (x,y)	=	Kovarianz von x und y
r_{xy}	=	Korrelationskoeffizient nach Bravais-Pearson
ρ	=	Spearman'scher Rangkorrelationskoeffizient
K	=	Kontingenzkoeffizient
P_L	=	Preisindex nach Laspeyres
P_P	=	Preisindex nach Paasche
M_L	=	Mengenindex nach Laspeyres
M_P	=	Mengenindex nach Paasche
W	=	Wertindex
Tr	=	Trend
G	=	glatte Komponente
S	=	Saisonkomponente
U	=	Restschwankungen
p	=	Wahrscheinlichkeit
μ	=	Mittelwert einer Stichprobe
s^2	=	Varianz der Stichprobe
s	=	Standardabweichung der Stichprobe
$\hat{\mu}$	=	Schätzung des Mittelwerts
$\tilde{s}^2$	=	Schätzung der Varianz
H_0	=	Nullhypothese
H_1	=	Alternativhypothese

6.2 Fachwörter englisch – deutsch

Alternative hypothesis	Alternativhypothese
Analysis of variance (ANOVA)	Varianzanalyse
Arithmetic mean	Aithmetisches Mittel
Biased estimator	Verzerrter Schätzer
Binomial coefficient	Binomialkoeffizient
Binomial distribution	Binomialverteilung
Central limit theorem	Zentraler Grenzwertsatz
Chi-square distribution	Chi-Quadrat-Verteilung
Chi-square test	Chi-Quadrat-Test
Conditional probability	Bedingte Wahrscheinlichkeit
Confidence interval	Konfidenzintervall
Contingency table	Kontingenztabelle
Correlation	Korrelation
Correlation coefficient	Korrelationskoeffizient
Covariance	Kovarianz
Critical value	Kritischer Wert
Cumulative probability	Kumulierte Wahrscheinlichkeit
Degrees of freedom	Freiheitsgrade
Density	Dichte
Density function	Dichtefunktion
Disjoint	Disjunkt
Discrete (distribution)	Diskrete (Verteilung)
Distribution	Verteilung
Distribution function	Verteilungsfunktion
Empirical (variance)	empirisch(e) (Varianz)
Error (of type I or II)	Fehler erster oder zweiter Art
Estimate	Schätzer
Estimation	Schätzung
Event	Ereignis
Expectation value	Erwartungswert
Expected number/frequency	erwartete Häufigkeit
Exponential distribution	Exponentialverteilung
F-distribution	F-Verteilung
F-test	F-Test
Freedom, degrees of	Freiheitsgrade
Geometric distribution	Geometrische Verteilung
Goodness of fit	Güte der Schätzung
Hypothesis	Hypothese

Independent (events, stoch. var.s)	Unabhängig(e) (Ereignisse, Zufallsvariablen)
Intersection	Schnittmenge
Law of large numbers	Gesetz der großen Zahl
Least square method	Methode der kleinsten Quadrate
Level of significance	Signifikanzniveau
Likelihood (function)	Likelihood-Funktion
Linear regression	Lineare Regression
Map	Abbildung
Marginal (density, distribution)	marginal(e) (Dichte, Funktion)
Maximum likelihood estimator	Maximum-likelihood Schätzer
Mean	Mittelwert
Mean square	quadratischer Mittelwert
Median	Median
ML-estimator	ML-Schätzer
Moments	Momente
Multiple regression	Multible Regression
Non-parametric test	Nicht-parametrischer Test
Normal distribution	Normalverteilung

Normed normal distribution	normierte Normalverteilung
Normed sum	normierte Summe
Null hypothesis	Null-Hypothese
One-sided test	einseitiger Test
Outlier	Ausreißer
Point estimation	Punktschätzung
Poisson distribution	Poisson-Verteilung
Pooled variance	zusammengefasste Varianz
Probability	Wahrscheinlichkeit
Probability function	Wahrscheinlichkeitsfunktion
Quartile	Quartil
Random variable	Zufallsvariable
Rank	Rang
Rank sum	Rangsumme
Reject	Verwerfen
Sample	Stichprobe
Sample correlation coefficient	Stichprobenkorrelationskoeffizient
Sample mean	Stichprobenmittelwert
Sample size	Stichprobengröße
Sample variance	Stichprobenvarianz
Sampling distribution	Zufallsverteilung
Set	Menge
Significance level	Signifikanzniveau
Slope	Steigung
Standard deviation	Standardabweichung
Statistic	Statistik
Stochastic variable	Zufallsvariable
Student's t	Student-t
Test	Test
Two-sided test	zweiseitiger Test
Type I or II error	Fehler erster oder zweiter Art
Unbiased estimator	Unverzerrter Schätzer
Uniform distribution	Normalverteilung
Union	Vereinigung
Variance	Varianz

6.3 Statistische Tafeln

Standardnormalverteilung

z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
0.0	.5000	.5040	.5080	.5120	.5160	.5199	.5239	.5279	.5319	.5359
0.1	.5398	.5438	.5478	.5517	.5557	.5596	.5636	.5675	.5714	.5753
0.2	.5793	.5832	.5871	.5910	.5948	.5987	.6026	.6064	.6103	.6141
0.3	.6179	.6217	.6255	.6293	.6331	.6368	.6406	.6443	.6480	.6517
0.4	.6554	.6591	.6628	.6664	.6700	.6736	.6772	.6808	.6844	.6879
0.5	.6915	.6950	.6985	.7019	.7054	.7088	.7123	.7157	.7190	.7224
0.6	.7257	.7291	.7324	.7357	.7389	.7422	.7454	.7486	.7517	.7549
0.7	.7580	.7611	.7642	.7673	.7704	.7734	.7764	.7794	.7823	.7852
0.8	.7881	.7910	.7939	.7967	.7995	.8023	.8051	.8079	.8106	.8133
0.9	.8158	.8186	.8212	.8238	.8264	.8289	.8315	.8340	.8365	.8398
1.0	.8413	.8438	.8461	.8485	.8508	.8531	.8554	.8577	.8599	.8621
1.1	.8643	.8665	.8686	.8708	.8729	.8749	.8770	.8790	.8810	.8830
1.2	.8849	.8869	.8888	.8907	.8925	.8944	.8962	.8980	.8997	.9015
1.3	.9032	.9049	.9066	.9082	.9099	.9115	.9131	.9147	.9162	.9177
1.4	.9192	.9207	.9222	.9236	.9251	.9265	.9279	.9292	.9306	.9319
1.5	.9332	.9345	.9357	.9370	.9382	.9394	.9406	.9418	.9429	.9441
1.6	.9452	.9463	.9474	.9484	.9495	.9505	.9515	.9525	.9535	.9545
1.7	.9554	.9564	.9573	.9582	.9591	.9599	.9608	.9616	.9625	.9633
1.8	.9641	.9649	.9656	.9664	.9671	.9678	.9686	.9693	.9699	.9706
1.9	.9713	.9719	.9726	.9732	.9738	.9744	.9750	.9756	.9761	.9767
2.0	.9772	.9778	.9783	.9788	.9793	.9798	.9803	.9808	.9812	.9817
2.1	.9821	.9826	.9830	.9834	.9838	.9842	.9846	.9850	.9854	.9857
2.2	.9861	.9864	.9868	.9871	.9875	.9878	.9881	.9884	.9887	.9890
2.3	.9893	.9896	.9898	.9901	.9904	.9906	.9909	.9911	.9913	.9916
2.4	.9918	.9920	.9922	.9925	.9927	.9929	.9931	.9932	.9934	.9936
2.5	.9938	.9940	.9941	.9943	.9945	.9946	.9948	.9949	.9951	.9952
2.6	.9953	.9955	.9956	.9957	.9959	.9960	.9961	.9962	.9963	.9964
2.7	.9965	.9966	.9967	.9968	.9969	.9970	.9971	.9972	.9973	.9974
2.8	.9974	.9975	.9976	.9977	.9977	.9978	.9979	.9979	.9980	.9981
2.9	.9981	.9982	.9982	.9983	.9984	.9984	.9985	.9985	.9986	.9986
3.0	.9987	.9987	.9987	.9988	.9988	.9989	.9989	.9989	.9990	.9990

Binomialverteilung

	p	0.1		0.2		0.3		0.4		0.5	
n	xi	f(xi)	F(xi)	f(xi)	F(xi)	f(xi)	F(xi)	f(xi)	F(xi)	f(xi)	F(xi)
3	0	.7290	.7290	.5120	.5120	.3430	.3430	.2160	.2160	.1250	.1250
	1	.2430	.9720	.3840	.8960	.4410	.7840	.4320	.6480	.3750	.5000
	2	.0270	.9990	.0960	.9920	.1890	.9730	.2880	.9360	.3750	.8750
	3	.0010	1	.0080	1	.0270	1	.0640	1	.1250	1
4	0	.6561	.6561	.4096	.4096	.2401	.2401	.1296	.1296	.0625	.0625
	1	.2916	.9477	.4096	.8192	.4116	.6517	.3456	.4752	.2500	.3125
	2	.0486	.9963	.1536	.9728	.2646	.9163	.3456	.8208	.3750	.6875
	3	.0036	.9999	.0256	.9984	.0756	.9919	.1536	.9744	.2500	.9375
	4	.0001	1	.0016	1	.0081	1	.0256	1	.0625	1
5	0	.5905	.5905	.3277	.3277	.1681	.1681	.0778	.0778	.0313	.0313
	1	.3281	.9185	.4096	.7373	.3602	.5282	.2592	.3370	.1563	.1875
	2	.0729	.9914	.2048	.9421	.3087	.8369	.3456	.6826	.3125	.5000
	3	.0081	.9995	.0512	.9933	.1323	.9692	.2304	.9130	.3125	.8125
	4	.0005	1	.0064	.9997	.0284	.9976	.0768	.9898	.1563	.9688
	5	.0000	1	.0003	1	.0024	1	.0102	1	.0313	1
6	0	.5314	.5314	.2621	.2621	.1176	.1176	.0467	.0467	.0156	.0156
	1	.3543	.8857	.3932	.6554	.3025	.4202	.1866	.2333	.0938	.1094
	2	.0984	.9842	.2458	.9011	.3241	.7443	.3110	.5443	.2344	.3438
	3	.0146	.9987	.0819	.9830	.1852	.9295	.2765	.8208	.3125	.6563
	4	.0012	.9999	.0154	.9984	.0595	.9891	.1382	.9590	.2344	.8906
	5	.0001	1	.0015	.9999	.0102	.9993	.0369	.9959	.0938	.9844
	6	.0000	1	.0001	1	.0007	1	.0041	1	.0156	1
8	0	.4305	.4305	.1678	.1678	.0576	.0576	.0168	.0168	.0039	.0039
	1	.3826	.8131	.3355	.5033	.1977	.2553	.0896	.1064	.0313	.0352
	2	.1488	.9619	.2936	.7969	.2965	.5518	.2090	.3154	.1094	.1445
	3	.0331	.9950	.1468	.9437	.2541	.8059	.2787	.5941	.2188	.3633
	4	.0046	.9996	.0459	.9896	.1361	.9420	.2322	.8263	.2734	.6367
	5	.0004	1	.0092	.9988	.0467	.9887	.1239	.9502	.2188	.8555
	6	.0000	1	.0011	.9999	.0100	.9987	.0413	.9915	.1094	.9648
	7	.0000	1	.0001	1	.0012	.9999	.0079	.9993	.0313	.9961
	8	.0000	1	.0000	1	.0001	1	.0007	1	.0039	1
10	0	.3487	.3487	.1074	.1074	.0282	.0282	.0060	.0060	.0010	.0010
	1	.3874	.7361	.2684	.3758	.1211	.1493	.0403	.0464	.0098	.0107

	2	.1937	.9298	.3020	.6778	.2335	.3828	.1209	.1673	.0439	.0547
	3	.0574	.9872	.2013	.8791	.2668	.6496	.2150	.3823	.1172	.1719
	4	.0112	.9984	.0881	.9672	.2001	.8497	.2508	.6331	.2051	.3770
	5	.0015	.9999	.0264	.9936	.1029	.9527	.2007	.8338	.2461	.6230
	6	.0001	1	.0055	.9991	.0368	.9894	.1115	.9452	.2051	.8281
	7	.0000	1	.0008	.9999	.0090	.9984	.0425	.9877	.1172	.9453
	8	.0000	1	.0001	1	.0014	.9999	.0106	.9983	.0439	.9893
	9	.0000	1	.0000	1	.0001	1	.0016	.9999	.0098	.9990
	10	.0000	1	.0000	1	.0000	1	.0000	1	.0000	1

Poisson-Verteilung

	$\lambda = 0.1$		$\lambda = 0.2$		$\lambda = 0.3$		$\lambda = 0.4$		$\lambda = 0.5$	
x_i	$f(x_i)$	$F(x_i)$	$f(x_i)$	$F(x_i)$	$f(x_i)$	$F(x_i)$	$f(x_i)$	$F(x_i)$	$f(x_i)$	$F(x_i)$
0	.9048	.9048	.8187	.8187	.7408	.7408	.6703	.6703	.6065	.6065
1	.0905	.9953	.1637	.9825	.2222	.9631	.2681	.9384	.3033	.9098
2	.0045	.9998	.0164	.9989	.0333	.9964	.0536	.9921	.0758	.9856
3	.0002	1	.0011	.9999	.0033	.9997	.0072	.9992	.0126	.9982
4	.0000	1	.0001	1	.0003	1	.0007	.9999	.0016	.9998
5							.0001	1	.0002	1

	$\lambda = 0.6$		$\lambda = 0.7$		$\lambda = 0.8$		$\lambda = 0.9$		$\lambda = 1$	
x_i	$f(x_i)$	$F(x_i)$	$f(x_i)$	$F(x_i)$	$f(x_i)$	$F(x_i)$	$f(x_i)$	$F(x_i)$	$f(x_i)$	$F(x_i)$
0	.5488	.5488	.4966	.4966	.4493	.4493	.4066	.4066	.3679	.3679
1	.3293	.8781	.3476	.8442	.3595	.8088	.3659	.7725	.3679	.7358
2	.0988	.9769	.1217	.9659	.1438	.9526	.1647	.9371	.1839	.9197
3	.0198	.9966	.0284	.9942	.0383	.9909	.0494	.9865	.0613	.9810
4	.0030	.9996	.0050	.9992	.0077	.9986	.0111	.9977	.0153	.9963
5	.0004	1	.0007	.9999	.0012	.9998	.0020	.9997	.0031	.9994
6			.0001	1	.0002	1	.0003	1	.0005	.9999
7									.0001	1

	$\lambda = 1.5$		$\lambda = 2$		$\lambda = 3$		$\lambda = 4$		$\lambda = 5$	
x_i	$f(x_i)$	$F(x_i)$	$f(x_i)$	$F(x_i)$	$f(x_i)$	$F(x_i)$	$f(x_i)$	$F(x_i)$	$f(x_i)$	$F(x_i)$
0	.2231	.2231	.1353	.1353	.0498	.0498	.0183	.0183	.0067	.0067
1	.3347	.5578	.2707	.4060	.1494	.1991	.0733	.0916	.0337	.0404
2	.2510	.8088	.2707	.6767	.2240	.4232	.1465	.2381	.0842	.1247
3	.1255	.9344	.1804	.8571	.2240	.6472	.1954	.4335	.1404	.2650
4	.0471	.9814	.0902	.9473	.1680	.8153	.1954	.6288	.1755	.4405
5	.0141	.9955	.0361	.9834	.1008	.9161	.1563	.7851	.1755	.6160
6	.0035	.9991	.0120	.9955	.0504	.9665	.1042	.8893	.1462	.7622
7	.0008	.9998	.0034	.9989	.0216	.9881	.0595	.9489	.1044	.8666
8	.0001	1	.0009	.9998	.0081	.9962	.0298	.9786	.0653	.9319
9			.0002	1	.0027	.9989	.0132	.9919	.0363	.9682
10					.0008	.9997	.0053	.9972	.0181	.9863
11					.0002	.9999	.0019	.9991	.0082	.9945
12					.0001	1	.0006	.9997	.0034	.9980
13							.0002	.9999	.0013	.9993
14							.0001	1	.0005	.9998
15									.0002	.9999
16									.0000	1

Chi-Quadrat-Verteilung

n ist die Anzahl der Freiheitsgrade; die Spaltenüberschriften beschreiben $F(\chi^2)$

n	0,001	0,005	0,01	0,025	0,05	0,1	0,25	0,995	0,999
1	0	0	0	0,001	0,004	0,016	0,102	7,879	10,827
2	0,002	0,01	0,02	0,051	0,103	0,211	0,575	10,597	13,815
3	0,024	0,072	0,115	0,216	0,352	0,584	1,213	12,838	16,266
4	0,091	0,207	0,297	0,484	0,711	1,064	1,923	14,86	18,466
5	0,21	0,412	0,554	0,831	1,145	1,61	2,675	16,75	20,515
6	0,381	0,676	0,872	1,237	1,635	2,204	3,455	18,548	22,457
7	0,599	0,989	1,239	1,69	2,167	2,833	4,255	20,278	24,321
8	0,857	1,344	1,647	2,18	2,733	3,49	5,071	21,955	26,124
9	1,152	1,735	2,088	2,7	3,325	4,168	5,899	23,589	27,877
10	1,479	2,156	2,558	3,247	3,94	4,865	6,737	25,188	29,588
11	1,834	2,603	3,053	3,816	4,575	5,578	7,584	26,757	31,264
12	2,214	3,074	3,571	4,404	5,226	6,304	8,438	28,3	32,909
13	2,617	3,565	4,107	5,009	5,892	7,041	9,299	29,819	34,527
14	3,041	4,075	4,66	5,629	6,571	7,79	10,165	31,319	36,124
15	3,483	4,601	5,229	6,262	7,261	8,547	11,037	32,801	37,698
16	3,942	5,142	5,812	6,908	7,962	9,312	11,912	34,267	39,252
17	4,416	5,697	6,408	7,564	8,672	10,085	12,792	35,718	40,791
18	4,905	6,265	7,015	8,231	9,39	10,865	13,675	37,156	42,312
19	5,407	6,844	7,633	8,907	10,117	11,651	14,562	38,582	43,819
20	5,921	7,434	8,26	9,591	10,851	12,443	15,452	39,997	45,314
21	6,447	8,034	8,897	10,283	11,591	13,24	16,344	41,401	46,796
22	6,983	8,643	9,542	10,982	12,338	14,041	17,24	42,796	48,268
23	7,529	9,26	10,196	11,689	13,091	14,848	18,137	44,181	49,728
24	8,085	9,886	10,856	12,401	13,848	15,659	19,037	45,558	51,179
25	8,649	10,52	11,524	13,12	14,611	16,473	19,939	46,928	52,619
26	9,222	11,16	12,198	13,844	15,379	17,292	20,843	48,29	54,051
27	9,803	11,808	12,878	14,573	16,151	18,114	21,749	49,645	55,475
28	10,391	12,461	13,565	15,308	16,928	18,939	22,657	50,994	56,892
29	10,986	13,121	14,256	16,047	17,708	19,768	23,567	52,335	58,301
30	11,588	13,787	14,953	16,791	18,493	20,599	24,478	53,672	59,702
40	17,917	20,707	22,164	24,433	26,509	29,051	33,66	66,766	73,403
50	24,674	27,991	29,707	32,357	34,764	37,689	42,942	79,49	86,66
60	31,738	35,534	37,485	40,482	43,188	46,459	52,294	91,952	99,608
70	39,036	43,275	45,442	48,758	51,739	55,329	61,698	104,215	112,317
80	46,52	51,172	53,54	57,153	60,391	64,278	71,145	116,321	124,839
90	54,156	59,196	61,754	65,647	69,126	73,291	80,625	128,299	137,208
100	61,918	67,328	70,065	74,222	77,929	82,358	90,133	140,17	149,449

Chi-Quadrat-Verteilung (Fortsetzung)

n	0,5	0,75	0,9	0,95	0,975	0,99
1	0,455	1,323	2,706	3,841	5,024	6,635
2	1,386	2,773	4,605	5,991	7,378	9,21
3	2,366	4,108	6,251	7,815	9,348	11,345
4	3,357	5,385	7,779	9,488	11,143	13,277
5	4,351	6,626	9,236	11,07	12,832	15,086
6	5,348	7,841	10,645	12,592	14,449	16,812
7	6,346	9,037	12,017	14,067	16,013	18,475
8	7,344	10,219	13,362	15,507	17,535	20,09
9	8,343	11,389	14,684	16,919	19,023	21,666
10	9,342	12,549	15,987	18,307	20,483	23,209
11	10,341	13,701	17,275	19,675	21,92	24,725
12	11,34	14,845	18,549	21,026	23,337	26,217
13	12,34	15,984	19,812	22,362	24,736	27,688
14	13,339	17,117	21,064	23,685	26,119	29,141
15	14,339	18,245	22,307	24,996	27,488	30,578
16	15,338	19,369	23,542	26,296	28,845	32
17	16,338	20,489	24,769	27,587	30,191	33,409
18	17,338	21,605	25,989	28,869	31,526	34,805
19	18,338	22,718	27,204	30,144	32,852	36,191
20	19,337	23,828	28,412	31,41	34,17	37,566
21	20,337	24,935	29,615	32,671	35,479	38,932
22	21,337	26,039	30,813	33,924	36,781	40,289
23	22,337	27,141	32,007	35,172	38,076	41,638
24	23,337	28,241	33,196	36,415	39,364	42,98
25	24,337	29,339	34,382	37,652	40,646	44,314
26	25,336	30,435	35,563	38,885	41,923	45,642
27	26,336	31,528	36,741	40,113	43,195	46,963
28	27,336	32,62	37,916	41,337	44,461	48,278
29	28,336	33,711	39,087	42,557	45,722	49,588
30	29,336	34,8	40,256	43,773	46,979	50,892
40	39,335	45,616	51,805	55,758	59,342	63,691
50	49,335	56,334	63,167	67,505	71,42	76,154
60	59,335	66,981	74,397	79,082	83,298	88,379
70	69,334	77,577	85,527	90,531	95,023	100,425
80	79,334	88,13	96,578	101,879	106,629	112,329
90	89,334	98,65	107,565	113,145	118,136	124,116
100	99,334	109,141	118,498	124,342	129,561	135,807

Student-t-Verteilung

n	t(0.995)	t(0.99)	t(0.975)	t(0.95)	t(0.9)	t(0.85)	t(0.8)
1	63,656	31,821	12,706	6,314	3,078	1,963	1,376
2	9,925	6,965	4,303	2,92	1,886	1,386	1,061
3	5,841	4,541	3,182	2,353	1,638	1,25	0,978
4	4,604	3,747	2,776	2,132	1,533	1,19	0,941
5	4,032	3,365	2,571	2,015	1,476	1,156	0,92
6	3,707	3,143	2,447	1,943	1,44	1,134	0,906
7	3,499	2,998	2,365	1,895	1,415	1,119	0,896
8	3,355	2,896	2,306	1,86	1,397	1,108	0,889
9	3,25	2,821	2,262	1,833	1,383	1,1	0,883
10	3,169	2,764	2,228	1,812	1,372	1,093	0,879
11	3,106	2,718	2,201	1,796	1,363	1,088	0,876
12	3,055	2,681	2,179	1,782	1,356	1,083	0,873
13	3,012	2,65	2,16	1,771	1,35	1,079	0,87
14	2,977	2,624	2,145	1,761	1,345	1,076	0,868
15	2,947	2,602	2,131	1,753	1,341	1,074	0,866
16	2,921	2,583	2,12	1,746	1,337	1,071	0,865
17	2,898	2,567	2,11	1,74	1,333	1,069	0,863
18	2,878	2,552	2,101	1,734	1,33	1,067	0,862
19	2,861	2,539	2,093	1,729	1,328	1,066	0,861
20	2,845	2,528	2,086	1,725	1,325	1,064	0,86
21	2,831	2,518	2,08	1,721	1,323	1,063	0,859
22	2,819	2,508	2,074	1,717	1,321	1,061	0,858
23	2,807	2,5	2,069	1,714	1,319	1,06	0,858
24	2,797	2,492	2,064	1,711	1,318	1,059	0,857
25	2,787	2,485	2,06	1,708	1,316	1,058	0,856
26	2,779	2,479	2,056	1,706	1,315	1,058	0,856
27	2,771	2,473	2,052	1,703	1,314	1,057	0,855
28	2,763	2,467	2,048	1,701	1,313	1,056	0,855
29	2,756	2,462	2,045	1,699	1,311	1,055	0,854
30	2,75	2,457	2,042	1,697	1,31	1,055	0,854
40	2,704	2,423	2,021	1,684	1,303	1,05	0,851
60	2,66	2,39	2	1,671	1,296	1,045	0,848
120	2,617	2,358	1,98	1,658	1,289	1,041	0,845
¥	2,576	2.326	1,96	1,645	1,282	1.036	0,842

6.5 Weiterführende Literatur

Wie wir schon in der Einleitung gesagt haben – gute Bücher über theoretische Statistik veralten nicht so rasch, wenn ihr pädagogisches beziehungsweise didaktisches Konzept stimmt. Es gibt über die letzten Jahrzehnte eine recht große Zahl von Lehrbüchern der theoretischen Statistik, und natürlich eine unabsehbare Menge an Spezialliteratur zu einzelnen Themenbereichen oder Fragestellungen. Die folgende Zusammenstellung bietet nur einen ersten Einstieg in die weiterführende Literatur.

Als nach wie vor hervorragendes Lehrbuch für den Anfänger ist zu nennen:

> Wagenführ, Rolf (1971): Statistik leicht gemacht, Bd. I: Einführung in die deskriptive Statistik. 6. Aufl. Köln: Bund-Verlag
>
> Wagenführ, Rolf/Tiede, Manfed/Voß, Werner (1971): Statistik leicht gemacht, Bd. II: Einführung in die induktive Statistik. Köln:Bund-Verlag

Als weiteres, modernes Lehrbuch für Anfänger:

> Quatember, Andreas (2005): Statistik ohne Angst vor Formeln, München: Pearson

Unter den Lehrbüchern möchten wir noch folgende herausgreifen:

Bourier, Günther (2005): Beschreibende Statistik: praxisorientierte Einführung; mit Aufgaben und Lösungen, 6. Aufl., Wiesbaden: Gabler

Fahrmeir, Ludwig/Künstler, Rita/Pigeot, Iris/Tutz, Gerhard (2004): Statistik – Der Weg zur Datenanalyse. 5. Aufl. Berlin/Heidelberg u.a.: Springer, dazu diess./Caputo, Angelika/Lang, Stefan (2004): Arbeitsbuch Statistik, 4. Aufl., Berlin/Heidelberg u.a.: Springer

Sachs, Lothar (2002): Angewandte Statistik – Anwendung statistischer Methoden 10. Aufl., Berlin/Heidelberg: Springer

Leiner, Bernd (2000): Einführung in die Statistik, 8. Aufl., München: R. Oldenbourg

Menges, Günter (1982): Die Statistik. Wiesbaden: Gabler

Menges, Günter (1973): Grundriß der Statistik, Bd. 1: Theorie, 2. Aufl., Opladen: Westdeutscher Verlag

Menges, Günter/Skala, Heinz (1973): Grundriß der Statistik, Bd. 2: Daten – Ihre Gewinnung und Verarbeitung. Opladen: Westdeutscher Verlag

Speziell für die Anwendungen der Statistik in den Wirtschaftswissenschaften ist gedacht:

Schira, Josef (2005): Statistische Methoden der VWL und BWL – Theorie und Praxis. 2. Aufl., München: Pearson

Moosmüller, Gertrud (2004): Methoden der empirischen Wirtschaftsforschung. München: Pearson

Eine große Zahl von Lehrbüchern beschäftigt sich auch speziell mit den Anwendungsgebieten der sozialwissenschaftlichen Forschung:

Kriz, Jürgen (1973): Statistik in den Sozialwissenschaften – Einführung und kritische Diskussion. Reinbek: Rowohlt

Gaensslen, Hermann/Schubö, Walter (1976): Einfache und komplexe statistische Analyse. München/Basel: Ernst Reinhardt

Opp, Karl-Dieter/Schmidt, Peter (1976): Einführung in die Mehrvariablenanalyse – Grundlagen der Formulierung und Prüfung komplexer sozialwissenschaftlicher Aussagen. Reinbek: Rowohlt

Spezielle Literatur zu bestimmten Erhebungs- und Auswertungstechniken gibt es vor allem auch für die Sozialwissenschaften:

Bühner, Markus (2006): Einführung in die Test- und Fragebogenkonstruktion. 2. Aufl. München: Pearson

Bogner, Alexander/Littig, Beate/Menz, Wolfgang (Hrsg.) (2005): Das Experteninterview – Theorie, Methode, Anwendung. 2. Aufl., Wiesbaden: VS Verlag für Sozialwissenschaften

Brüsemeister, Thomas (2000): Qualitative Forschung. Opladen: Westdeutscher Verlag

Als Lernhilfe für das Programm MS Excel sei auf die Online-Hilfe verwiesen; für SPSS gibt es ein umfassendes Lehrbuch:

Bühl, Achim (2006): SPS 14 – Einführung in die moderne Datenanalyse. 10. Aufl., München: Pearson

Wer sich näher mit der Methodik der Manipulationsmöglichkeiten, Fälschungen und einseitigen Anwendungen statistischer Verfahren oder tendenziösen Interpretationen von Ergebnissen befassen möchte, dem seien folgende Quellen empfohlen:

Huff, Derroll (1954): How to Lie With Statistics. New York: W.W. Norton

Statistisches Landesamt Baden-Württemberg (Hrsg.) (1996): Ich glaube nur der Statistik… Was Winston Churchill über die Zahlen und die Statistik wirklich sagte. Stuttgart: Selbstverlag

Krämer, Walter/Krämer, Denis/Trenkler, Götz (1998): Das neue Lexikon der populären Irrtümer. Frankfurt: Eichborn

Neue Zürcher Zeitung – NZZ Folio vom Januar 2006: Statistik – Zählen und gezählt werden

Zur Geschichte der Statistik sei genannt:

Pearson, Egon S./Kendall, Maurice G. (Hrsg.) (1970). Studies in the History of Statistics and Probability. London: Griffin

Cantor, Moritz (1907): Vorlesung über Geschichte der Mathematik, 4 Bde, 3. Aufl., Leipzig: Teubner, Neudruck Vaduz: Sändig Reprint 2004

Zu den im vorliegenden Buch genannten Anwendungsbereichen der Statistik lassen sich folgende Quellen als Einstiegsliteratur benennen:

a) Zur Bevölkerungsstatistik beziehungsweise Demographie
Bertelsmann Stiftung (Hrsg.) (2006): Wegweiser Demographischer Wandel 2020. Gütersloh: Verlag Bertelsmann Stiftung

b) Zur Wirtschaftsstatistik
Haslinger, Franz (1986): Volkswirtschaftliche Gesamtrechnung. 4. Aufl., München/Wien: Oldenbourg
Brümmerhoff, Dieter (2002): Volkswirtschaftliche Gesamtrechnungen. 7. Aufl., München: Oldenbourg

c) Zur Sozialstatistik
Geißler, Rainer (2006): Die Sozialstruktur Deutschlands – Zur gesellschaftlichen Entwicklung mit einer Bilanz zur Vereinigung. 4 Aufl., Wiesbaden: VS Verlag für Sozialwissenschaften

d) Zur Demoskopie
Noelle, Elisabeth (1963): Umfragen in der Massengesellschaft – Einführung in die Methoden der Demoskopie. Reinbek: Rowohlt

e) Zur amtlichen Statistik: Es gibt eine große Zahl von regelmäßig veröffentlichten internationalen und nationalen Quellen, in denen amtliche Statistiken publiziert werden. Eine detaillierte Auflistung würde den Rahmen hier sprengen. Die meisten internationalen Organisationen im UN-Bereich veröffentlichen Jahresberichte mit ausführlichen statistischen Anhängen. Explizit nennen möchten wir:
United Nations Development Programme (fortlaufend): Bericht über die menschliche Entwicklung (englisch: Human Development Report)
World Bank (fortlaufend): Weltentwicklungsbericht (englisch: World Development Report)
World Bank (fortlaufend): World Development Indicators
International Labour Office (fortlaufend): World Labour Report
World Health Organization (fortlaufend): World Health Report
International Monetary Fund (fortlaufend): World Economic Outlook
Europäische Zentralbank (fortlaufend): Jahresberichte und Monatsberichte

Alle diese Veröffentlichungen werden von den entsprechenden Organisationen jeweils selbst herausgegeben und im Eigenverlag publiziert. An europäischen und nationalen amtlichen Quellen ist zu nennen:

Eurostat (fortlaufend): Europa im Blick der Statistik
Statistisches Bundesamt (Hrsg.) (fortlaufend): Statistisches Jahrbuch für die Bundesrepublik Deutschland. Stuttgart: Metzler-Poeschel

Sowie die Fachserien des Statistischen Bundesamts und die Veröffentlichungen der Statistischen Landesämter. Die amtlichen Quellen bieten mittlerweile in der Regel auch ein umfassendes Internet-Angebot an Daten (siehe nächster Abschnitt).

…und wer es zum guten Schluss nicht immer ganz ernst haben will:

DekaBank (Hrsg.) (2006): Die Glaskugel muss ins Eckige! Ökonomische Ansichten zur Fußball-WM. Frankfurt: Selbstverlag

6.6 Internet-Quellen

Wie bereits gesagt, gibt es eine außerordentlich große Vielfalt an statistischen Informationen, die über das Internet zugänglich sind. Wir können an dieser Stelle nur die wichtigsten „Einstiegspunkte" in einer internetgestützten Recherche angeben.

Die Statistiken der United Nations und ihrer angeschlossenen Organisationen erschließen sich über die United Nations Statistics Division:

http://unstats.un.org/unsd/default.htm

Auf der Website der UNSD gibt es auch eine Link-Seite mit Zugängen zu sämtlichen (!) nationalen Statistik-Ämtern dieser Erde:

http://unstats.un.org/unsd/methods/inter-natlinks/sd_natstat.htm

Das Europäische Statistische Amt ist über folgende Seite erreichbar:

http://ec.europa.eu/eurostat/

Das Statistische Bundesamt bietet zusätzlich einen europäischen Datenservice:

http://www.eds-destatis.de/

Das Angebot des Statistischen Bundesamtes in Deutschland ist zentral verfügbar über:

http://www.destatis.de/

Das Statistische Bundesamt und die Statistischen Ämter bieten ein gemeinsames Internet-Portal:

http://www.statistik-portal.de/Statistik-Portal/

Über dieses Portal sind wiederum die einzelnen Statistischen Landesämter zu erreichen

http://www.statistik-portal.de/Statistik-Portal/LinksUebersicht.asp

Diese Seite erschließt auch die Zugänge zu den Statistik-Ämtern anderer Länder und supranationaler Organisationen. Einige wenige grundlegende Daten über Gemeinden in Deutschland lassen sich ebenfalls über dieses Portal abfragen:

http://www.statistik-portal.de/Statistik-Portal/gemeindeverz.asp

Ansonsten empfiehlt es sich jedoch, statistische Informationen über Kommunen entweder über die Datenangebote der Landesämter oder der Websites der Kommunen selbst zu recherchieren – in der Regel über: *www.name-der-kommune.de* Viele weitere Daten über Kommunen lassen sich zentral über das Online Portal deutscher Kommunen erschließen:

http://www.kommon.de/index.html

7 Noten

1. Aus Gründen der Sprachästethik verwenden wir im Folgenden nicht immer sowohl die weibliche als auch die männliche Form, wenn beide Geschlechter angesprochen werden sollen – wir bemühen uns zu wechseln.
2. Vgl. Wagenführ, Rolf (1971): Statistik leicht gemacht, Bd. I: Einführung in die deskriptive Statistik. 6. Aufl. Köln: Bund-Verlag, z.B. 92, 96, 99, 163.
3. Der gerne und viel zitierte Satz: „Ich glaube nur der Statistik, die ich selbst gefälscht habe“ wird zumeist Winston Churchill zugeschrieben, stammt allerdings vermutlich von Joseph Goebbels, der den englischen Premier als Lügner darstellen wollte und ihm dazu das entsprechende Zitat unterschob: Nachzulesen in Statistisches Landesamt Baden-Württemberg (Hrsg.) (1996): Ich glaube nur der Statistik … Was Winston Churchill über die Zahlen und die Statistik wirklich sagte. Stuttgart: Selbstverlag, oder auch in Krämer, Walter/Krämer, Denis/Trenkler, Götz (1998): Das neue Lexikon der populären Irrtümer. Frankfurt: Eichborn: vgl. auch http://www.fscklog.com/banales/index.html
4. Eine Übersicht zur Geschichte der Statistik findet sich bei Menges, Günter (1982): Die Statistik. Wiesbaden: Gabler, Kap. 1; vgl. auch Pearson, Egon S./Kendall, Maurice G. (Hrsg.) (1970): Studies in the History of Statistics and Probability. London: Griffin und als klassisches Werk Cantor, Moritz (1907): Vorlesung über Geschichte der Mathematik, 4 Bde, 3. Aufl. Leipzig: Teubner, Neudruck Vaduz: Sändig Reprint 2004.
5. Vgl. dazu Wagenführ, Rolf (1971): op.cit, 14.
6. Menges, Günter (1982), op.cit., 20.

7. Vgl. ausführlicher auch Schira, Josef (2005): Statistische Methoden der VWL und BWL – Theorie undPraxis. 2. Aufl., München: Pearson, 19ff.; Menges, Günter/Skala, Heinz (1973): Grundriß der Statistik, Bd. 2: Daten – Ihre Gewinnung und Verarbeitung. Opladen: Westdeutscher Verlag, 28ff.
8. Siehe z.B. Haslinger, Franz (1986): Volkswirtschaftliche Gesamtrechnung. 4. Aufl., München/Wien: Oldenbourg, 43.
9. Siehe Brümmerhoff, Dieter (2002): Volkswirtschaftliche Gesamtrechnungen. 7. Aufl., München: Oldenbourg, 28.
10. Vgl. Schira (2005), op.cit., 20f.
11. Menges (1982), op.cit., 89.
12. Zum Folgenden vgl. ausführlich ibid., 97 ff.
13. Vgl. ausführlich Menges/Skala (1973), op.cit., Bd. 2, 67.
14. Der Mikrozensus ist die amtliche Repräsentativstatistik über die Bevölkerung und den Arbeitsmarkt, an der jährlich 1% aller Haushalte in Deutschland beteiligt sind (laufende Haushaltsstichprobe). Insgesamt nehmen rund 370.000 Haushalte mit 820.000 Personen am Mikrozensus teil; darunter etwa 160.000 Personen in rund 70.000 Haushalten in den neuen Bundesländern und Berlin-Ost. Im Statistischen Bundesamt erfolgt die organisatorische und technische Vorbereitung des Mikrozensus. Die Durchführung der Befragung und die Aufbereitung obliegt den statistischen Landesämtern; vgl. www.destatis.de/micro/d/micro_c1.htm
15. Das SOEP ist eine seit 1984 laufende jährliche Wiederholungsbefragung von Deutschen, Ausländern und Zuwanderern in den alten und neuen Bundesländern. Die Stichprobe umfasste im Erhebungsjahr 2004 mehr als 12.000 Haushalte mit fast 24.000 Personen. Themenschwerpunkte sind unter anderem Haushaltszusammensetzung, Erwerbs- und Familienbiographie, Erwerbsbeteiligung und berufliche Mobilität, Einkommensverläufe, Gesundheit und Lebenszufriedenheit. Der SOEP-Datensatz wird universitären Forschungseinrichtungen im In- und Ausland für Forschung und Lehre in SPSS-, SAS-, Stataund ASCII- Format weitergegeben; vgl. www.diw.de/deutsch/sop/index.html
16. Vgl. ausführlich Schira (2005), op.cit., 22 ff.; Menges (1982), op.cit., 26 ff.
17. Online-Schulungen sind sowohl in der PC- als auch in der Apple Macintosh-Version in das Programm integriert und lassen sich – sofern man die Theorie „hinter" den Auswertungsmodulen beherrscht, in einem „learning by doing"-Verfahren erlernen.
18. Derzeit aktuell ist SPSS 14, als umfassendes Lehrbuch vgl. Bühl, Achim (2006): SPSS 14 – Einführung in die moderne Datenanalyse. 10. Aufl., München: Pearson.
19. Vgl. dazu ausführlich Wagenführ (1972), op.cit., 61 ff.
20. Ibid., 65.
21. Angaben nach Statistisches Bundesamt (Hrsg.) (2005): Statistisches Jahrbuch für die Bundesrepublik Deutschland. Stuttgart: Metzler-Poeschel, @@@
22. Vgl. dazu Wagenführ, Rolf (1971), op.cit., Bd. II, 27 ff.
23. Wagenführ, Rolf (1971), op.cit., Bd. II, 54.
24. Zur Erinnerung: Die Standardabweichung ergibt sich aus der Wurzel aus der Varianz!
25. Vgl. Kap. 3.4.1.