S T A T I S T I K

Harald Strelec

Gregor Laaha, Friedrich Leisch, Bernhard Spangl

Universität für Bodenkultur

Department für Raum, Landschaft und Infrastruktur
Institut für Statistik

Peter-Jordan-Straße 82, 1190 Wien

http://statistik.boku.ac.at/

2. März 2020

In memoriam
Harald Strelec

Vorwort

Dieses Skriptum wurde zum größten Teil von meinem Vorgänger Univ.Prof. Harald Strelec verfasst und ich bin dankbar, seinen in jahrelanger Lehre wohlerprobten Text weiterverwenden zu können. Im Vergleich zur letzten Version aus dem Jahr 2006 wurden in den Kapiteln 1–4 nur einige Grafiken modernisiert bzw. fehlende Grafiktypen ergänzt, und der Abschnitt über Wahrscheinlichkeitsnetze durch die heute gebräuchlicheren QQ-Diagramme ersetzt.

In Kapitel 5 ist nun einheitlich bei einseitigen Tests immer μ_0 Teil der Nullhypothese, d.h. die Nullhypothese ist immer von Form „kleiner gleich“ oder „größer gleich“, während die Alternative die entsprechende scharfe Ungleichung „größer“ oder „kleiner“ ist. Tabelle 5.13 enthielt fehlerhafte Werte, diese wurden korrigiert und die Tabelle insgesamt erweitert.

Aus Kapitel 6 wurden die seit Jahren nicht mehr unterrichteten Abschnitte über Varianzanalyse mit zufälligen Effekten entfernt. Das Kapitel über mehrfache Gruppenvergleiche wird derzeit ebenfalls nicht in den Grundvorlesungen unterrichtet und daher weiter nach hinten geschoben. Kapitel 7 wurde um zahlreiche anschauliche Grafiken ergänzt sowie die Formeln zur Berechnung der Regressionsparameter deutlich vereinfacht. In Kapitel 8 wurden die Rang- und Vorzeichentests vor die Abschnitte über Verteilungstests gereiht.

Das Skriptum ist nicht nur als Text zur reinen Prüfungsvorbereitung gedacht, sondern auch als erstes Nachschlagewerk in Sachen Statistik. Der Text ist daher etwas umfangreicher als der Stoff der Grundvorlesungen aus Statistik an der BOKU. Des weiteren werden je nach Wichtigkeit für die jeweilige Studienrichtung manche Abschnitte genauer oder weniger genau durchgenommen. Welche Teile des Skriptums (nicht) prüfungsrelevant sind wird in den einzelnen Lehrveranstaltungen bekannt gegeben.

Wien, im April 2013 Friedrich Leisch

Inhaltsverzeichnis

.

Kapitel 1

Einleitung

In vielen Situationen des beruflichen und privaten Alltags lässt sich bei kritischer Betrachtungsweise das Phänomen Ungewissheit, Unsicherheit oder gar Zufälligkeit in unterschiedlich starker Ausprägung erkennen. Zur Beschreibung und Analyse derartiger Situationen dient die *Wahrscheinlichkeitsrechnung*. Diese stellt einen unmittelbaren Zweig der Mathematik dar. Ein Teil dieser Lehrveranstaltung ist diesem Thema gewidmet. Zur Modellbildung sind dabei i.a. endlich– oder unendlichdimensionale Parameter notwendig.

Eine Aufgabe der *Statistik* besteht darin, diese Parameter auf Grund von realen Beobachtungen *(Stichproben)* zu *schätzen* oder Aussagen darüber zu *testen*. Auch dieser Zweig der *Stochastik* ist häufig eher mathematisch ausgerichtet und wird manchmal als *schließende Statistik* bezeichnet.

Für die *angewandte Statistik* bilden die Probleme der realen Wirklichkeit den Ausgangspunkt. Zentrales Objekt sind real erhobene Daten. Sie dienen bereits häufig der Modellgewinnung *(explorative Datenanalyse, exploratory data analysis, EDA)* und auch die Methoden der angewandten Statistik orientieren sich an ihnen. Dies bedingt die Berücksichtigung spezifischer Beobachtungssituationen und anwendungsbedingter Problemstellungen. Damit stellen angewandte und mathematische Statistik zwar nicht konkurrierende, aber doch z.T. sehr unterschiedlich orientierte Zweige der Statistik dar.

Schließlich fällt auch die Gewinnung, Beschreibung und Aufbereitung von realen Daten in den Aufgabenbereich der Statistik. Die *beschreibende Statistik* stellt dazu eine Reihe teilweise traditioneller, teilweise moderner, EDV–orientierter Methoden zur Verfügung. Es hat sich bewährt, mit diesen eine Einführung in die Stochastik zu beginnen, weil

1) die *Datenaufbereitung* grundsätzlich den ersten Schritt in der Datenanalyse darstellen sollte und

2) viele Begriffe der Wahrscheinlichkeitsrechnung durch analoge in der beschreibenden Statistik veranschaulicht und motiviert werden.

Die folgenden Beispiele sollen eine Motivation für typische Fragestellungen in der angewandten Statistik vermitteln und einen Überblick über dabei einzusetzende Methoden bieten. Die Auswahl von Zweigen der angewandten Statistik beschränkt sich im folgenden auf die wesentlichen, die auch in der praktischen Anwendung am weitesten verbreitet sind und wirklich mit Erfolg eingesetzt werden können.

Beispiel 1.1 Auf einer Obstplantage wurden 80 Äpfel der Sorte 'Golddelicious/Klasse I' auf ihr Gewicht (in *dag*) hin untersucht. Dabei konnte man die in Tab. 1.1 zusammengestellten Werte beobachten.

Tabelle 1.1: *Prüfdaten*

14.7	16.0	15.8	15.6	17.4	16.2	16.0	15.8	13.9	13.5
15.2	13.8	13.2	16.5	15.3	17.7	17.3	16.2	16.2	13.8
18.5	17.0	14.7	15.3	14.6	15.3	19.0	14.2	17.7	14.4
13.8	19.0	14.5	13.4	14.7	13.7	14.8	13.4	15.8	14.3
16.4	18.1	14.4	14.9	19.4	14.6	15.3	15.6	15.6	15.0
16.3	15.9	14.8	15.4	16.1	15.4	15.0	14.8	15.7	14.2
15.3	16.0	15.5	16.7	18.7	13.4	15.4	16.4	14.7	17.5
16.0	15.7	15.7	16.3	17.7	16.0	18.5	15.6	15.5	14.9

Es stellt sich unmittelbar die Frage nach dem Aussagegehalt des obigen Datensatzes:

- Entspricht die Produktion den Erwartungen?
- Wie liegt die Produktion, wodurch lässt sich dies ausdrücken?
- Kann die Produktion als homogen angesehen werden, wie streut sie, wie lässt sich das Streuverhalten beschreiben?
- Lassen sich Aussagen über den Anteil des Obstes treffen, der unter 170 g liegt? Wie groß ist dieser?
- Gelten derartige Aussagen auch wirklich "sicher"? Wenn nein, warum nicht und was sonst?

Eine erste, und oft schon weitreichende Antwort vermittelt eine grafische bzw. pseudografische Aufbereitung des betrachteten Datensatzes mit Hilfe der Verfahren der *beschreibenden Statistik*. Geeignete Kenngrößen helfen, oft umfangreiches Datenmaterial einfach und knapp zu beschreiben und vor allem vergleichbar zu machen. Ein eigenes Kapitel beschreibt diesen in der Praxis sehr oft anzutreffenden Zweig der Statistik.

Häufig ist aber die bloße Beschreibung von Beobachtungsdaten nicht ausreichend. Naheliegende Fragen sind etwa:

- Hat sich die Produktion im letzten Jahr (signifikant) verbessert? Streut das Gewicht weniger?
- Ist die Produktion der ausländischen Konkurrenz auf Grund des zur Verfügung stehenden Datenmaterials homogener?
- Sollte sich das obige Datenmaterial (siehe Übungen) als nicht homogen herausstellen, kann man vielleicht Einflussfaktoren finden, durch die sich die Inhomogenitäten zumindest weitgehend erklären lassen?

Es gilt dann, rational abgesicherte Aussagen (daher kommt in der Statistik auch Mathematik vor!) zu treffen, wie sie in der *schließenden Statistik* behandelt werden.

◇◇◇

Beispiel 1.2 Im Rahmen einer Studentenbefragung wurde das Interesse der Studenten an einem selbständigen Beruf ausgelotet. Dabei wurden u.a. auch Fragen nach dem Beruf der Eltern, der eigenen Schulbildung und einer allfälligen Praxis gestellt. Die Tab. 1.2 enthält einen kleinen Ausschnitt aus diesem Datensatz.

Die erste Frage, die sich unmittelbar stellt, ist die nach allfälligen *Zusammenhängen* von Variablen. Die Abhängigkeit von zwei, drei oder mehr Variablen wird durch die Analyse von *Kontingenztafeln* untersucht. Sie stellen ein effizientes Mittel zur Analyse *multivariater nominaler* Daten dar.

Tabelle 1.2: *Studentenumfrage — selbständiger Beruf*

1	*2a*	*2b*	*3*	*4*	*1*	*2a*	*2b*	*3*	*4*	*1*	*2a*	*2b*	*3*	*4*	*1*	*2a*	*2b*	*3*	*4*
1	3	2	1	0	4	3	4	2	1	2	1	2	2	1	2	1	3	1	0
2	3	3	4	0	2	2		3	0	1	3	2	2	1	2	3	2	1	1
2	3	4	2	1	4	1	3	1	1	1	3	2	2	0	3	2	2	1	1
2	4	3	2	1	2	3	3	1	1	2	1	3	2	0	3	2		1	1
2	3	3	1	0	1	3	4	2	1	2	3	3	1	0	4	3	4	1	0
2	2	3	1	1	3	3	4	3	0	1	4	2	4	1	2	1	3	1	0
3	3	3	1	0	3	1	4	2	1	1	1	3	1	0	3	3	2	1	0
1	3	2	2	1	4	4	1	1	0	3	3	4	1	0	2	3	3	2	1
3	3	3	1	0	2	1	4	2	0	2	4	4	1	0	2	3	3	1	0
2	3	3	1	0	2	3	3	2	0	3	1	3	2	0	2	3	2	2	1
2	1	3	1	0	5	3	3	1	0	2	4	4	1	0	2	3	3	1	0
2	3	3	1	0	4	3	2	2	1	1	3		1	1	2	4	4	2	1
3	3	3	1	0	3	3	3	1	0	2	3	3	1	1	2	2	2	1	0
2	3	4	1	1	2	3	3	1	0	2	1	3	4	1	2	3	4	1	0
3	3	3	1	0	2	4	3	1	1	2	1	3	1	1	2	1	3	2	0

1 ... Absicht, sich selbständig zu machen
- 1 ja
- 2 eventuell
- 3 eher nein
- 4 nein
- 5 weiß nicht

2a ... Beruf der Mutter
- 1 nicht erwerbstätig
- 2 selbständig erwerbstätig
- 3 Angestellte
- 4 Arbeiterin

2b ... Beruf des Vaters
Codierung analog zu Frage 2a

3 ... Eigene Schulbildung
- 1 AHS oder ähnliche
- 2 HTL oder ähnliche
- 3 HAK
- 4 andere

4 ... Praxiserfahrung
- 0 nein
- 1 ja

An diesem Beispiel lässt sich ein weiteres Problem unmittelbar verdeutlichen. Ist die Gesamtheit der Studenten nicht zu *heterogen*, um sie in einer gemeinsamen Analyse zu untersuchen? Wäre es nicht besser, *Strukturen* in der Grundgesamtheit zu suchen, für die sich die betrachteten Variablen *homogener* verhalten und für die daher eine statistische Analyse schärfer und damit aussagekräftiger wäre?

Die *Clusteranalyse* (oder *automatische Klassifikation*) bietet Möglichkeiten, Stichproben auf Grund geeigneter Kriterien in homogene Teilgruppen *(Cluster)* zu strukturieren. Die diesen entsprechenden Teile der Grundgesamtheit ergeben sich im Idealfall als *natürliche* Gruppen, im schlechtesten Fall sind sie inhaltlich nicht beschreibbar, in jedem Fall sind sie aber ein *geeigneter Ausgangspunkt* für die weitere effiziente statistische Analyse der Daten.

◇◇◇

Beispiel 1.3 Aus Lomb et al. (*Atm. Environment* **21**, 1987) bzw. Krapfenbauer und Hollermann (*Ozon in der Troposphäre – ...*, 1993) stammen die Daten aus Tab. 1.3. Sie beschreiben die Emission von NMHC–Stoffen (in μg je g und h) bei Laubbäumen in Abhängigkeit von der Temperatur (in $°C$).

Tabelle 1.3: *NMHC–Emission und Temperatur*

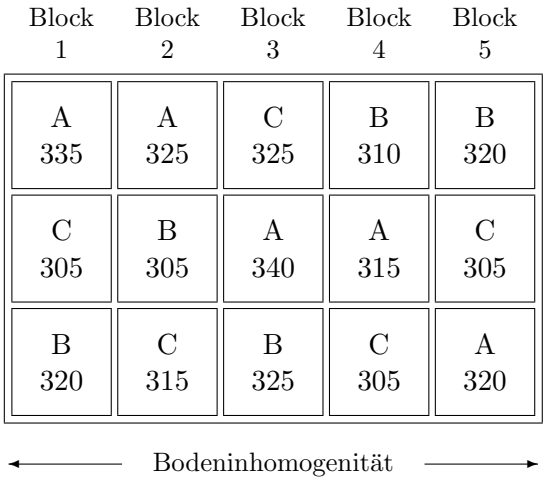
Temperatur (in $°C$)	NMHC–Emission (in $\mu g/g\,h$)				
12.5	0.7				
15.0	0.3				
17.5	1.0				
20.0	21.0	1.6			
22.5	40.0	136.0			
25.0	7.4	1.0	4.6	22.0	
27.5	12.0	4.6	117.0	1.6	6.3
30.0	22.0	100.0	4.6	22.0	10.0
32.5	34.0	22.0	2.2	63.0	10.0
35.0	10.0	19.0			
37.5	0.9	29.0			
40.0	100.0				

Dieses Beispiel stellt eine typische Situation bei der Durchführung und Auswertung von Versuchen dar. Hier liegt das Schwergewicht neben der Beschreibung von Versuch und Daten bei der Analyse möglicher Zusammenhänge und deren formelmäßigen Fassung. Die entsprechenden Modelle werden im Kapitel über *Regressionsanalyse* behandelt.

◇◇◇

Beispiel 1.4 Im Rahmen eines Zuckerrübenversuches wurden drei Rübensorten hinsichtlich ihres ha–Ertrages untersucht (Bätz et al., 1987). Um den Bodeneinfluss möglichst auszuschalten, bildete man fünf Teilflächen ("Blöcke") zu je drei Teilstücken. Für jeden Block wurden die drei Rübensorten zufällig den drei Teilstücken zugeordnet *(einfaktorielle Blockanlage)*. Die beobachteten ha–Erträge (in dt) findet man in Abb. 1.4, aus der auch die Zuteilung der Rübensorten auf die Teilstücke hervorgeht.

Abbildung 1.1: *Einfaktorielle Blockanlage*



Die wichtigste Frage ist hier die Klärung, ob es im (durchschnittlichen) ha–Ertrag zwischen den drei Rübensorten *merkliche* Unterschiede gibt. Ein weiterer, im allgemeinen nicht so wichtiger Aspekt kann die Untersuchung einer möglichen Inhomogenität zwischen den einzelnen Blöcken sein.

Diese und ähnliche Fragen lassen sich mit Methoden der *Varianzanalyse* systematisch behandeln und einer Beantwortung zuführen. Ein wesentlicher Aspekt bei der Analyse derartiger Probleme liegt in der *Planung* der Datengewinnung *(Versuchsplanung, experimental design)*. Sie dient der klaren Zuordnung von *zufallsbedingter Schwankung* auf der einen Seite und *systematischem Einfluss* auf der anderen Seite, um möglichst scharfe Aussagen zu erzielen. Die Varianzanalyse wiederum ist ein Spezialfall des linearen Regressionsmodells. Im linearen Modell kann der Einfluß mehrerer Einflußvariablen auf eine Zielvariable untersucht werden. Zusätzlich zu Blöcken und Rübensorten könnte etwa noch relevant sein, wieviel gedüngt wurde.

◇◇◇

Beispiel 1.5 Zwei Hühnerfarmen prüfen ihre Transporteinheiten (Kiste = 48 Packungen á 10 Eier = 480 Eier) auf die Einhaltung der Gewichtsklasse. Auf Grund branchenweiter Übereinstimmung sollten dabei nicht mehr als 1.5 % der Eier falsch klassiert sein.

Betrieb A organisiert die Prüfung so, dass aus einer Kiste (Los) 50 Eier zufällig entnommen und gewogen werden. Falls nicht mehr als zwei Eier beanstandet werden, gibt man das Los frei, im anderen Fall muss es vollständig durchgeprüft werden.

Der Konkurrenzbetrieb gestaltet die Prüfung anders. Zunächst werden 32 Eier zufällig ausgewählt und gewogen. Falls alle geprüften Eier entsprechen, wird das Los freigegeben, bei drei oder mehr falsch klassierten Eiern muss die Kiste total durchkontrolliert werden. Liegen aber bei den 32 geprüften Eiern nur eines oder zwei außerhalb der Gewichtsklasse, so werden weitere 32 Eier aus der Kiste entnommen und ebenfalls gewogen. Sind unter den jetzt *insgesamt* 64 überprüften Eiern mehr als drei falsch klassiert, wird das Los zunächst zurückgewiesen und muss vollständig kontrolliert werden, im anderen Fall wird es freigegeben.

Diese typische Situation aus der *(statistischen) Qualitätskontrolle (quality control)* führt unmittelbar zu folgenden Fragen:

- Was lässt sich allgemein über die Brauchbarkeit derartiger Prüfverfahren aussagen, wie wirkungsvoll und wie wirtschaftlich sind sie?
- Welches der beiden im Beispiel beschriebenen Verfahren ist vorzuziehen? Auf Grund welcher Unterscheidungskriterien?
- Gibt es möglicherweise effizientere Methoden?

◇ ◇ ◇

Beispiel 1.6 In einem Forst mit Fichtenmonokultur wurden bei Forstarbeiten (Schlägerungen, Ausholzen) laufend Angaben über das Alter der Bäume erhoben. Die Tab. 1.4 zeigt einen Ausschnitt des Datenmaterials, wobei sich die mit '+' versehenen Jahresangaben auf abgestorbene Bäume beziehen.

Tabelle 1.4: *Alter von Fichten*

10+	35	29	43	39	45	27	102+
42	33	33	89+	49	41	33	42
21	13+	56	36	51	23	51	26
58	37	41	39	31	35	43+	29
44	41	32+	72+	25	25	55	19+

- Kann man aus diesen Angaben die Lebensdauer von Fichten beschreiben?
- Lässt sich unter Ausnützung weiterer Information das Lebensdauerverhalten von Fichten in Abhängigkeit von Umweltfaktoren (Schadstoffeintrag, klimatische Verhältnisse, Höhenlage) beschreiben?

Derartige Fragestellungen zählen zur *Lebensdaueranalyse (survival analysis)*. Die Methoden ähneln sehr denen der *Zuverlässigkeitsanalyse (reliability analysis)* in der *technischen Statistik*.

◇ ◇ ◇

Kapitel 2

Beschreibende Statistik

Jeder statistischen Analyse sollte eine Sichtung des Datenmaterials vorangehen, da ohne sie die Gefahr von Fehlschlüssen erheblich steigt. Geeignete Hilfsmittel stellt die beschreibende Statistik in Form grafischer Datenaufbereitung und durch Angabe von Kenngrößen zur Verfügung. Sie sollen als Grundlage zur Beurteilung der Daten und ihrer Qualität dienen, sowie Anregungen zur Formulierung von Modellen und Hypothesen vermitteln. Zuvor sollen aber noch einige Bemerkungen über Merkmale und zum Häufigkeitsbegriff erörtert werden.

2.1 Populationen, Merkmale & Skalen

2.1.1 Populationen und Stichproben

Eine typische Situation in der Statistik ist, dass verschiedene *Merkmale* (auch Variablen genannt) an mehreren *Objekten* gemessen oder beobachtet werden. Im Beispiel mit den Äpfeln ist jeder Apfel ein Objekt, die gemessene Variable ist das Gewicht des Apfels. Bei der Studentenbefragung sind die Studenten die Objekte, beobachtete Merkmale sind die Absicht, sich selbständig zu machen, Berufe von Mutter und Vater, Schulbildung der Studierenden und Praxiserfahrung.

Die Menge aller denkmöglich vermessbaren bzw. beobachtbaren Objekte bildet die sogenannte *Grundgesamtheit*, manchmal auch *Gesamtpopulation* genannt. Im Beispiel der Studentenbefragung können dies je nach Forschungsfrage z.B. alle Studierenden der BOKU oder auch alle Studierenden irgendeiner Universität in Österreich (deutschsprachigen Ländern, Europa, Asien, ...) sein. Sind Daten für alle Objekte der Grundgesamtheit vorhanden, spricht man von einem sogenannten *Zensus*. Dann müssen die Daten „nur“ noch effizient aufgearbeitet und dargestellt werden, um Informationen aus den Daten zu gewinnen.

In der Praxis ist es aus technischen Gründen meistens weder möglich noch sinnvoll die Merkmale aller Objekte der Grundgesamtheit zu erfassen. Es macht z.B. keinen Sinn das Gewicht aller in Österreich gewachsenen Äpfel einzeln zu erfassen. Man begnügt sich daher mit einer Teilmenge der Grundgesamtheit, einer sogenannten `Stichprobe`. Diese sollte natürlich möglichst „repräsentativ“ für die Grundgesamtheit sein. Wie man genau zu guten Stichproben kommt ist nicht Teil der Grundlagenvorlesung aus Statistik, zum Thema Stichprobentheorie gibt es eigene Lehrbücher. Ein häufig verwendetes und einfaches Prinzip sind sogenannte `Zufallsstichproben`, bei der jedes Objekt der Grundgesamtheit (theoretisch) mit derselben Wahrscheinlichkeit Teil der Stichprobe ist.

2.1.2 Skalentypen

Die beobachteten Merkmale können je nachdem was sie messen unterschiedliche *Ausprägungen* haben. Mit Ausprägungen werden dabei alle für dieses Merkmal möglichen Werte bezeichnet. Mögliche Ausprägungen des Merkmals Geschlecht sind beim Menschen „Mann“ und „Frau“, während das Gewicht eine positive reelle Zahl ist.

Meistens unterscheidet man vier Typen von Skalen und damit verbundenen Merkmalen. Mit wachsender Komplexität sind dies:

Nominalskala: Die Werte bzw. die Ausprägungen, die ein derartig skaliertes Merkmal annehmen kann, sind einfach *Namen*. Beispiele sind etwa 'Obstsorte' (Apfel, Birne, Zwetschke, ...), 'gelesene Tageszeitung' (Krone, Presse, Standard ...) oder 'Studienrichtung' (890, 915, ...). Die Ausprägungen haben keine innere Ordnung: man kann sie zwar z.B. alphabetisch sortieren, dies ist jedoch keine inhaltliche Sortierung. Alphabetisch liegt Birne zwischen Apfel und Zwetschke, ist aber natürlich keine Kreuzung der beiden Obstsorten.

Ordinalskala: Die Ausprägungen sind wieder Namen oder Teilsätze, allerdings besteht unter den Ausprägungen eine Ordnung. Ein Beispiel wären die Bezeichnungen 'kalt', 'lauwarm', und 'heiß' für die Temperatur von Kaffee. Ein wichtiges Merkmal von Ordinalskalen ist, dass Differenzen in der Regel keine Bedeutung haben bzw. die Abstände zwischen benachbarten Ausprägungen nicht gleich sind. ('Sehr gut', 'Gut') und ('Genügend', 'Nicht genügend') sind benachbarte Paare der Ordinalskala Schulnoten. Die numerische Differenz zwischen 1 und 2 bzw. 4 und 5 ist jeweils 1, aber zwischen 4 und 5 liegt auch bestanden vs. nicht bestanden, die tatsächliche Differenz ist daher viel größer.

Intervallskala: Die Intervallskala ist die einfachere der beiden Skalen für numerische Merkmale, also Merkmale die mit Zahlen beschrieben werden. Die Haupteigenschaft der Intervallskala ist dass der Nullpunkt mehr oder weniger willkürlich festgelegt wurde, sodass dem absoluten Wert eines Merkmals weniger Bedeutung zukommt als der Differenz von Werten. Typische Beispiele dafür sind Temperatur (in Celsius oder Fahrenheit) und Datum.

Verhältnisskala: Im Gegensatz zur Intervallskala kommt dem Bezugspunkt hier Bedeutung zu, die absoluten Skalenwerte sind direkt vergleichbar. Beispiele sind Temperatur in ^{o}K, Einkommen oder Fehleranzahl auf Druckseiten.

Merkmale der ersten beiden Skalentypen werden als *kategorielle* oder *qualitative* Merkmale bezeichnet, die der letzten beiden Skalentypen sind *numerische* oder *quantitative* Merkmale.

Manchmal unterscheidet man numerische Merkmale auch in *Zählmerkmale (diskrete Merkmale)* und *Messmerkmale (stetige* oder *kontinuierliche Merkmale)*. Im ersten Fall sind die möglichen Werte einfach ganze Zahlen, im zweiten Fall sind sie Teil eines Kontinuums (z.B. Intervall im eindimensionalen, Fläche im zweidimensionalen Fall).

2.2 Häufigkeiten

Liegt ein beobachteter Datensatz vor, stellt sich oft die Frage nach der Häufigkeit von Werten, Situationen oder allgemein von "Ereignissen". Als Vorgriff auf Kapitel 3 sei bereits an dieser Stelle dieser Begriff eingeführt.

Definition: Alle möglichen *beobachtbaren* Werte im Zuge einer Datenerfassung werden als *Merkmalraum* M bezeichnet. Ein *Ereignis* ist eine Teilmenge $E \subset M$, wobei aber nur solche Teilmengen in Frage kommen, die sich auf sinnvolle Weise mit anderen Ereignissen wieder zu einem Ereignis verbinden lassen. Das System $\mathbf{E}$ aller Ereignisse zu einem gegebenen Merkmalraum M nennt man *Ereignisfeld*. Sinnvolle Verknüpfungen sind

UND:	$E_1, E_2 \in \mathbf{E}$	$\Rightarrow$	"E_1 und E_2"	$\widehat{=}\quad E_1 \cap E_2 \in \mathbf{E}$
ODER:	$E_1, E_2 \in \mathbf{E}$	$\Rightarrow$	"E_1 oder E_2"	$\widehat{=}\quad E_1 \cup E_2 \in \mathbf{E}$
NICHT:	$E_1 \in \mathbf{E}$	$\Rightarrow$	"nicht E_1"	$\widehat{=}\quad E_1^c \in \mathbf{E}$.

Besondere Ereignisse sind die *Elementarereignisse* $E = \{x\} \in \mathbf{E}$ für $x \in M$, das *sichere Ereignis* $E = M \in \mathbf{E}$ und das Gegenteil davon, das *unmögliche Ereignis* $E = \emptyset \in \mathbf{E}$.

Beispiel 2.1 In Beispiel 1.1 bietet sich $M = [10, 20]$ als denkbarer Merkmalraum an. Das Ereignis (die Situation) *"Gewicht unter 15 dag"* wird durch das Intervall $[10, 15)$ beschrieben.

◇◇◇

Beispiel 2.2 In Beispiel 1.2 ist bei Frage 1 der Merkmalraum sinnvollerweise $M = \{1, 2, 3, 4, 5\}$ zu wählen. Das Ereignis (die Situation) *"positiv zu selbständigem Beruf eingestellt"* wäre dann durch $\{1, 2\}$ zu beschreiben.

◇◇◇

Definition: Es liegt ein Datensatz $x_1, x_2, \ldots, x_n$ vom Umfang n vor. Unter der *absoluten Häufigkeit* $h_a(E)$ eines bestimmten Ereignisses E bei diesen n Beobachtungswerten versteht man einfach die Anzahl von Werten, bei denen das Ereignis eingetreten ist oder für die die Situation zutrifft. Formal lautet sie daher:

$$h_a(E) = \#\{x_i : \ x_i \in E\} \ .$$

Da die so definierte Häufigkeit unmittelbar vom Datenumfang n abhängt und damit für Vergleichszwecke ungeeignet ist, wird meistens die *relative Häufigkeit* $h_r(E)$ oder die *prozentuelle Häufigkeit* $h_p(E)$ verwendet, für die gilt:

$$\begin{aligned} h_r(E) &= h_a(E)/n \\ h_p(E) &= 100\, h_a(E)/n\,\% \end{aligned}$$

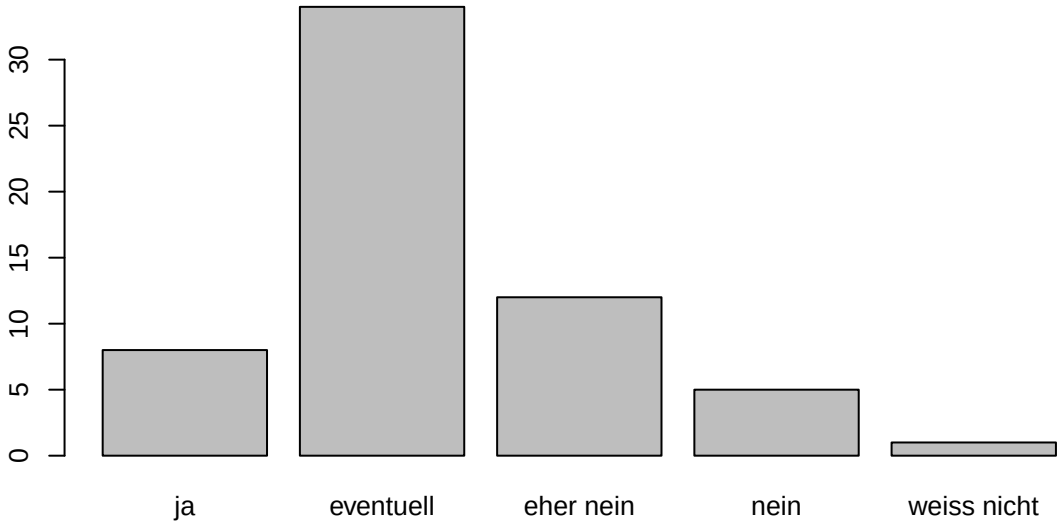
2.3 Grafische Datenaufbereitung

Die klassischen Darstellungsformen für Häufigkeitsverteilungen von Daten (Stichproben) sind das *Balken-* und *Kreisdiagramm* für qualtitative Merkmale, das *Stab–* oder *Säulendiagramm* für diskrete Größen bzw. das *Histogramm* für kontinuierliche Merkmale.

2.3.1 Balkendiagramm

Die einfachste und am häufigsten verwendete Visualisierung von kategoriellen Merkmalen sind Balkendiagramme, wobei die Höhe der Balken den absoluten oder relativen Häufigkeiten der einzelnen Balken entspricht. Betrachtet man das Merkmal „Absicht selbständig machen“ aus Studierendenbefragung von Bsp. 1.2 ergibt sich Abb. 2.1.

Abbildung 2.1: *Balkendiagramm*



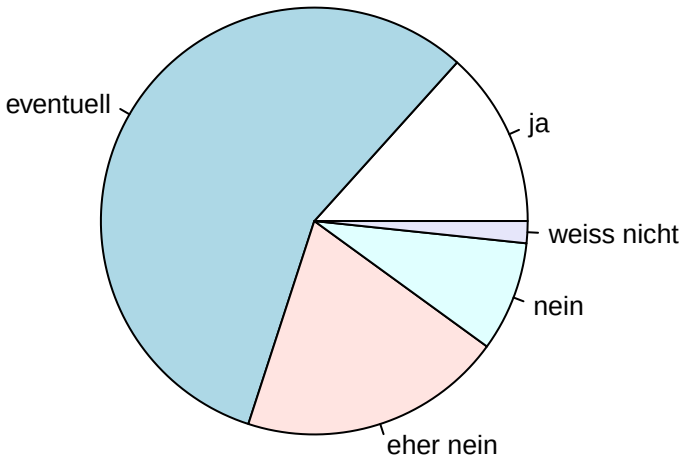
2.3.2 Kreisdiagramm

In Medien und Geschäftspräsentationen sind auch Kreis- oder Tortendiagramme in verschiedensten Variationen sehr beliebt. Hier werden die Anteile jeder Ausprägung als Kreissegment dargestellt. Aus wissenschaftlicher Sicht sind Kreisdiagramme ein sehr schlechter Weg um Daten zu visualisieren. Das menschliche Auge kann lineare Maße recht gut schätzen, ist aber schlecht beim Schätzen von relativen Flächen oder Winkeln. Ein Balkendiagramm ist fast immer zu bevorzugen, eine der wenigen Ausnahmen ist die Darstellung absoluter Mehrheiten, d.h., ob ein Kreissegment oder die Summe mehrerer benachbarter Segmente mehr als 50% der Fläche einnimmt. Im Beispiel sieht man im Kreisdiagramm, dass mehr als die Hälfte der Studierenden einen selbständigen Beruf „eventuell“ in Betracht zieht.

2.3.3 Strichliste

Oft sind Ereignisse einfach Klassen, die die Form von Intervallen besitzen. Häufigkeiten dafür wurden traditionell über *Strichlisten* ermittelt, und obwohl heute natürlich Computer die

Abbildung 2.2: *Kreisdiagramm*



Auszählung für uns übernehmen, ist es für das Verständnis doch oft hilfreich, kurz über Strichlisten nachzudenken. Dabei werden die Werte des Datensatzes den einzelnen Klassen gemäß Tab. 2.3 zugeordnet und meist durch einen Strich markiert. Um das Zählen zu erleichtern und auch um einen optischen Eindruck von der Häufigkeitsverteilung zu gewinnen, werden häufig Gruppen – meist Fünfergruppen – gebildet, indem z.B. die jeweils fünfte Beobachtung in einer Gruppe von fünf Werten die vier vorangegangenen waagrecht verbindet.

Teilt man den Datensatz aus Bsp. 1.1 so in 8 Klassen, wie man ihn für die Erstellung eines Histogramms benötigt, erhält man eine Strichliste, die Abb. 2.3 zeigt.

Abbildung 2.3: *Strichliste*

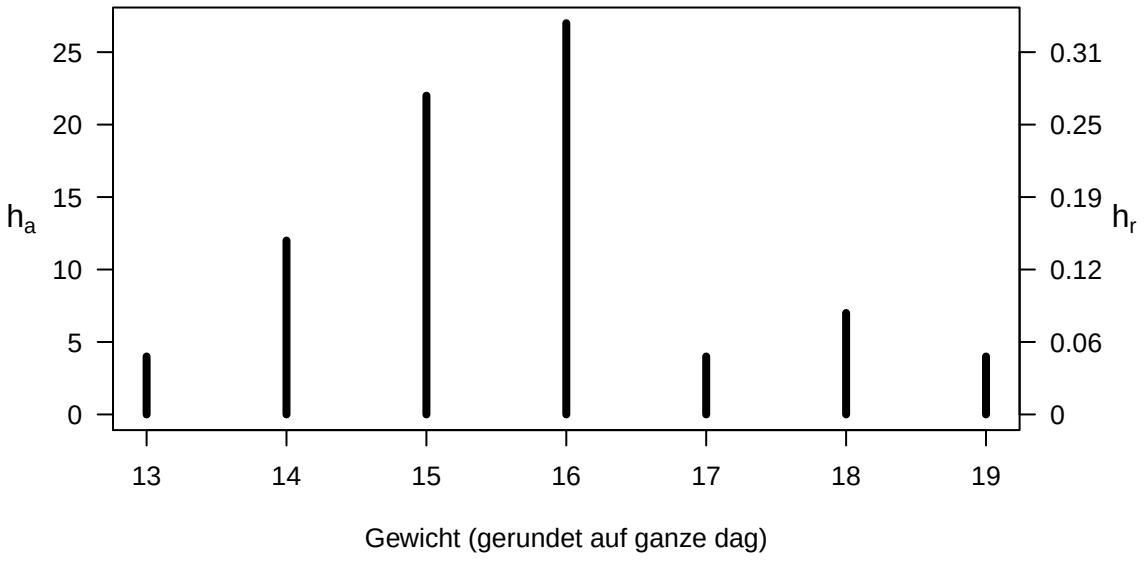
Klasse	Häufigkeit	h_a
13.1 – 13.9	𝍸 𝍸	10
13.9 – 14.7	𝍸 𝍸 ||	12
14.7 – 15.5	𝍸 𝍸 𝍸 |||	18
15.5 – 16.3	𝍸 𝍸 𝍸 𝍸 ||	22
16.3 – 17.1	𝍸	5
17.1 – 17.9	𝍸 |	6
17.9 – 18.7	||||	4
18.7 – 19.5	|||	3
Summe		80

◇◇◇

2.3.4 Stabdiagramm

Über jedem Wert aus dem Wertevorrat eines diskreten numerischen Merkmals wird die (absolute, relative, prozentuelle) Häufigkeit seines Auftretens im Datensatz aufgetragen. Es stellt eine Schätzung für die theoretische *Wahrscheinlichkeitsfunktion* einer Zufallsgröße dar (siehe Abschnitt 3.3). Rundet man das Gewicht der Äpfel aus Bsp. 1.1 auf ganze *dag* erhält man diskrete Daten. Abb. 2.4 zeigt das zugehörige Stabdiagramm, wobei h_a absolute und h_r relative Häufigkeiten bezeichnet.

Abbildung 2.4: *Stabdiagramm*



2.3.5 Histogramm

Erster — und sensibelster — Schritt bei der Histogrammerstellung ist die *Klassierung* des Wertebereiches einer kontinuierlichen Größe mit entsprechender Aufteilung der betrachteten Daten. Anzahl, Lage und Breite der Klassen lassen ein großes Ausmaß an Subjektivität zu. Sinnvoll wird eine Klassierung erst ab einem gewissen Mindeststichprobenumfang. Ein üblicher Vorschlag geht von einer Mindestanzahl von 30 Daten aus.

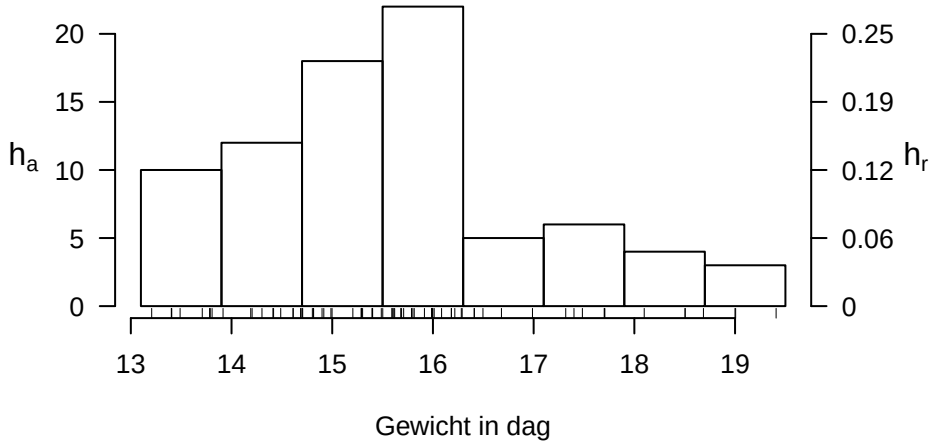
Als Faustregel für die Klassenanzahl k bei Stichproben vom Umfang n findet man häufig

$$\sqrt[3]{n} \leq k \leq \sqrt[2]{n} \quad ,$$

wobei man mit wachsendem n von der oberen Grenze weg zur unteren tendiert (z.B.: $n = 40/k = 6$, $n = 500/k = 15$, $n = 10\,000/k = 25$). Das Minimum und das Maximum der Beobachtungswerte sollten nicht als Grenzen der äußeren Klassen fungieren, sondern im Inneren dieser Klassen liegen. Üblicherweise wählt man gleich breite Klassen, es sind aber auch unterschiedlich breite Klassen zulässig, soferne man sich an die eigentliche Definition des Histogramms hält.

Das Histogramm stellt das empirische Gegenstück zur Dichtefunktion der zugrundeliegenden theoretischen Wahrscheinlichkeitsverteilung des beobachteten Merkmals dar (siehe

Abbildung 2.5: *Histogramm 1*



Abschnitt 3.3). Daher ergibt sich der Funktionswert des Histogramms über einer Klasse, also die Höhe des über einer Klasse aufzutragenden Rechtecks, derart, dass der *Flächeninhalt* dieses Rechtecks der relativen bzw. prozentuellen Häufigkeit von Daten in der betrachteten Klasse entspricht. Bei *gleicher* Klassenbreite ist diese Höhe klarerweise proportional zum Flächeninhalt und damit zur Häufigkeit. Da Softwarepakete i. Allg. automatisch gleich breite Klassen festlegen, findet man speziell dort häufig die Ordinate eines Histogramms mit der Häufigkeit skaliert. Die tatsächliche Bedeutung sollte aber bewusst bleiben.

Ein Histogramm mit gleichen Klassenbreiten für die Daten aus Bsp. 1.1 wäre demnach folgendermaßen zu erstellen:

$n = 80 \quad \Rightarrow \quad k = 8$ (oder 9)
Minimum: 13.2 Maximum: 19.4
Klassierung: $19.4 - 13.2 = 6.2 \quad \Rightarrow$ Klassenbreite: $0.8\ dag = 8\ g$
13.1, 13.9, 14.7, 15.5, 16.3, 17.1, 17.9, 18.7, 19.5

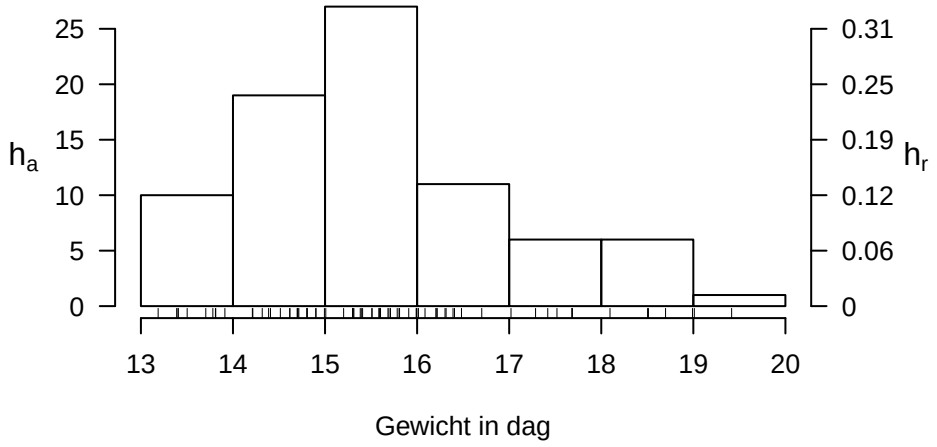
Die angeführten Werte begrenzen bzw. trennen die Klassen, wobei die rechte Grenze vereinbarungsgemäß dazugehören soll. Abb. 2.5 zeigt dieses Histogramm, wobei wie im Stabdiagramm in Abb. 2.4 sowohl absolute wie auch relative Häufigkeiten h_a und h_r markiert sind.

Die Form des Histogramms kann mitunter stark von den gewählten Klassenbreiten und -grenzen abhängen. Abb. 2.6 zeigt ein Histogramm derselben Daten, allerdings mit nur 7 Klassen und ganzzahligen Klassengrenzen. Während im ersten Histogramm der Sprung nach unten rechts von 16 sehr ausgeprägt ist, ist dies im zweiten Histogramm weniger stark ausgeprägt. Die kleinen vertikalen Striche unterhalb des zweiten Histogramms markieren die Beobachtungen (und erklären auch die unterschiedlichen Sprunghöhen).

2.3.6 Box–Plot

Eine einfache und deshalb auch sehr beliebte Darstellung von Daten ist in Form des *Box–Plots* gegeben (vgl. Abb. 2.7). Die Box wird dabei von den beiden Stichprobenquartilen gebildet und durch den Stichprobenmedian (siehe jeweils Abschnitt 2.4) getrennt. Das heißt, die Box überdeckt die mittlere Hälfte der Daten, links und rechts (bzw. ober und unter bei vertikalen Box-Plots) der Box liegen jeweils ein Viertel der Daten. Nach Tukey besitzen die seitlichen *Bänder (whiskers)* jeweils eine Länge von maximal dem 1,5–Fachen des Interquartilsabstandes,

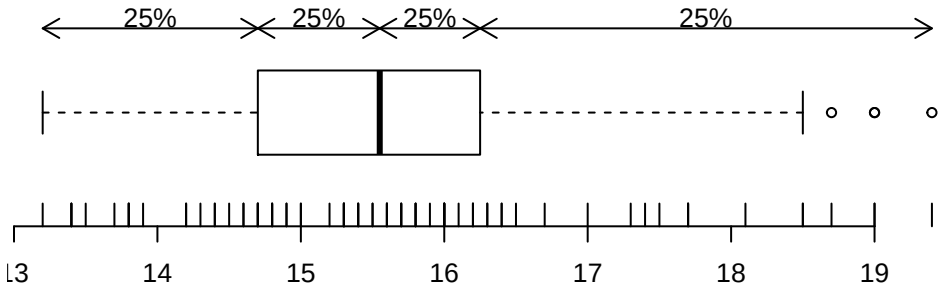
Abbildung 2.6: *Histogramm 2*



reichen aber nicht weiter als bis zum ersten bzw. letzten Wert und enden in jedem Fall bei einem beobachteten Wert. Diese Definition ist aber durchaus nicht die einzige. Box-Plots sind hervorragend für den Gruppenvergleich bzgl. eines betrachteten Merkmals geeignet, weil sie in knapper und übersichtlicher Form Lage, Konzentration, Streuverhalten und Symmetrie einer (Häufigkeits–) Verteilung beschreiben.

Der zum Bsp. 1.1 gehörende Box-Plot besitzt die in Abb. 2.7 ersichtliche Form.

Abbildung 2.7: *Box-Plot*



◇◇◇

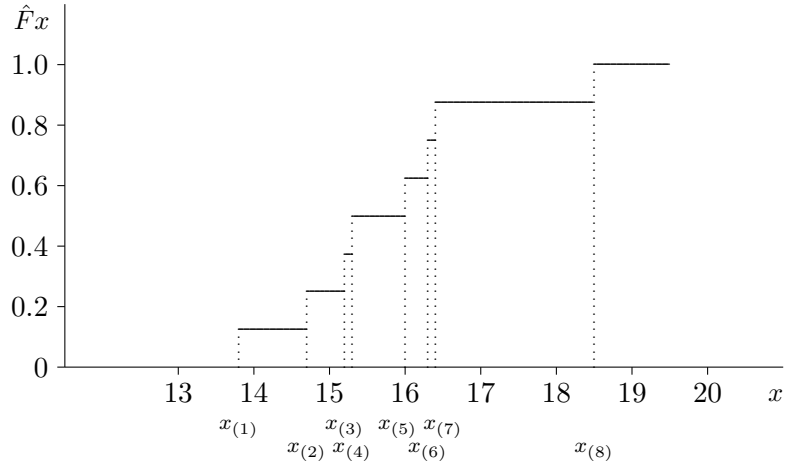
2.3.7 Empirische Verteilungsfunktion

Das Gegenstück zur theoretischen Verteilungsfunktion *(VF)* (siehe Abschnitt 3.3) eines Merkmals X bildet die *empirische Verteilungsfunktion (eVF)* für Stichproben. Sie ist als

$$\hat{F}_n(x_1, \ldots, x_n;\, x) := \frac{1}{n} \sum_{i=1}^{n} \mathrm{I}_{(-\infty, x]}(x_i) \qquad \text{für } x \in \boldsymbol{R},$$

definiert, also den Anteil der Stichprobe $x_1, \ldots, x_n$, der kleiner oder gleich dem Argument $x \in \boldsymbol{R}$ ist. Das ergibt den typisch treppenartigen Verlauf, wie ihn Abb. 2.8 für die Daten aus Bsp. 1.1 zeigt.

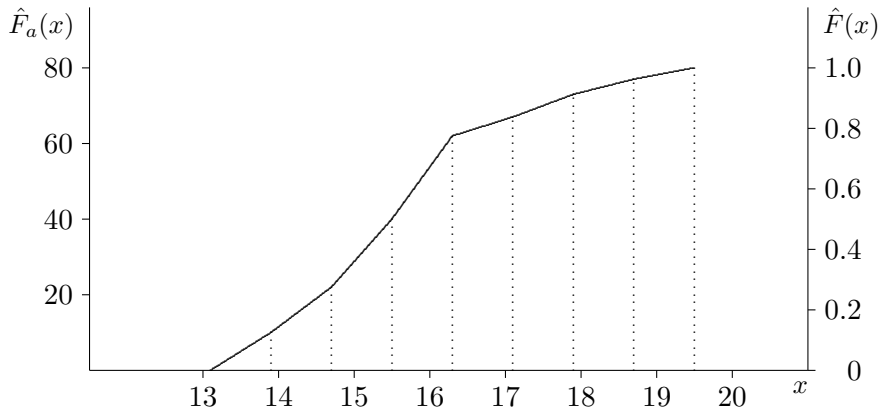
Abbildung 2.8: *empirische Verteilungsfunktion*



2.3.8 Summenpolygon

Für klassierte Daten erhält man meist eine abgewandelte Form der *eVF*, indem nur an den oberen Klassengrenzen die Häufigkeit von Daten betrachtet wird, die kleiner oder gleich eben dieser Grenze sind. Der betrachtete Funktionswert ergibt sich somit als Summe der Häufigkeiten der Klassen, die unterhalb dieser Grenze liegen. Zwischen diesen Gitterwerten wird linear interpoliert, d.h. die Datenhäufigkeit "gleichmäßig" auf die jeweiligen Klassen "verteilt". Dadurch erklärt sich die Bezeichnung *Summenpolygon* für diese modifizierte *eVF*. Für die Daten aus Bsp. 1.1 zeigt Abb. 2.9 das nach dem Histogramm aus Abb. 2.5 erstellte Summenpolygon.

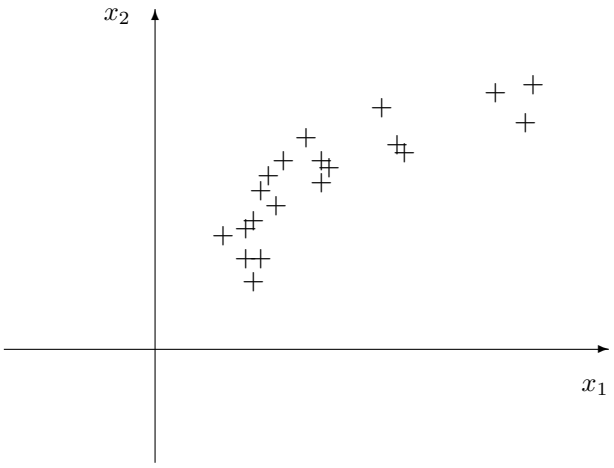
Abbildung 2.9: *Summenpolygon*



2.3.9 Streudiagramm

Für zweidimensionale Daten zu stetigen Merkmalen ist das sogenannte *Streudiagramm (scatter plot)* zur grafischen Aufbereitung geeignet. Es stellt eigentlich ein übliches Koordinatensystem dar, in dem die Datenpunkte (allenfalls bezeichnet) eingetragen sind (vgl. Abb. 2.10).

Abbildung 2.10: *Streudiagramm*



2.4 Kenngrößen

Neben Grafiken wie Histogrammen zur Beschreibung von Verteilungen sind oft auch numerische Beschreibungen notwendig, sogenannte *statistische Kennzahlen*. Insbesonders bei großen Datensätzen mit sehr vielen Merkmalen können nicht Grafiken aller Merkmale betrachtet werden. Die absoluten oder relativen Häufigkeiten der Ausprägungen codieren die Stichprobe ohne Informationsverlust (bis auf die Reihenfolge). Für kategorische Größen mit wenigen Kategorien sind die Häufigkeiten der Ausprägungen auch die üblichste numerische Beschreibung der Daten. Bei numerischen Merkmalen gibt es in der Regel viele verschiedene Ausprägungen, eine Möglichkeit der Kompression ist die Klassierung der Daten. Es gibt jedoch noch eine Reihe weiterer Kennzahlen für numerische Größen.

2.4.1 Ränge und Quantile

Sortiert man die Stichprobe der Größe nach

$$x_{(1)} \leq x_{(2)} \leq \ldots \leq x_{(n-1)} \leq x_{(n)}$$

erhält man die sogenannten *Rangstatistiken* $x_{(i)}$ (auch kurz Ränge genannt). Der kleinste beobachtete Wert $x_{(1)}$ ist das Minimum der Stichprobe, der größte $x_{(n)}$ das Maximum.

Die *Quantile* ergeben sich als die Umkehrfunktion der empirischen Verteilungsfunktion

$$q_\alpha = \hat{F}_n^{-1}(\alpha) \quad \Leftrightarrow \quad \frac{\#\{x_i \leq q_\alpha\}}{n} = \alpha \in [0,1]$$

Die Quantile können dazu benutzt werden um Fragen vom Typ „Kleiner als welcher Wert sind ein Fünftel der Stichprobe?“ zu beantworten. In diesem Beispiel wäre $q_{0.2}$ die Antwort.

Die allgemeine Idee zur Definition des α-Quantils q_α $(0 < \alpha < 1)$ ist, dass es die Daten so in zwei Teile trennt, dass ungefähr $\alpha \cdot 100\%$ der Daten links von q_α und $(1-\alpha) \cdot 100\%$ der Daten rechts von q_α liegen. Es muss also gelten

$$\begin{aligned} q_\alpha &= x_{([n\alpha]+1)}, && \text{wenn } n\alpha \text{ nicht ganzzahlig,} \\ q_\alpha &\in [x_{(n\alpha)}, x_{(n\alpha+1)}], && \text{wenn } n\alpha \text{ ganzzahlig.} \end{aligned}$$

Dabei ist $[n\alpha]$ die zu $n\alpha$ nächste kleinere ganze Zahl. Für ganzzahliges $n\alpha$ wird oft einfach die Intervallmitte $(x_{(n\alpha)} + x_{(n\alpha+1)})/2$ genommen. Obige Quantilsdefinition ist gut geeignet zur „manuellen“ Berechnung von Quantilen da sie sehr einfach ist. Ein Nachteil obiger Formel ist, dass sie wie die empirische Verteilungsfunktion eine Treppenfunktion mit Sprungstellen in den Beobachtungen $x_{(i)}$ darstellt, also nicht stetig ist. Viele Softwarepakete (inklusive R) interpolieren daher die empirische Verteilungsfunktion zur Berechnung von Quantilen linear, um eine stetige Abbildung von α auf q_α zu erhalten, vergleiche auch Abb. 2.8 und 2.9.

Spezielle Quantile die oft verwendet werden sind

- Die *Quartile* $q_{0.25}$ (1. Quartil), $q_{0.5}$ (2. Quartil, Median), und $q_{0.75}$ (3. Quartil), die den Datensatz in vier Viertel mit jeweils gleich vielen Beobachtungen zerteilen.
- Die *Dezile* $q_{0.1}$, $q_{0.2}$, $\ldots$ $q_{0.9}$ teilen den Datensatz in zehn gleich große Teilmengen.

Weniger gebräuchlich sind *Terzile* (unteres, mittleres und oberes Drittel) und *Quintile* (fünf Teilmengen). Ränge und Quantile setzen voraus, dass der Datensatz der Größe nach geordnet werden kann, machen also erst ab einer Ordinalskala Sinn. Für nominalskalierte Merkmale sind Ränge und Quantile nicht sinnvoll berechenbar.

Beispiel 2.3 Für die Daten aus Bsp. 1.1 mit $n = 80$ erhält man das 20%-Quantil wegen $n\alpha = 16$ als

$$q_{0.20} = (x_{(16)} + x_{(17)})/2 = (14.5 + 14.6)/2 = 14.55$$

und das 33%-Quantil wegen $n\alpha = 26.4$ als

$$q_{0.33} = x_{(27)} = 14.9$$

◇◇◇

2.4.2 Lagemaße

Die bekanntesten Kenngrößen eines Datensatzes zur Lagebeschreibung eines kontinuierlichen Merkmals sind wohl der *arithmetische Mittelwert*, bekannt als *Stichprobenmittelwert*,

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i \tag{2.1}$$

und der *Stichprobenmedian*

$$\tilde{x} = q_{0.5} = \begin{cases} x_{((n+1)/2)} & \text{falls } n \text{ ungerade} \\ (x_{(n/2)} + x_{(n/2+1)})/2 & \text{falls } n \text{ gerade.} \end{cases} \tag{2.2}$$

Ein weiteres Lagemaß ist der *Modalwert* (oder kurz: *Modus*) eines Datensatzes. Er wird von Statistik-Programmen seltener angegeben und bedeutet in Übertragung seiner Definition bei theoretischen Verteilungen einfach "häufigster Wert". Da dieser bei Stichproben selten eindeutig ist, gibt es meist das Ergebnis „not unique". Eine alternative Definition ist bei klassierten Stichproben möglich, wo man den Modalwert als repräsentativen Wert (etwa das Zentrum) der häufigsten Klasse oder der häufigsten Klassen (die liegen dann meist benachbart und das Zentrum definiert sich über den gemeinsamen Bereich dieser Klassen) festlegt.

Beispiel 2.4 Für die Daten aus Bsp. 1.1 erhält man

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{1}{80}\,1252.6 = 15.67$$

und

$$\tilde{x} = (x_{(40)} + x_{(41)})/2 = (15.5 + 15.6)/2 = 15.55\ .$$

Die modale Klasse im Histogramm in Abb. 2.5 ist das Intervall $(15.5, 16.3]$, sodass man als Modalwert 15.9 erhält.

◇◇◇

Die Verwendung von $\tilde{x}$ hat nicht nur wegen der einfachen Berechnung seine Berechtigung, sondern vor allem dann, wenn Verdacht auf *Ausreißer* (abliegende Werte, untypische Werte) besteht. Während bei der Mittelbildung jeder Wert, somit auch jeder ausreißerverdächtige Wert mit dem gleichen Gewicht $1/n$ eingeht, tragen letztere zum Stichprobenmedian praktisch nichts bei. Daher stellt dieser ein *robustes* Verfahren zur Schätzung des theoretischen Mittelwertes bei symmetrischen Verteilungen dar.

2.4.3 Streuungsmaße

Traditionell ist die *Stichproben–Standardabweichung (sample standard deviation, SSD)*

$$s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2} \tag{2.3}$$

bei Analysen zur Streuungsbeschreibung immer angegeben. Dabei ist gerade hier die Empfindlichkeit gegenüber Ausreißern ziemlich groß (*quadrierte* Abstände!). Häufig wird sie in der numerisch bedenklichen Form

$$s = \sqrt{\frac{1}{n-1}\Big(\sum_{i=1}^{n} x_i^2 - n\,\bar{x}^2\Big)} \tag{2.4}$$

ausgewertet, bei der durch die Subtraktion zweier möglicherweise großer Werte von annähernd gleicher Größenordnung beachtliche Auslöschungseffekte in den letzten, aber schließlich signifikanten Stellen erfolgen können. Bei manuellen Berechnungen mit kleinen Stichproben wie in Übungenbeispielen ist Formel 2.4 aber oft einfacher und der Einsatz unbedenklich, da meist ohnehin nur auf wenige Nachkommastellen genau gerechnet wird. Analoges gilt für das Quadrat s^2, die sogenannte *Stichprobenvarianz*.

Für den Fall, dass $|\bar{x}| \gg 0$ gilt, wird gerne die 'normierte' Stichprobenstandardabweichung

$$\text{cv} = \frac{s}{\bar{x}} \ ,$$

der sogenannte *Variationskoeffizient (*engl. *coefficient of variation)*, zur Beschreibung der *relativen* Streuung herangezogen.

Ein sehr robustes Maß liegt in Form des *Medmed* vor, der als *Median* der absoluten Abweichungen vom Stichproben–*Median* erklärt ist (auch MAD, *mean absolute deviation*):

$$Medmed = \text{med}_{i=1(1)n} \{|x_i - \tilde{x}|\} \quad . \tag{2.5}$$

Durch die innere Medianbildung wird die Lage relativ einflusslos geschätzt und bei der äußeren Medianbildung bleiben die großen Abweichungen unberücksichtigt.

Eine weitere sehr robuste Streuungskenngröße ist durch den *Interquartilsabstand (interquartile range, IQR)*

$$IQR = q_{0.75} - q_{0.25} \tag{2.6}$$

gegeben, wobei $q_{0.25}$ und $q_{0.75}$ wieder das 1. bzw. 3. Quartil darstellen. Im Box-Plot entspricht der IQR der Größe der Box.

Schließlich soll noch ein Streuungsmaß erwähnt werden, das in der Praxis leider immer noch sehr verbreitetet ist, von dessen Gebrauch man aber eher abraten muss. Die sogenannte *Spannweite* (auch Spannbreite, engl. *range*)

$$R = x_{(n)} - x_{(1)} \tag{2.7}$$

beschreibt den Gesamtstreubereich der Daten und ist von den beiden Extremwerten stark abhängig, sodass sie das glatte Gegenteil von robust ist. Die Beliebtheit dieser Maßzahl rührt aus Zeiten, wo die einfache Berechnung im Vordergrund stand; in Zeiten des Computers ist dieses Argument aber überholt.

Beispiel 2.5 Für die Daten aus Bsp. 1.1 erhält man zunächst

$$\begin{aligned} s^2 &= \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 = \frac{1}{n-1}\left(\sum_{i=1}^{n} x_i^2 - 2\bar{x}\sum_{i=1}^{n} x_i + n\bar{x}^2\right) \\ &= \frac{1}{n-1}\left(\sum_{i=1}^{n} x_i^2 - n\bar{x}^2\right) = \frac{1}{n-1}\left(\sum_{i=1}^{n} x_i^2 - \frac{1}{n}(\sum_{i=1}^{n} x_i)^2\right) \\ &= \frac{1}{79}\left(19774.36 - \frac{1}{80}1252.6^2\right) \\ &= 2.0478 \end{aligned}$$

und damit $s = 1.43$. Der *MAD*–Wert ergibt sich als

$$MAD = \frac{1}{n}\sum_{i=1}^{n} |x_i - \tilde{x}| = \frac{86.6}{80} = 1.08 \ ,$$

und der *IQR*–Wert wegen

$$\begin{aligned} q_{0.25} &= (x_{(20)} + x_{(21)})/2 = (14.7 + 14.7)/2 = 14.7 \\ q_{0.75} &= (x_{(60)} + x_{(61)})/2 = (16.2 + 16.3)/2 = 16.25 \end{aligned}$$

zu

$$IQR = q_{0.75} - q_{0.25} = 16.25 - 14.7 = 1.55 \ .$$

Der Vollständigkeit halber sei auch noch die Spannweite berechnet:

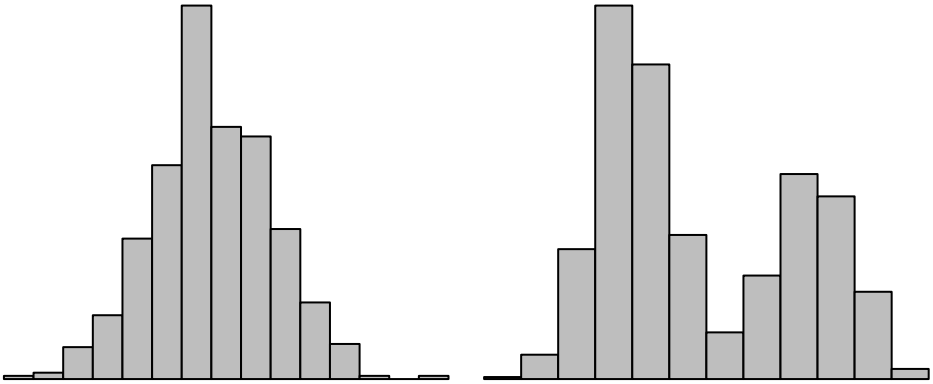
$$R = x_{(80)} - x_{(1)} = 19.4 - 13.2 = 6.2$$

◇◇◇

2.4.4 Beurteilung der Gestalt

Die *Gestalt* einer Verteilung lässt sich am einfachsten durch Grafiken beurteilen, siehe Abbildungen 2.4.4 und 2.4.4. Zeigt das Histogramm nur ein eindeutiges Maximum spricht man von einer *unimodalen* Verteilung, bei zwei lokalen Maxima von `bimodal`, bei mehreren lokalen Maxima von *multimodal*. Die Anzahl der Klassen und Klassenränder können einen starken Einfluss auf die Anzahl der lokalen Maxima eines Histogramms haben, daher ist ein wenig Vorsicht geboten und im Zweifel sollten auch andere Darstellungen der Daten herangezogen werden (Streudiagramm etc.).

Abbildung 2.11: *Uni- und bimodale Verteilung*



Bei unimodalen Verteilungen unterscheidet man üblicherweise zwischen symmetrischen und schiefen Verteilungen. Zur numerischen Beurteilung der Schiefe kann das sogenannte dritte Stichprobenmoment

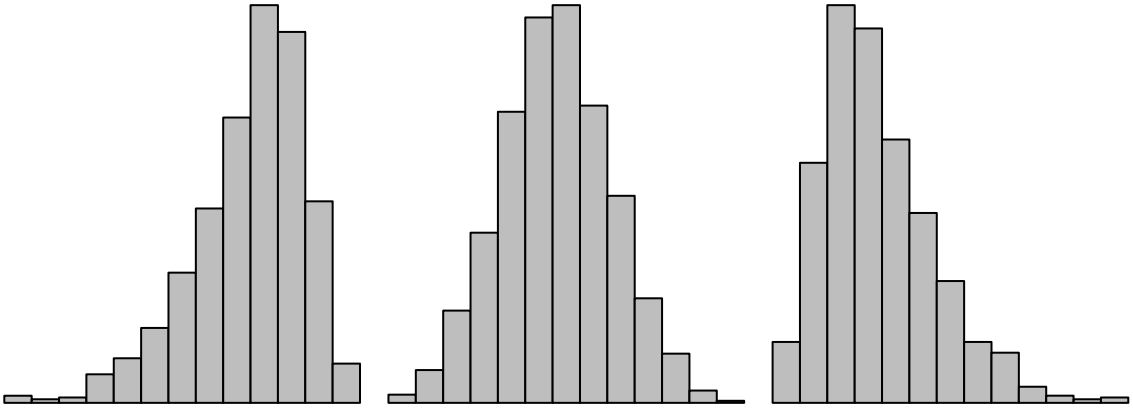
$$\hat{\gamma}_1 = \frac{1}{n} \sum_{i=1}^{n} \left(\frac{x_i - \bar{x}}{s} \right)^3 \quad . \tag{2.8}$$

verwendet werden, dieses ist jedoch noch ausreißeranfälliger als die Standardabweichung (Abstände hoch drei).

Es gilt folgende Zuordnung:

$$\gamma_1 \left\{ \begin{array}{c} < \\ \approx \\ > \end{array} \right\} 0 \;\Rightarrow\; \text{Verteilung ist} \left\{ \begin{array}{l} \text{rechtssteil} = \text{linksschief} \\ \text{symmetrisch} \\ \text{linkssteil} = \text{rechtsschief.} \end{array} \right.$$

Abbildung 2.12: *Linksschiefe, symmetrische und rechtsschiefe Verteilung*



Beispiel 2.6 Für die Daten aus Bsp. 1.1 erhält man

$$\begin{aligned} \hat{\gamma}_1 &= \frac{1}{n}\sum_{i=1}^{n}\left(\frac{x_i-\bar{x}}{s_*}\right)^3 \\ &= \frac{1}{n}\left(\sum_{i=1}^{n}x_i^3 - \frac{3}{n}(\sum_{i=1}^{n}x_i^2)(\sum_{i=1}^{n}x_i) + \frac{2}{n^2}(\sum_{i=1}^{n}x_i)^3\right)/s_*^3 \\ &= \frac{1}{80}\left(314831.278 - \frac{3}{80}\,19774.36\times 1252.6 + \frac{2}{80^2}\,1252.6^3\right)/1.42^3 \\ &= 0.644\ , \end{aligned}$$

wobei zu beachten ist, dass hier aus Konsistenzgründen s durch s_* ersetzt ist, das über eine Division durch n anstelle durch $n-1$ in der Formel (2.3) gewonnen wurde. Die positive Schiefe passt zur Form der Histogramme in Abbildungen 2.5 und 2.6.

◇◇◇

Kapitel 3

Wahrscheinlichkeitsrechnung

Dieser Abschnitt dient der Modellbildung *zufallsbeeinflusster* oder *unsicherheitsbehafteter* Phänomene. Wie im Einführungsabschnitt bereits angedeutet, ist dieser Zweig der Stochastik eher mathematisch orientiert. Er kann als Wahrscheinlichkeitstheorie auch tatsächlich völlig unabhängig von Anwendungsmöglichkeiten als eigenständige mathematische Disziplin betrieben werden. Hier sollen aber die Modelle und Rechenregeln nur soweit diskutiert werden, wie es für die Behandlung statistischer Probleme notwendig ist.

3.1 Wahrscheinlichkeit

3.1.1 Definition, Grundregeln

Die Ausgangssituation ist ein Versuch, der – zumindest gedanklich – beliebig wiederholbar ist. Dem Versuch ist ein Merkmalraum M und ein Ereignisfeld $\mathbf{E}$ zugeordnet (vgl. Abschnitt 2.2). Führt man den Versuch n–mal aus, so zeigt die relative Häufigkeit $h_r(E)$ eines betrachteten Ereignisses $E \in \mathbf{E}$ mit wachsender Wiederholungszahl n ein konvergenzähnliches Verhalten *(empirisches Gesetz der großen Zahlen)*. Außerdem gilt, wie man sich leicht anschaulich überlegt,

H1) $0 \leq h_r(E) \leq 1$ für alle $E \in \mathbf{E}$,

H2) $h_r(\emptyset) = 0$ und $h_r(M) = 1$,

H3) $E_1, E_2 \in \mathbf{E}$ und $E_1 \cap E_2 = \emptyset$ (E_1 und E_2 "schließen einander aus")
$\Rightarrow\ h_r(E_1 \cup E_2) = h_r(E_1) + h_r(E_2)$ *(Additivität)*

Das motiviert zur (axiomatischen) Einführung eines von n unabhängigen Begriffes zur Beschreibung der Eintrittschance eines Ereignisses E in folgender

Definition: Unter einer *Wahrscheinlichkeitsverteilung (*engl. *probability distribution)* P versteht man die Zuordnung einer *Wahrscheinlichkeit (*engl. *probability)* $P(E)$ für jedes Ereignis $E \in \mathbf{E}$, die folgenden Bedingungen genügen muss :

P1) $0 \leq P(E) \leq 1$ für alle $E \in \mathbf{E}$,

P2) $P(\emptyset) = 0$ und $P(M) = 1$,

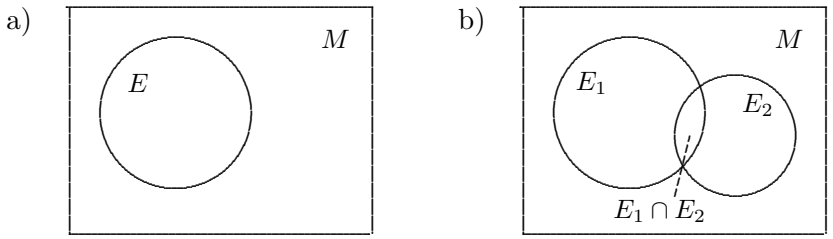
P3) $\left.\begin{array}{l} E_1, E_2, \ldots \in \mathbf{E} \\ E_i \cap E_j = \emptyset \quad \text{für alle } i \neq j \end{array}\right\} \Rightarrow P(E_1 \cup E_2 \cup \ldots) = P(\bigcup_{i=1}^{\infty} E_i) = \sum_{i=1}^{\infty} P(E_i)$

(σ–Additivität, abzählbare Additivität)

Das Tripel $(M, \mathbf{E}, P)$ heißt dann *Wahrscheinlichkeitsraum (*engl. *probability space)* (W–Raum).

Die Rechenregeln für Wahrscheinlichkeiten lassen sich sehr leicht veranschaulichen, wenn man als Merkmalraum einen Teil (meistens ein Rechteck) der Ebene auszeichnet (Abb.3.1). Ereignisse sind dann einfach Teilflächen und die Wahrscheinlichkeit wird als Flächenanteil des Ereignisses am gesamten Merkmalraum aufgefasst (*geometrische* Wahrscheinlichkeitsinterpretation).

Abbildung 3.1: *Geometrische Wahrscheinlichkeit*



Es gelten folgende einfache Rechenregeln:

1) "Gegenwahrscheinlichkeit":

$$P(E^c) = 1 - P(E) \tag{3.1}$$

2) "Additionsregel"

a)
$$P(E_1 \cup E_2) = P(E_1) + P(E_2) - P(E_1 \cap E_2) \tag{3.2}$$

b)
$$\begin{aligned} P(E_1 \cup E_2 \cup \ldots \cup E_k) &= \\ &= P(E_1) + P(E_2) + \ldots + P(E_k) \\ &\quad - P(E_1 \cap E_2) - P(E_1 \cap E_3) - \ldots - P(E_{k-1} \cap E_k) \\ &\quad + P(E_1 \cap E_2 \cap E_3) + \ldots + P(E_{k-2} \cap E_{k-1} \cap E_k) \\ &\quad \ldots \\ &= \sum_{l=1}^{k} (-1)^{l+1} \sum_{\{i_1,\ldots,i_l\} \subset \{1,\ldots,k\}} P(E_{i_1} \cap E_{i_2} \cap \ldots \cap E_{i_l}) \end{aligned} \tag{3.3}$$

Beweis:

1) Wegen $E \cap E^c = \emptyset$, gilt auf Grund der Additivität (P3)

$$1 = P(M) = P(E \cup E^c) = P(E) + P(E^c) \ ,$$

also die behauptete Beziehung.

2) Anschaulich entnimmt man der Skizze in Abb.3.1, dass sich die Gesamtfläche für $E_1 \cup E_2$ als Summe der Einzelflächen für E_1 bzw. E_2 ergibt, wobei aber der offensichtlich doppelt belegte Anteil des "Zwickels" von E_1 und E_2 $(= E_1 \cap E_2)$ einmal abgezogen werden muss.

Formal gilt zunächst

$$E_1 \cup E_2 = (E_1 - E_2) \cup (E_1 \cap E_2) \cup (E_2 - E_1)$$

und diese drei (geklammerten) Ereignisse schließen einander paarweise aus. Also erhält man wegen der Additivität

$$P(E_1 \cup E_2) = P(E_1 - E_2) + P(E_1 \cap E_2) + P(E_2 - E_1) \ .$$

Da aber

$$P(E_1 - E_2) + P(E_1 \cap E_2) = P(E_1)$$

und

$$P(E_2 - E_1) + P(E_1 \cap E_2) = P(E_2)$$

gilt, ergibt sich die gesuchte Beziehung.

Den allgemeinen Fall für $k > 2$ erhält man durch Induktion. Beispielsweise gilt wegen der vorher gezeigten Beziehung

$$P(E_1 \cup E_2 \cup E_3) = P((E_1 \cup E_2) \cup E_3) = P(E_1 \cup E_2) + P(E_3) - P((E_1 \cup E_2) \cap E_3) \ .$$

Nun hat man aber

$$\begin{aligned} P((E_1 \cup E_2) \cap E_3) &= P((E_1 \cap E_3) \cup (E_2 \cap E_3)) \\ &= P(E_1 \cap E_3) + P(E_2 \cap E_3) - P((E_1 \cap E_3) \cap (E_2 \cap E_3)) \\ &= P(E_1 \cap E_3) + P(E_2 \cap E_3) - P(E_1 \cap E_2 \cap E_3) \end{aligned}$$

und damit wieder die behauptete Beziehung. △

Verlangt man, dass bei einem Versuch mit nur endlich vielen Ausgängen $M = \{\omega_1, \omega_2, \ldots, \omega_m\}$ allen diesen Situationen dieselbe Eintrittschance zukommt, dann gilt:

Klassische Wahrscheinlichkeitsdefinition: Ist $E = \{\omega_{i_1}, \omega_{i_2}, \ldots, \omega_{i_g}\} \subset M$ ein zum betrachteten Versuch gehörendes Ereignis, dann ergibt sich die Wahrscheinlichkeit für den Eintritt von E zu:

$$P(E) = \frac{g}{m} = \frac{\text{"Anzahl der günstigen Fälle"}}{\text{"Anzahl der möglichen Fälle"}} \tag{3.4}$$

Beweis:

Wegen der Forderung der Chancengleichheit muss unter Ausnützung der Addivität

$$1 = P(M) = P(\bigcup_{i=1}^{m} \{\omega_i\}) = \sum_{i=1}^{m} P(\{\omega_i\}) = m \times p$$

und somit $p = \frac{1}{m}$ gelten. Nun führt aber der neuerliche Einsatz der Addivität zu

$$P(E) = P(\bigcup_{l=1}^{g} \{\omega_{i_l}\}) = \sum_{l=1}^{g} P(\{\omega_{i_l}\}) = g \times p$$

und damit zur behaupteten Formel. △

Beispiel 3.1 Ein (fairer) Würfel wird einmal geworfen. Also ist $M = \{1, 2, 3, 4, 5, 6\}$ und jede Teilmenge davon ist als Ereignis denkbar. Die Wahrscheinlichkeit einer ungeraden Augenanzahl – $E = \{1, 3, 5\} \subset M$ – ergibt sich wegen $m = 6$ und $g = 3$ zu

$$P(E) = \frac{3}{6} = 0.5$$

◇◇◇

Beispiel 3.2 Es werden zwei Würfel gleichzeitig geworfen. Damit ist $M = \{(1,1), (1,2), \ldots, (1,6), (2,1), \ldots, (2,6), \ldots, (6,1), \ldots, (6,6)\}$ und wiederum sind alle Teilmengen sinnvolle Ereignisse. Die Wahrscheinlichkeit einer Augensumme unter 5 ergibt sich wegen

$$E = \{(1,1), (1,2), (1,3), (2,1), (2,2), (3,1)\}$$

nach (3.4) zu

$$P(E) = \frac{6}{36} = \frac{1}{6} = 0.167$$

◇◇◇

3.1.2 Einige Grundbegriffe der Kombinatorik

Die Ermittlung der Anzahl möglicher bzw. günstiger Fälle verlangt häufig die Kenntnis einfacher kombinatorischer Grundbegriffe. Die wichtigsten werden im folgenden aufgelistet:

Permutationen ohne Wiederholung:

Jede Anordnung (Vertauschung) von n (verschiedenen) Elementen heißt Permutation. Für die Buchstaben a, b und c stellen beispielsweise abc, acb, bac, bca, cab, cba mögliche Permutationen (ohne Wiederholung) dar. Die Anzahl derartiger Vertauschungsmöglichkeiten ergibt sich zu

$$P_n = n! = n \times (n-1) \times (n-2) \times \ldots \times 2 \times 1$$

(verbal: n Faktorielle). Bei $n = 3$ erhält man die $n! = 3! = 6$ oben genannten Anordnungen.

Permutationen mit Wiederholung:

Lässt man bei den n anzuordnenden Elementen auch gleichartige zu, also n_1-mal ein Element a_1, n_2-mal ein Element a_2 usw. und schließlich n_k-mal das Element a_k, so sind Anordnungen bei denen gleichlautende Elemente vertauscht werden, natürlich ident. Beispielsweise ergeben sich für aab noch die alternativen Anordnungen aba und baa, also insgesamt drei. Allgemein gilt für die Anzahl von Permutationen mit Wiederholung

$$\bar{P}_{n;n_1,\ldots,n_k} = \frac{n!}{n_1! \ldots n_k!} \quad .$$

Für das Beispiel gilt $n_1 = 2$ und $n_2 = 1$ und damit

$$\bar{P}_{3;2,1} = \frac{3!}{2!1!} = 3 \quad ,$$

was mit der obigen Bemerkung übereinstimmt.

Kombinationen ohne Wiederholung:

Jede Anzahl von k (verschiedenen) Elementen aus einer Grundmenge von n Elementen nennt man eine "Kombination von n Element zur Klasse k" (ohne Wiederholung). Beispielsweise können bei $n = 4$ Tennisspielern die Paarungen ($k = 2$) 12, 13, 14, 23, 24 und 34 gebildet werden. Es gibt

$$K_{n,k} = \frac{n!}{k!(n-k)!} =: \binom{n}{k}$$

(Binomialkoeffizient!) derartiger Kombinationen. Für das Beispiel ergibt dies

$$K_{4,2} = \frac{4!}{2!2!} = \frac{4 \times 3 \times 2}{2 \times 2} = 6 \quad ,$$

also die oben genannten Fälle.

Beispiel 3.3 Eine Transportkiste enthält $N = 100$ Hühnereier, unter denen sich $A = 5$ angeschlagene Eier (Ausschusseinheiten) befinden. Um die Qualität dieses Loses zu überprüfen, werden zufällig $n = 10$ Eier (Stichprobe) ausgewählt und untersucht. Wie groß ist die Wahrscheinlichkeit, dass dabei $a = 2$ kaputte Eier gefunden werden?

Lösung: Insgesamt gibt es $m = \binom{N}{n}$ mögliche Stichprobenkonstellationen, von denen aber nur solche für das betrachtete Ereignis günstig sind, die genau a Ausschusseinheiten enthalten. Dazu müssen offensichtlich aus den vorhandenen A angeschlagenen Eiern a ausgewählt werden und aus den $N - A$ intakten Eiern $n - a$. Für den ersten der beiden Auswahlgänge gibt es $\binom{A}{a}$ Auswahlmöglichkeiten, für den zweiten $\binom{N-A}{n-a}$. Da jede Auswahl für die Ausschusseinheiten mit jeder für die intakten Eier kombiniert werden kann, um eine zulässige Stichprobensituation zu erreichen, gibt es $g = \binom{A}{a}\binom{N-A}{n-a}$ für das betrachtete Ereignis "günstige" Situationen. Somit ergibt sich die gesuchte Wahrscheinlichkeit allgemein zu

$$P(''a'') = \frac{\binom{A}{a}\binom{N-A}{n-a}}{\binom{N}{n}}$$

und mit den konkreten Werten als

$$\begin{aligned} P(''2'') &= \frac{\binom{5}{2}\binom{100-5}{10-2}}{\binom{100}{10}} = \frac{5!}{2!3!}\frac{10!90!}{100!}\frac{95!}{8!87!} \\ &= \frac{5 \times 4}{2 \times 1} \times \frac{10 \times 9 \ldots 2 \times 1}{100 \times 99 \ldots 92 \times 91} \times \frac{95 \times 94 \ldots 88}{8 \times 7 \ldots 2 \times 1} \\ &= 0.0702 \end{aligned}$$

Kombinationen mit Wiederholung:

Können bei einer Kombination von k aus n Elementen auch gleichartige vorkommen (mit Wiederholung bzw. mit Zurücklegen), so sind die verschiedenen Kombinationen gegeben durch

$$K_{n,k} = \binom{n+k-1}{k}$$

Variation ohne Zurücklegen:

Um k Elemente aus einer Grundmenge von n Elementen auszuwählen, wobei jedoch die Reihenfolge eine Rolle spielt, so nennt man dies eine geordnete Auswahl bzw. Variation oder auch k-Permutation. Die Anzahl der k-Permutationen ist gegeben durch

$$P_{n,k} = n \times (n-1) \times (n-2) \ldots (n-k+1) = \frac{n!}{(n-k)!}$$

Variation mit Zurücklegen:

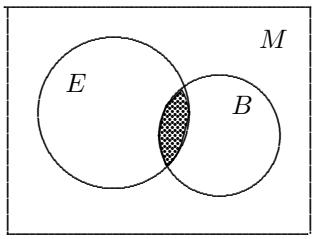
Zieht man bei einer Variation mit Zurücklegen, so muss man berücksichtigen, dass ein Element mehrere Male vorkommen kann. Da jedoch die Reihenfolge hier ebenfalls eine Rolle spielt, gibt es n Möglichkeiten für jedes Element k. Daraus ergibt sich die Anzahl der k-Permutationen mit folgender Formel

$$P_{n,k} = n^k$$

3.2 Bedingte Wahrscheinlichkeit

Durch Zusatzinformation über den Versuchsverlauf wird möglicherweise die Wahrscheinlichkeit für das Auftreten eines betrachteten Ereignisses beeinflusst. Fragt man bei einem Studenteninterview u.a. nach der absolvierten Schule, so wird die Antwort (das Ereignis) "HTL" vielleicht in 20 % der Fälle gegeben werden. Steht der Interviewer aber zufällig vor der Universität Wien, wird dieser Prozentsatz drastisch sinken, steht er vor der Technischen Universität Wien, wird er deutlich höher sein. Die Wahrscheinlichkeit wird in beiden Fällen *relativiert* oder *bedingt* durch zusätzliche Information bzw. den Eintritt eines *bedingenden Ereignisses* (siehe Abb.3.2)

Abbildung 3.2: *Bedingte Wahrscheinlichkeit*



Definition: Unter der *bedingten Wahrscheinlichkeit* $P(E \mid B)$ eines Ereignisses E, bedingt durch den Eintritt eines Ereignisses B mit $P(B) > 0$ versteht man den durch

$$P(E \mid B) = \frac{P(E \cap B)}{P(B)} \tag{3.5}$$

erklärten Ausdruck.

Beispiel 3.4 Beim Würfeln mit zwei (fairen) Würfeln lautet eine mögliche Frage etwa nach der Wahrscheinlichkeit, dass die Augensumme größer als sieben ist, wenn mindestens einer

der beiden Würfeln eine Sechs zeigt. Damit ist das betrachtete Ereignis E "Augensumme > 7" und das bedingende Ereignis B "mindestens 1×6". Nun gilt

$$\begin{aligned} P(B) &= P(\{(6,1),\ldots,(6,6),(1,6),\ldots,(5,6)\}) = \frac{11}{36} \\ P(E) &= P(\{(2,6),(3,5),(3,6),\ldots,(6,2),\ldots,(6,6)\}) = \frac{15}{36} \\ P(E \cap B) &= P(\{(6,2),\ldots,(6,6),(2,6),\ldots,(5,6)\}) = \frac{9}{36} \quad . \end{aligned}$$

Damit erhält man für die durch B bedingte Wahrscheinlichkeit von E

$$P(E \mid B) = \frac{9/36}{11/36} = \frac{9}{11} = 0.818 \quad ,$$

was im Vergleich zur (nichtbedingten, absoluten) Wahrscheinlichkeit $P(E) = 0.417$ natürlich deutlich höher ausfällt.

Bemerkung: Je nach dem Verhältnis der durch B bedingten zur nichtbedingten Wahrscheinlichkeit von E spricht man von

$$P(E \mid B) \left\{ \begin{array}{c} < \\ = \\ > \end{array} \right\} P(E) \;\Leftrightarrow\; \left\{ \begin{array}{l} B \text{ behindert } E \\ B \text{ beeinflusst } E \text{ nicht (d.h. unabhängig)} \\ B \text{ begünstigt } E \quad . \end{array} \right.$$

Falls ein Ereignis B das Ereignis E nicht beeinflusst, sind diese beiden Ereignisse offensichtlich unabhängig. Um von der Voraussetzung $P(B) > 0$ loszukommen, erklärt man die Unabhängigkeit von Ereignissen allgemein durch folgende

Definition: Zwei Ereignisse E_1 und E_2 heißen *unabhängig*, wenn

$$P(E_1 \cap E_2) = P(E_1) \times P(E_2) \tag{3.6}$$

zutrifft. Die Ereignisse $E_1, E_2, \ldots, E_n$ nennt man unabhängig, wenn für jede Auswahl von k Ereignissen $E_{i_1}, \ldots, E_{i_k}$ $(k \leq n)$

$$P(\bigcap_{l=1}^{k} E_{i_l}) = \prod_{l=1}^{k} P(E_{i_l}) \tag{3.7}$$

gilt.

Bemerkung: Es gilt nun tatsächlich für $P(B) > 0$:

$$E \text{ und } B \text{ sind unabhängig } \Leftrightarrow\; P(E \mid B) = P(E)$$

Beweis:
a) E und B unabhängig $\Rightarrow\; P(E \cap B) = P(E) \times P(B) \;\Rightarrow\; P(E \mid B) = \frac{P(E \cap B)}{P(B)} = P(E)$.
b) $P(E \mid B) = P(E) \;\Rightarrow\; P(E) = \frac{P(E \cap B)}{P(B)} \;\Rightarrow\; P(E \cap B) = P(E) \times P(B)$. $\triangle$

Beispiel 3.5 Beim zweimaligen Werfen einer Münze bezeichnet E_1 "Adler beim 1. Wurf" und E_2 "Kopf beim 2. Wurf". Es gilt

$$\begin{aligned} P(E_1) &= P(\{(A,A),(A,K)\}) = \frac{1}{2} \\ P(E_2) &= P(\{(A,K),(K,K)\}) = \frac{1}{2} \\ P(E_1 \cap E_2) &= P(\{(A,K)\}) = \frac{1}{4} \quad , \end{aligned}$$

also sind E_1 und E_2 unabhängig.

◇◇◇

Die folgenden Aussagen stellen klassische Regeln der Wahrscheinlichkeitsrechnung dar und finden sich häufig trotz ihrer einfachen Form und Herleitung als "Theoreme" bezeichnet.

Multiplikationsregel:

Für eine beliebige Folge $E_1, E_2, \ldots, E_n \in \mathbf{E}$ von Ereignissen gilt

a) $$P(E_1 \cap E_2) = P(E_1) \times P(E_2|E_1) \; ; \tag{3.8}$$

b) $$P(E_1 \cap E_2 \cap \cdots \cap E_n) = \tag{3.9}$$
$$= \; P(E_1) \times P(E_2|E_1) \times P(E_3|E_1 \cap E_2) \times \cdots \times P(E_n|E_1 \cap \cdots \cap E_{n-1})$$

Beweis:

a) Die Beziehung (3.8) folgt unmittelbar aus der Definition der bedingten Wahrscheinlichkeit.

b) Die Formel (3.9) leitet sich induktiv aus a) ab, wie man sich etwa für den Fall $n = 3$ einfach überzeugen kann:

$$\begin{aligned} &P(E_1 \cap E_2 \cap E_3) = \\ &= \; P((E_1 \cap E_2) \cap E_3) = P(E_1 \cap E_2) \times P(E_3|E_1 \cap E_2) = \\ &= \; P(E_1) \times P(E_2|E_1) \times P(E_3|E_1 \cap E_2) \; . \end{aligned}$$

△

Satz von der vollständigen Wahrscheinlichkeit:

Lässt sich der Merkmalraum in endlich oder höchstens abzählbar unendlich viele Teilereignisse ("Hypothesen") $H_1, H_2, \ldots$ zerlegen, die einander paarweise ausschließen, so gilt für ein beliebiges Ereignis E:

$$P(E) = \sum_{i=1}^{\infty} P(E|H_i)\, P(H_i) \; . \tag{3.10}$$

Beweis:

Wie aus Abb.3.3 ersichtlich ist, gilt zunächst

$$P(E) = P(E \cap M) = P(E \cap (\cup_{i=1}^{\infty} H_i)) = P(\cup_{i=1}^{\infty} (E \cap H_i))$$

und zusammen mit der Multiplikationsregel für

$$P(E \cap H_i) = P(H_i)\, P(E|H_i)$$

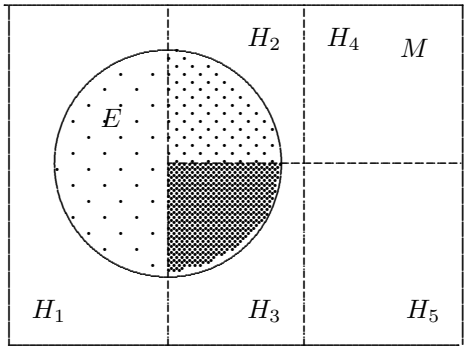
die Beziehung (3.10). △

Beispiel 3.1: Ein willkürlich aus der Menge aller Studenten herausgegriffener Student besitzt eine eindeutig bestimmte *Stammuniversität* (A: Universität Wien, B: Universität Graz,..., E: TU Wien, F: TU Graz, usw.). Für jede Universität ist der Anteil an HTL–Absolventen (Ereignis E) bekannt. Außerdem kennt man die Aufteilung der Studenten auf die 12 österreichischen Universitäten (Hypothesen $H_i,\ i = 1, \ldots, 12$). Die Frage nach dem österreichweiten Anteil ($\widehat{=}$ Wahrscheinlichkeit) der HTL–Absolventen unter den Studenten wird mit Hilfe von (3.10) beantwortet.

◇◇◇

Gleichsam die umgekehrte Fragestellung liegt vor, wenn das Ereignis E eintritt und die Frage interessiert, mit welcher Wahrscheinlichkeit die zugrundeliegenden Hypothesen dabei auftreten. Die Antwort gibt der folgende

Abbildung 3.3: *Zerlegung in Hypothesen*



Satz von Bayes [beïs]:

In Fortsetzung des Satzes von der vollständigen Wahrscheinlichkeit gilt

$$P(H_j|E) = \frac{P(E|H_j)\,P(H_j)}{\sum_{i=1}^{\infty} P(E|H_i)\,P(H_i)}\ . \tag{3.11}$$

Beweis:

In der Formel

$$P(H_j|E) = \frac{P(E \cap H_j)}{P(E)}$$

wendet man im Zähler die Multiplikationsregel an, ersetzt den Nenner durch die Beziehung (3.10) und erhält damit die angegebene Beziehung. △

Beispiel 3.6 Vier Obstbauern beliefern zur Erntezeit einen Großhändler mit einer bestimmten Apfelsorte. Eine Obstkiste sollte ca. 20 kg wiegen. Die Anlieferung verteilt sich auf die Bauern ungefähr im Verhältnis $A : B : C : D = 2 : 3 : 3 : 4$. Manchmal wiegt eine Obstkiste zuwenig. Bei A passiert das in einem von 20 Fällen, bei B liegt die Wahrscheinlichkeit um 4 %, bei C um 5 % und für D ist der Anteil zu leichter Kisten ungefähr bei 0.03. Der Händler findet am Abend zufällig eine Kiste, die eindeutig zu leicht ist. Allerdings ist beim Stapeln der Kisten die Herkunft durcheinander gekommen. Bei wem soll er sich beschweren?

Lösung: Offensichtlich gilt mit L ("Kiste zu leicht")

$$\begin{array}{ll} P(A) = 2/12 & P(L|A) = 0.05 \\ P(B) = 3/12 & P(L|B) = 0.04 \\ P(C) = 3/12 & P(L|C) = 0.05 \\ P(D) = 4/12 & P(L|D) = 0.03 \end{array}$$

und damit

$$P(L) = 0.05 \times \frac{2}{12} + 0.04 \times \frac{3}{12} + 0.05 \times \frac{3}{12} + 0.03 \times \frac{4}{12} = \frac{0.49}{12} = 0.041\ .$$

Mit dem Satz von Bayes ergibt sich dann

$$P(A|L) = \frac{0.05 \times 2/12}{0.49/12} = 0.204$$

und analog $P(B|L) = 0.245$, $P(C|L) = 0.306$ und $P(D|L) = 0.245$, sodass am ehesten C der Urheber sein könnte.

$\diamond\diamond\diamond$

3.3 Determinierte und zufällige Größen

Eine Reihe von Größen (Merkmalen), die in Natur, Technik oder Gesellschaft beobachtet werden können, haben *festen* Charakter, sie sind *determiniert*. Beispiele dafür sind sind etwa

- die Chromosomenanzahl bei (nicht an Gendefekten leidenden) Lebewesen
- die Anzahl von Elektronen bei chemischen Elementen
- die Anzahl von Parlamentsabgeordenten oder
- der ÖH–Beitrag.

Ein beliebig herausgegriffener Patient besitzt 2 mal 23 Chromosomen, ein willkürlich betrachtetes Sauerstoffatom hat 8 Elektronen, die Zahl der Abgeordneten im österreichischen Nationalrat beträgt zur Zeit 183 und der ÖH–Betrag eines beliebigen Studenten lautet im Studienjahr 2011/2012 € 17.00. In allen genannten Fällen kann das Ergebnis nur so lauten, wie es angegeben ist.

Demgegenüber gibt es die überwiegende Zahl von Phänomenen, wo der Beobachtungswert von vornherein *nicht* feststeht. Beispiele dafür sind etwa

- die Körpergröße einer beliebigen Person,
- die Anzahl von Rindern bei einem willkürlich ausgewählten Landwirt,
- die Milchmenge einer beliebigen Kuh während einer Laktationsperiode,
- der Treibstoffverbrauch eines beliebigen VW–Golf.

In allen Fällen lässt sich der zu beobachtende Wert des Merkmals *nicht mit Sicherheit* angeben. Man kann bloß von der *Wahrscheinlichkeit* sprechen, mit der er in einem *bestimmten Bereich* liegt. Derartige Größen, die durch *Unsicherheit* charakterisiert sind, deren Beobachtungen in gewissem Sinn *zufälliger* Natur sind, nennt man daher *zufällige Größen*. Sie können nur durch Angabe der Wahrscheinlichkeit beschrieben werden, mit der man Werte in einem bestimmten Bereich beobachten kann.

Definition: Eine (eindimensionale) Größe X heißt *zufällig (Zufallsgröße (ZG), Zufallsvariable, stochastische Größe)*, wenn durch Aussagen der Form

$$X \leq x,\ x \in \boldsymbol{R} \tag{3.12}$$

Ereignisse beschrieben werden, deren Eintreten mit einer bestimmten *Wahrscheinlichkeit* zu beobachten ist. Analog spricht man von einer (*m–dimensionalen ZG*) $\mathbf{X} = (X_1, X_2, \ldots, X_m)$, falls durch

$$X_1 \leq x_1, \ldots, X_m \leq x_m \quad \text{mit } x_i \in \boldsymbol{R} \text{ für } i = 1, \ldots, m \tag{3.13}$$

ein (sinnvolles) Ereignis beschrieben werden kann, das mit einer bestimmten Wahrscheinlichkeit zu beobachten ist.

Die Werte, die eine *ZG* X annehmen kann, bilden den *Wertevorrat* bzw. den *Merkmalraum* M_X dieser *ZG*.

Eine *ZG* lässt sich nur dadurch sinnvoll beschreiben, dass man für *jedes* Ereignis der Form (3.12) bzw. (3.13) die oben erwähnte Wahrscheinlichkeit angibt. Man spricht dann von der *Wahrscheinlichkeitsverteilung* P_X der *ZG* X. Ein wichtiges Hilfsmittel zur Beschreibung dieser Verteilung ist durch die sogenannte Verteilungsfunktion einer *ZG* gegeben:

Definition: Unter der *Verteilungsfunktion (*engl. *cumulative distribution function) (VF)* F_X einer *ZG* X versteht man die durch

$$F_X(x) := P_X((-\infty, x]) = P(X \leq x) \quad \text{für } x \in \boldsymbol{R} \tag{3.14}$$

definierte Funktion auf $\boldsymbol{R}$.

Für eine *VF* gilt offensichtlich:

VF 1) $0 \leq F_X(x) \leq 1 \quad \forall x \in \boldsymbol{R}$;

VF 2) F_X ist monoton wachsend mit

$$\lim_{x \to -\infty} F_X(x) = 0 \quad \text{und} \quad \lim_{x \to \infty} F_X(x) = 1 \, ;$$

VF 3) $P_X((a, b]) = P(a < X \leq b) = F_X(b) - F_X(a) \quad \text{für } -\infty < a \leq b < \infty \, .$

Beweis:
Da die *VF* eine Wahrscheinlichkeit beschreibt, ergibt sich unmittelbar VF 1. Die zweite Beziehung folgt, weil $P(X \leq x_1) \leq P(X \leq x_2)$ für $x_1 < x_2$. Schließlich gilt wegen

$$(-\infty, b] = (-\infty, a] \cup (a, b]$$

und der Additivität der Wahrscheinlichkeit

$$F_X(b) = P_X((-\infty, b]) = P_X((-\infty, a]) + P_X((a, b]) = F_X(a) + P_X((a, b]) \, ,$$

woraus die letzte Beziehung folgt. △

Die zwei wichtigsten Typen von *ZGen* sind die *diskreten* und die *stetigen* (oder auch *kontinuierlichen*) *ZGen*.

Definition: Eine *ZG* X heißt *diskret*, wenn es höchstens abzählbar unendlich viele Werte gibt, die X annehmen kann. Es gilt also offenbar mit $M_X = \{x_1, x_2, \ldots\}$ und $p_X(x_i) = P(X = x_i)$

$$\sum_{x_i \in M_X} p_X(x_i) = 1 \, .$$

Die Funktion p_X, die jedem möglichen Wert von X seine Wahrscheinlichkeit zuordnet, heißt *Wahrscheinlichkeitsfunktion (W-Fkt)*. Sie definiert die Verteilung von X vollständig.

Beispiel 3.7 Der zugrundeliegende Versuch besteht im Würfeln mit zwei (fairen) Würfeln. Die *ZG* X beschreibt die Augensumme. Damit ist $M_X = \{2, 3, \ldots, 11, 12\}$ und X demnach diskret. Für die *W-Fkt* erhält man die Werte in der folgenden Tabelle

i	2	3	4	5	6	7	8	9	10	11	12
$p_X(i)$	1/36	2/36	3/36	4/36	5/36	6/36	5/36	4/36	3/36	2/36	1/36

Dabei ergibt sich etwa $p_X(5)$ als

$$p_X(5) = P(X = 5) = P(\{(1\text{-}4), (2\text{-}3), (3\text{-}2), (4\text{-}1)\}) = 4/36 \ .$$

◇◇◇

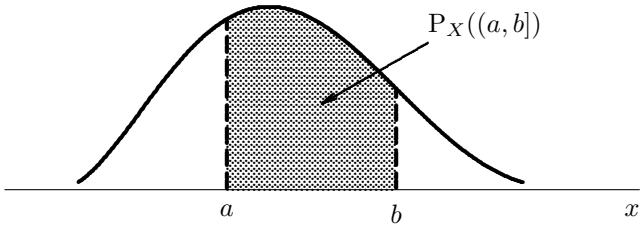
Definition: Eine *ZG X* heißt *stetig* (oder *kontinuierlich*), wenn ihr Wertebereich die ganzen reellen Zahlen $\boldsymbol{R}$ oder eine überabzählbare Teilmenge davon (positive reelle Zahlen, Intervall, ...) ist.

Gibt es eine *nichtnegative* Funktion $f_X : \boldsymbol{R} \to \boldsymbol{R}_+$ gibt, sodass für jedes $-\infty < a \leq b < \infty$ gilt:

$$P(a < X \leq b) = \int_a^b f_X(x)\, dx \tag{3.15}$$

so heißt diese *Dichtefunktion (DF)*, kurz *Dichte*. Wir beschäftigen uns nur mir stetigen Zufallsgrößen für die es auch eine Dichte gibt. Die Wahrscheinlichkeit eines Ereignisses für X lässt sich dann als Integral der *DF* über die Ereignismenge darstellen (siehe Abb.3.4).

Abbildung 3.4: *Dichtefunktion*



Beispiel 3.8 Das Abwiegen einer Person erfolgt auf einer Personenwaage mit Digitalanzeige in *kg*. Die *ZG X* beschreibt die Differenz zwischen tatsächlichem Gewicht und der Anzeige. Wenn man die Anzeige als gerundeten Wert auffasst, ist hier $M_X = [-0.5, 0.5)$ und X stetig mit der *DF* aus Abb.3.5. Der (konstante) Wert h der *DF* im Bereich zwischen -0.5 und 0.5 ergibt sich wegen

$$1 = \int_{-\infty}^{\infty} f_X(x)\, dx = \int_{-0.5}^{0.5} h\, dx = 1 \times h$$

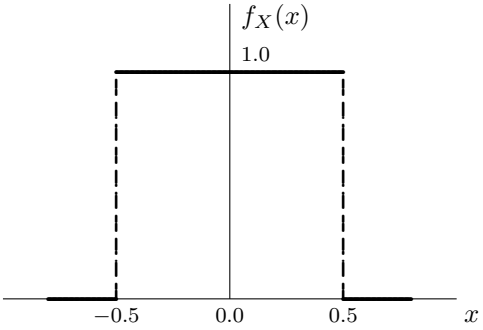
einfach als $h = 1$.

◇◇◇

3.4 Momente von Zufallsgrößen

Ähnlich wie bei den Häufigkeitsverteilungen in der beschreibenden Statistik, sollen aussagekräftige Kenngrößen auch bei Wahrscheinlichkeitsverteilungen möglichst viel an Information vermitteln. Damit würde es ausreichen, einige wenige Parameter anstelle einer Funktion (*VF, DF, W-Fkt*) zur einigermaßen zufriedenstellenden Beschreibung einer *ZG* heranzuziehen. Zu den wichtigsten dieser Parameter gehören die Momente einer *ZG* bzw. Verteilung.

Abbildung 3.5: *S(-0.5,0.5)–Gleichverteilung*



3.4.1 Mittelwert, Erwartung

Der analoge Vorgang zur Durchschnittsbildung einer Stichprobe ist im Fall einer (theoretischen) Wahrscheinlichkeitsverteilung die gewichtete Mittelung aller möglichen Werte, die eine betrachtete *ZG* X annehmen kann. Das führt bei diskreten *ZGen* auf

$$\mathrm{E}(X) = \mu_X = \sum_{x_i \in M_X} x_i \, p_X(x_i) \tag{3.16}$$

und bei stetigen *ZGen* auf

$$\mathrm{E}(X) = \mu_X = \int_{-\infty}^{\infty} x \, f_X(x) \, dx \tag{3.17}$$

und heißt *Mittelwert* oder *Erwartungswert* von X.

Beispiel 3.9 Für die *ZG* aus Bsp.3.8 erhält man den Mittelwert

$$\mu_X = 2 \times \frac{1}{36} + 3 \times \frac{2}{36} + \cdots + 11 \times \frac{2}{36} + 12 \times \frac{1}{36} = \frac{252}{36} = 7$$

und für die *ZG* aus Bsp.3.9 den Mittelwert

$$\mu_X = \int_{-0.5}^{0.5} x \cdot 1 \, dx = \left. \frac{x^2}{2} \right|_{-0.5}^{0.5} = 0.125 - 0.125 = 0 \ .$$

Definition: Ist $g: \boldsymbol{R} \to \boldsymbol{R}$ eine derartige reellwertige Funktion, dass mit X auch $Y := g(X)$ eine *ZG* darstellt, so versteht man unter der *Erwartung von* $g(X)$ den durch

$$\mathrm{E}(g(X)) = \mu_Y = \begin{cases} \sum_{x_i \in M_X} g(x_i) \, p_X(x_i) & \text{falls } X \text{ diskret} \\ \int_{-\infty}^{\infty} g(x) \, f_X(x) \, dx & \text{falls } X \text{ stetig} \end{cases} \tag{3.18}$$

definierten Wert.

3.4.2 Varianz, Standardabweichung, höhere Momente

Analog zur Stichprobensituation werden Varianz und Standardabweichung auch als (theoretische) Parameter über die durchschnittliche quadratische Abweichung vom (theoretischen) Mittelwert bzw. als Quadratwurzel davon erklärt. Daher verwendet man folgende

Definition: Die *Varianz* einer *ZG* X ist durch

$$\sigma_X^2 = \mathrm{Var}(X) = \mathrm{E}((X - \mu_X)^2) \tag{3.19}$$

erklärt. Es gilt also im diskreten Fall

$$\sigma_X^2 = \sum_{x_i \in M_X} (x_i - \mu_X)^2 \, p_X(x_i)$$

und bei einer stetigen *ZG*

$$\sigma_X^2 = \int_{-\infty}^{\infty} (x - \mu_X)^2 \, f_X(x) \, dx \ .$$

Die (positive) Wurzel σ_X von σ_X^2 wird als (theoretische) *Standardabweichung* von X bezeichnet. Sie stellt ein dimensionsgleiches Streuungsmaß für X dar.

Es gilt:

$$\sigma_X^2 = \mathrm{E}(X^2) - (\mathrm{E}(X))^2 \tag{3.20}$$

Beweis:
Die Erwartung ist linear, also gilt

$$\begin{aligned} \mathrm{E}((X - \mu_X)^2) &= \mathrm{E}(X^2 - 2\mu_X X + \mu_X^2) \\ &= \mathrm{E}(X^2) - 2\mu_X \mathrm{E}(X) + \mu_X^2 = \mathrm{E}(X^2) - \mu_X^2 \ . \end{aligned}$$

△

Eine wertvolle Hilfe zur Abschätzung von Wahrscheinlichkeiten – unabhängig von jeder Annahme über die Art der Verteilung – stellt die im folgenden formulierte Tschebyschev'sche Ungleichung dar.

Tschebyschev'sche Ungleichung:

Für die *ZG* X mit der Varianz σ_X^2 gilt

$$P(|X - \mu_X| > k\sigma_X) \leq \frac{1}{k^2} \tag{3.21}$$

Beweis:
Für eine stetige *ZG* erhält man

$$\begin{aligned} P(|X - \mu_X| \geq k\sigma_X) &= P\left(\frac{(X - \mu_X)^2}{k^2\sigma_X^2} \geq 1\right) = \int_{(X-\mu_X)^2/k^2\sigma_X^2 \geq 1} f_X(x) \, dx \\ &\leq \int_{(X-\mu_X)^2/k^2\sigma_X^2 \geq 1} \frac{(x - \mu_X)^2}{k^2\sigma_X^2} \, f_X(x) \, dx \\ &\leq \int_{-\infty}^{\infty} \frac{(x - \mu_X)^2}{k^2\sigma_X^2} \, f_X(x) \, dx = \frac{1}{k^2} \end{aligned}$$

△

Damit erhält man für große k–Werte ($k \geq 2$) brauchbare Abschätzungen für Überschreitungswahrscheinlichkeiten; z.B. gilt für $k = 3$

$$P(|X - \mu_X| \geq 3\sigma_X) \leq \frac{1}{9} = 0.111 \ .$$

Definition: Die durch

$$\mu_{X;k} := \mathrm{E}(X^k) \tag{3.22}$$

erklärte Kenngröße heißt *k–tes Moment* der *ZG* X. Die modifizierte Form

$$\mu'_{X;k} := \mathrm{E}((X - \mu_X)^k) \tag{3.23}$$

stellt das *k–te zentrierte Moment* von X dar.

Bemerkung: Die Varianz ist somit das zentrierte 2. Moment von X.

Ähnlich dem Stichprobenfall dienen höhere Momente zur verfeinerten Beschreibung der Form von Verteilungen. So beschreibt etwa das *standardisierte* 3. Moment

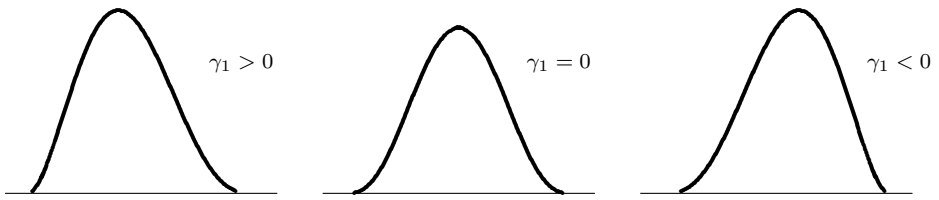
$$\gamma_1 := \mathrm{E}((X - \mu_X)^3/\sigma_X^3) = \frac{\mu'_{X;3}}{\sigma_X^3} \tag{3.24}$$

die Symmetrie einer Verteilung und wird daher auch *Schiefe (*engl. *skewness)* genannt. Es gilt

$$\gamma_1 \left\{ \begin{array}{c} < \\ = \\ > \end{array} \right\} 0 \text{ , falls Verteilung } (DF) \left\{ \begin{array}{l} \text{rechtssteil} \\ \text{symmetrisch} \\ \text{linkssteil} \end{array} \right. \tag{3.25}$$

(siehe Abb.3.6).

Abbildung 3.6: *Schiefe*

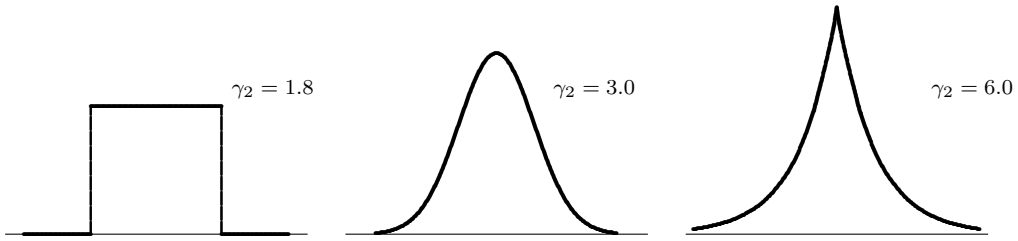


Zur Beschreibung der Verteilung in ihren Außenbereichen dient das *standardisierte* 4. Moment

$$\gamma_2 := \mathrm{E}((X - \mu_X)^4/\sigma_X^4) = \frac{\mu'_{X;4}}{\sigma_X^4} \ , \tag{3.26}$$

das als *Kurtosis* bezeichnet wird. Da γ_2 ein gerades Moment darstellt, ist es immer positiv. Es fällt umso größer aus, je größer die Wahrscheinlichkeit ist, dass im Verhältnis zum zentralen Bereich weit abliegende Werte beobachtet werden können. Die Abb.3.7 zeigt einige Beispiele

Abbildung 3.7: *Kurtosis*



für die Kurtosis spezieller Verteilungsformen. Für die Normalverteilung (siehe Abschnitt 6) ist $\gamma_2 = 3$. Daher wird manchmal die Kurtosis einer Verteilung darauf bezogen und die Größe

$$\gamma_{2;0} := \gamma_2 - 3$$

im Vergleich mit der Normalverteilung betrachtet. Diese reduzierte Kurtosis wird als *Exzess*, manchmal aber ebenfalls als Kurtosis bezeichnet. Auf den letzten Umstand ist bei der Auswertung mit statistischen Softwarepaketen zu achten.

3.5 Diskrete Zufallsgrößen (Verteilungen)

Im folgenden sind die wichtigsten diskreten Verteilungen zusammengestellt. Dabei werden jeweils der Wertebereich $M_X = \{x_1, x_2, \ldots\}$ der zugeordneten *ZG* X, die *W–Fkt*

$$p_X(x_i) = P(X = x_i) \quad \text{für } x_i \in M_X$$

mit

$$\sum_{x_i \in M_X} p_X(x_i) = 1$$

und die wichtigsten Momente angegeben. Außerdem sind Anwendungssituationen genannt.

3.5.1 Diskrete Gleichverteilung $D(m)$

Eine gleichverteilte *ZG* X beschreibt in codierter Form einen Versuch mit m gleichwahrscheinlichen Ausgängen, bei dem somit die klassische Wahrscheinlichkeitsdefinition angewendet werden kann. Es gilt daher $M_X = \{1, 2, \ldots, m\}$ und für die *W–Fkt*

$$p_X(i) = \frac{1}{m} \quad \text{für } i = 1, 2, \ldots, m \; . \tag{3.27}$$

Eür die ersten Momente gilt:

$$\mu_X = \frac{m+1}{2} \tag{3.28}$$

$$\sigma_X^2 = \frac{m^2 - 1}{12} \; . \tag{3.29}$$

Beweis:
Zunächst erhält man

$$\mu_X = \mathrm{E}(X) = \sum_{i=1}^{m} i\, p_X(i) = \sum_{i=1}^{m} \frac{i}{m} = \frac{m(m+1)}{2m} = \frac{m+1}{2} \ .$$

Für die Varianz berechnet man zunächst

$$\sigma^2_{X;0} = \mathrm{E}(X^2) = \sum_{i=1}^{m} \frac{i^2}{m} = \frac{m(m+1)(2m+1)}{6m} = \frac{(m+1)(2m+1)}{6}$$

und gewinnt dann

$$\begin{aligned} \sigma^2_X &= \sigma^2_{X;0} - \mu^2_X = \frac{(m+1)(2m+1)}{6} - \frac{(m+1)^2}{4} \\ &= \frac{1}{12}(m+1)(4m+2-3m-3) = \frac{(m+1)(m-1)}{12} \ . \end{aligned}$$ △

3.5.2 Alternativverteilung $A(p)$

Die *ZG* X kann nur zwei mögliche Zustände annehmen (Antwort: ja/nein; Qualität: gut/Ausschuss; Existenz: vorhanden/fehlend; Münzwurf: Kopf/Adler), die mit 0 und 1 codiert sind, sodass $M_X = \{0, 1\}$ ist. Es gilt

$$p_X(i) = \begin{cases} p & \text{falls } i = 1 \\ 1-p & \text{falls } i = 0 \end{cases} \tag{3.30}$$

mit $0 \le p \le 1$. Für $p = 0$ bzw. $p = 1$ liegt eine entartete Verteilung vor, weil dann *sicher* bloß *ein* Wert auftreten kann („0“ im ersten, „1“ im zweiten Fall).

Offensichtlich gilt:

$$\mu_X = \mathrm{E}(X) = 1 \cdot p + 0 \cdot (1-p) = p$$

und

$$\sigma^2_X = \mathrm{E}(X^2) - (\mathrm{E}(X))^2 = 1 \cdot p + 0 \cdot (1-p) - p^2 = p - p^2 = p(1-p) \ .$$

3.5.3 Binomialverteilung $Bi(n, p)$

Eine binomialverteilte *ZG* X liegt vor, wenn bei der n–maligen Durchführung eines bestimmten Versuches das Auftreten eines bestimmten Versuchsereignisses E (z.B.: Qualität – Ausschuss; Frage – Antwort ”ja”) gezählt wird, wobei diese Wiederholungen einander nicht beeinflussen, also unabhängig sind.

Damit gilt mit $M_X = \{0, 1, 2, \ldots, n\}$ und $p = P(E)$ für die Wahrscheinlichkeitsfunktion

$$p_X(i) = b_{n,p}(i) = \binom{n}{i} p^i (1-p)^{n-i} \quad \text{für } i = 0, 1, \ldots, n \ . \tag{3.31}$$

Beweis:
Eine Situation, für die das Ereignis E exakt i–mal eintritt, ist offensichtlich:

$$S_{11\ldots100\ldots0} = E_1 \cap E_2 \cap \cdots E_i \cap E^c_{i+1} \cap \cdots E^c_n \ ,$$

wobei der Index auf die Nummer der Versuchsdurchführung hinweist. Da die Ereignisse in der obigen Darstellung unabhängig sind, gilt:

$$\begin{aligned} P(S_{11\ldots100\ldots0}) &= \underbrace{p \cdot p \cdots p}_{i} \cdot \underbrace{(1-p) \cdot (1-p) \cdots (1-p)}_{n-i} \\ &= p^i (1-p)^{n-i} \ . \end{aligned}$$

Da es insgesamt $n!/(i!(n-i)!)$ Situationen gibt, in denen das Ereignis E genau i–mal eintritt, und diesen dieselbe, oben hergeleitete Wahrscheinlichkeit zukommt, ergibt sich die Formel für die *W–Fkt.* $\triangle$

Es gilt für die ersten Momente:

$$\mu_X = np \tag{3.32}$$

$$\sigma_X^2 = np(1-p) \ . \tag{3.33}$$

Beweis:
Zunächst erhält man

$$\begin{aligned} \mu_X = \mathrm{E}(X) &= \sum_{i=0}^{n} i \, p_X(i) = \sum_{i=0}^{n} i \binom{n}{i} p^i (1-p)^{n-i} \\ &= \sum_{i=0}^{n} \frac{i \times n!}{i!(n-i)!} p^i (1-p)^{n-i} \\ &= np \sum_{i=1}^{n} \frac{(n-1)!}{(i-1)!((n-1)-(i-1))!} p^{i-1} (1-p)^{(n-1)-(i-1)} \ . \end{aligned}$$

Ersetzt man in der letzten Summe den Laufindex i durch $j = i-1$, so ergibt sich

$$\mu_X = np \sum_{j=0}^{n-1} \binom{n-1}{j} p^j (1-p)^{n-1-j} \ .$$

In dieser Summe wird die *W–Fkt* einer $Bi(n-1,p)$–Verteilung über alle möglichen Werte $j = 0, 1, \ldots, n-1$ addiert, und das ergibt 1, woraus die Behauptung folgt.

Für die Varianzformel leitet man zunächst analog

$$\mathrm{E}(X(X-1)) = n(n-1)p^2$$

ab und erhält dann

$$\begin{aligned}\sigma_X &= \mathrm{E}((X-\mu_X)^2) = \mathrm{E}(X^2) - \mu_X^2 = \mathrm{E}(X(X-1)) + \mu_X - \mu_X^2 \\ &= n(n-1)p^2 + np - (np)^2 = np(np - p + 1 - np) = np(1-p)\ .\end{aligned}$$ △

Beispiel 3.10 Bei der Prüfung eines Loses vom Umfang N und mit einem Ausschussanteil p werden n Stichprobeneinheiten in folgender (wenig praktizierten) Art geprüft: nach Entnahme und Beobachtung der ersten Einheit (Ausschuss /in Ordnung) wird diese Einheit in das Los zurückgegeben. Danach wird eine zweite Einheit gezogen (dabei kann mit Wahrscheinlichkeit $1/N$ die bereits einmal gezogene Einheit wieder entnommen werden!), beobachtet und wieder zurückgelegt usw. Damit ist vor jeder Ziehung einer Einheit dieselbe Ausgangseinheit gegeben und die einzelnen Ziehungen sind somit unabhängig. Man spricht auch von "Ziehung *mit* Zurücklegen".

◇◇◇

3.5.4 Hypergeometrische Verteilung $H(N, A, n)$

Aus einer Grundmenge von N Elementen, von denen A eine spezielle Eigenschaft besitzen, werden n (Stichproben–) Einheiten *auf einmal* gezogen. Die Anzahl X von derart ausgezeichneten Elementen unter den n gezogenen Einheiten ist hypergeometrisch verteilt. Dabei ist $M_X = \{x_u, x_u+1, \ldots, x_o-1, x_o\}$ mit $x_u := \max(0, n-(N-A))$ und $x_o := \min(n, A)$, weil in der Stichprobe nicht mehr nichtausgezeichnete Einheiten auftreten, als die insgesamt $N-A$ vorhandenen, und nicht mehr ausgezeichnete vorkommen können als die insgesamt gegebenen A. Für die *W–Fkt* erhält man (siehe Bsp.3.3)

$$p_X(i) = h_{N,A,n}(i) = \frac{\binom{A}{i}\binom{N-A}{n-i}}{\binom{N}{n}} \quad \text{für } x_u \le i \le x_o\ . \tag{3.34}$$

Es gilt für die ersten Momente:

$$\mu_X = n\frac{A}{N} \tag{3.35}$$

$$\sigma_X^2 = n\frac{A}{N}(1-\frac{A}{N})(1-\frac{n-1}{N-1})\ . \tag{3.36}$$

Bemerkung 1: Die im Vergleich zur Binomialverteilung kleinere Varianz bei der hypergeometrischen Verteilung erklärt sich dadurch, dass bei letzterer mindestens soviele Einheiten geprüft werden wie im Fall der ersten. Damit ist der Informationsgehalt im Durchschnitt größer, was gleichbedeutend mit höherer Genauigkeit bzw. kleinerer Varianz ist.

Bemerkung 2: Je größer der Losumfang N im Vergleich zum Stichprobenumfang n wird, desto weniger fällt der Unterschied zwischen der Ziehung ohne und mit Zurücklegen auf. Tatsächlich gilt

$$\lim_{N\to\infty} h_{N,Np,n}(i) = b_{n,p}(i) \quad \text{für } i = 0, 1, \ldots, n\ . \tag{3.37}$$

Als Faustregel gilt, dass eine $H(N, A, n)$–Verteilung durch eine $Bi(n, A/N)$–Verteilung approximiert werden kann, wenn der Stichprobenanteil n/N maximal 10 % beträgt.

Beispiel 3.11 Betrachtet man die Situation aus Bsp.3.11 und wählt die üblicherweise praktizierte Variante, dass geprüfte Einheiten während der Prüfung nicht wieder in das Los vom Umfang N zurückgegeben werden, ist die Anzahl von Ausschusseinheiten in der Stichprobe $H(N, A, n)$–verteilt.

◇◇◇

3.5.5 Geometrische Verteilung $G(p)$

Wiederholt man einen Versuch solange, bis erstmals ein ausgezeichnetes Ereignis E (Würfeln einer "6" beim Mensch-ärgere-Dich-nicht–Spiel; erfolgreicher Telefonkontakt bei telefonischen Terminvereinbarungen) eintritt, und sind die einzelnen Versuchswiederholungen unabhängig, so ist die Anzahl X notwendiger Versuchsdurchführungen geometrisch verteilt mit dem Parameter $p = P(E)$. Es gilt $M_X = \{1, 2, \ldots\}$ und für die *W–Fkt* erhält man

$$p_X(i) = g_p(i) = (1-p)^{i-1}p \quad \text{für } i = 1, 2, \ldots \ . \tag{3.38}$$

Beweis:
Offensichtlich sind exakt i Versuche notwendig, wenn bei den ersten $i-1$ Versuchen stets das Gegenteil von E eintritt, beim i–ten Mal aber E. Also gilt unter Ausnützung der Unabhängigkeit

$$p_X(i) = P(\underbrace{E^c \cap E^c \cap \ldots E^c}_{i-1} \cap E) = \underbrace{(1-p)(1-p)\cdots(1-p)}_{i-1}\, p = (1-p)^{i-1}p \ . \qquad \triangle$$

Es gilt für die ersten Momente:

$$\mu_X = \frac{1}{p} \tag{3.39}$$

$$\sigma_X^2 = \frac{1-p}{p^2} \ . \tag{3.40}$$

Beweis:
Für den Mittelwert gilt mit $q := 1 - p$

$$\mu_X = \mathrm{E}(X) = \sum_{i=1}^{\infty} i\, p_X(i) = \sum_{i=1}^{\infty} i\,(1-p)^{i-1}p = p \sum_{i=1}^{\infty} i\, q^{i-1} \ .$$

Offensichtlich sind die Summanden die Ableitung von q^i. Differentiation und (unendliche) Summenbildung sind vertauschbar, wenn die Reihe absolut konvergiert. In der Tat ist

$$\sum_{i=1}^{\infty} q^i = \frac{q}{1-q} \leq \infty$$

konvergent, also gilt

$$\sum_{i=1}^{\infty} i\, q^{i-1} = \frac{d}{dq} \sum_{i=1}^{\infty} q^i = \frac{d}{dq}\, \frac{q}{1-q} = \frac{1-q-(-q)}{(1-q)^2} = \frac{1}{(1-q)^2} \ ,$$

woraus unmittelbar die angegebene Beziehung folgt. Für die Varianz verläuft der Beweis analog. $\triangle$

Beispiel 3.12 Im Falle des Würfelversuches mit dem Werfen einer "6" gilt wegen $P(''6'') = 1/6$

$$\mu_X = \frac{1}{p} = \frac{1}{1/6} = 6$$

$$\sigma_X^2 = \frac{1-p}{p^2} = \frac{5/6}{1/36} = 30 \ .$$

$\diamond\diamond\diamond$

3.5.6 Poisson–Verteilung $Po(\mu)$

Eine poissonverteilte *ZG* ergibt sich beispielsweise als *Anzahl* eingehender Anrufe in einer Telefonzentrale während einer bestimmten Zeitspanne oder als *Anzahl* von Labkrautsamen in einer bestimmten Volumseinheit Saatgut. Näherungsweise liegt eine solche Verteilung etwa bei der *Anzahl* von Druckfehlern auf einer Buchseite vor. Im allgemeinen erhält man eine Poissonverteilung für die Anzahl X von Eintritten eines bestimmten Ereignisses E (Anruf, Labkrautsamen, Druckfehler) während einer betrachteten Zeitspanne Δt (allgemein: Einheit; z.B. Volumen, Seitenanzahl), wenn die Wahrscheinlichkeit, dass E genau einmal in dieser Zeit (Einheit) eintritt, mit kleiner werdendem Intervall (Einheit) immer besser proportional zur Intervallänge (Größe der Einheit) wird, also z.B.

$$P(X = 1|\Delta t) = \lambda \cdot \Delta t + o(\Delta t)$$

gilt, und bei kurzen Intervallen E praktisch nie mehr als einmal auftritt, also

$$P(X > 1|\Delta t) = o(\Delta t)$$

erfüllt ist. Daher nennt man die Poissonverteilung auch manchmal die "Verteilung der seltenen Ereignisse". Schließlich muss die Anzahl von Ereigniseintritten in überlappungsfreien Zeitintervallen unabhängig sein.

Somit gilt $M_X = \{0, 1, 2, \ldots\}$ und für die *W–Fkt* erhält man mit $\mu = \lambda \cdot \Delta t$

$$p_X(i) = p_\mu(i) = \frac{\mu^i\, e^{-\mu}}{i!} \quad \text{für } i = 0, 1, 2, \ldots \ . \tag{3.41}$$

Es gilt für die ersten Momente:

$$\mu_X = \sigma_X^2 = \mu \ . \tag{3.42}$$

Damit ist der Parameter μ der Poissonverteilung als *durchschnittliche Anzahl* von Ereigniseintritten *je Einheit* (Zeiteinheit, Volumseinheit usw.) zu interpretieren.

Beweis:
Für den Mittelwert gilt

$$\mu_X = \mathrm{E}(X) = \sum_{i=0}^{\infty} i\, p_X(i) = \sum_{i=1}^{\infty} \frac{i\,\mu^i\, e^{-\mu}}{i!} = \mu \sum_{i=1}^{\infty} \frac{\mu^{i-1}\, e^{-\mu}}{(i-1)!} \ .$$

Ersetzt man in der letzten Summe den Laufindex durch $j := i - 1$, so summiert man dann alle Poissonwahrscheinlichkeiten von 0 bis ∞ auf, was eins ergibt. Damit folgt die behauptete Beziehung. Für die Varianz argumentiert man analog. △

Beispiel 3.13 Wie groß ist die Wahrscheinlichkeit, dass in einer Telefonzentrale, in der im Durchschnitt 12 Anrufe in der Stunde einlangen, innerhalb von 10 Minuten mehr als 2 Anrufe hereinkommen?

Lösung: Die Anzahl von innerhalb einer Stunde eingehenden Anrufen ist poissonverteilt mit $\mu = \lambda \cdot 1 = 12$, also gilt $\lambda = 12$ bei der Zeiteinheit 1 Stunde. Damit ist die Anzahl X von Anrufen innerhalb von 10 $min \mathrel{\widehat{=}} 1/6\ h$ poissonverteilt mit dem Parameter $\mu = 12 \cdot 1/6 = 2$. Damit erhält man

$$\begin{aligned} P(X > 2) &= 1 - P(X \le 2) = 1 - (p_2(0) + p_2(1) + p_2(2)) \\ &= 1 - e^{-2}\,(\frac{\mu^0}{0!} + \frac{\mu^1}{1!} + \frac{\mu^2}{2!}) = 0.323 \ . \end{aligned}$$

◇◇◇

Bemerkung: Ist $\mu > 0$ gegeben, so nähert sich eine $Bi(n, p_n = \mu/n)$–Verteilung mit wachsendem n immer mehr einer Poissonverteilung mit dem Parameter μ:

$$\lim_{n\to\infty} b_{n,\mu/n}(i) = p_\mu(i) \quad \text{für } i = 0, 1, \ldots, n \ . \tag{3.43}$$

Als Faustregel gilt, dass eine $Bi(n, p)$–Verteilung durch eine $Po(np)$–Verteilung approximiert werden kann, wenn $p \leq 0.1$ gilt.

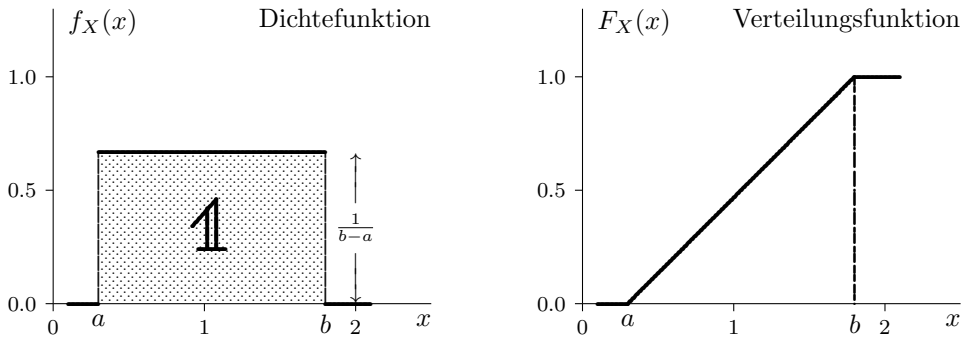
3.6 Stetige Zufallsgrößen (Verteilungen)

Einige der wichtigsten stetigen Verteilungen sind mit dem Wertebereich $M_X \subset \boldsymbol{R}$ der zugeordneten *ZG* X und ihrer *DF* zusammengestellt. Hinzu kommen die wichtigsten Momente und allfällige Anwendungssituationen.

3.6.1 Stetige Gleichverteilung $S(a, b)$

Bei einer im Intervall $(a, b) \subset \boldsymbol{R}$ stetig gleichverteilten *ZG* hängt die Wahrscheinlichkeit für Teilintervalle nur von deren Länge, aber nicht ihrer Lage ab. Damit muss die *DF* auf dem

Abbildung 3.8: *Stetige Gleichverteilung*



Wertebereich $M_X = (a, b)$ konstant sein und wegen der Normierungsbedingung gilt

$$f_X(x) = \begin{cases} \dfrac{1}{b-a} & \text{für } a < x < b \\ 0 & \text{sonst .} \end{cases} \tag{3.44}$$

Die *VF* ergibt sich zu:

$$F_X(x) = \begin{cases} 0 & \text{für } x \leq a \\ \dfrac{x-a}{b-a} & \text{für } a < x < b \\ 1 & \text{für } x \geq b \ . \end{cases} \tag{3.45}$$

Die Abb. 3.8 zeigt *DF* und *VF* der stetigen Gleichverteilung.

Für die ersten Momente erhält man:

$$\mu_X = \frac{a+b}{2} \tag{3.46}$$

$$\sigma_X^2 = \frac{(b-a)^2}{12} \ . \tag{3.47}$$

Beweis:
Für den Mittelwert gilt

$$\begin{aligned}\mu_X &= \int_{-\infty}^{\infty} x \cdot f_X(x)\,dx = \int_a^b \frac{x}{b-a}\,dx \\ &= \frac{x^2}{2(b-a)}\Big|_a^b = \frac{b^2-a^2}{2(b-a)} = \frac{a+b}{2}\ .\end{aligned}$$

Zur Berechnung der Varianz erhält man zunächst

$$\sigma^2_{X;0} = \int_a^b \frac{x^2}{b-a}\,dx = \frac{b^3-a^3}{3(b-a)} = \frac{a^2+ab+b^2}{3}$$

und damit die Varianz als

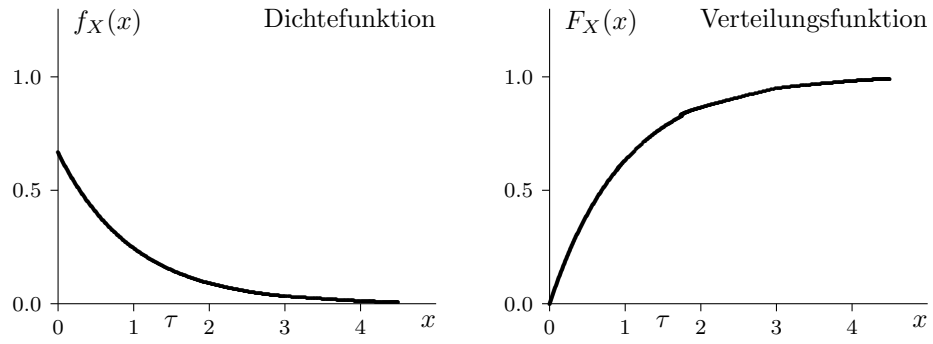
$$\begin{aligned}\sigma_X^2 &= \sigma^2_{X;0} - \mu_X^2 \\ &= \frac{a^2+ab+b^2}{3} - \frac{(a+b)^2}{4} \\ &= \frac{1}{12}(4a^2+4ab+4b^2-3a^2-6ab-3b^2) = \frac{(b-a)^2}{12}\ .\end{aligned}$$

△

3.6.2 Exponentialverteilung $Ex(\tau)$

Exponentialverteilte *ZGen* treten häufig bei der Lebensdauerbeschreibung, und hier insbesondere im Bereich elektronischer Bauteile, auf. Der Wertebereich $M_X = (0, \infty)$ umfasst die

Abbildung 3.9: *Exponentialverteilung*



positiven reellen Zahlen und die *DF* ist definiert durch

$$f_X(x) = \begin{cases} \frac{1}{\tau}e^{-x/\tau} & \text{für } x > 0 \\ 0 & \text{sonst .} \end{cases} \tag{3.48}$$

Die *VF* ergibt sich zu:

$$F_X(x) = \begin{cases} 0 & \text{für } x \le 0 \\ 1 - e^{-x/\tau} & \text{für } x > 0\ ., \end{cases} \tag{3.49}$$

und für die ersten beiden Momente erhält man

$$\mu_X = \tau \qquad \text{und} \qquad \sigma_X^2 = \tau^2\ . \tag{3.50}$$

Die Abb. 3.9 zeigt *DF* und *VF* einer Exponentialverteilung.

Zwischen der Poissonverteilung und der Exponentialverteilung besteht ein wichtiger Zusammenhang, der durch folgende Beziehung erklärt wird:

Poissonverteilung/Exponentialverteilung:

Ist die Anzahl X von Ereigniseintritten in einem Intervall der Länge Δt poissonverteilt mit dem Parameter $\mu = \lambda \cdot \Delta t$, so ist die Zeit Y zwischen zwei derartigen Ereigniseintritten exponentialverteilt mit dem Parameter $\tau = 1/\lambda$.

Beweis:
Die *VF* von Y ergibt sich offensichtlich als

$$\begin{aligned} F_Y(y) &= P(Y \le y) = 1 - P(Y > y) = 1 - P(X = 0 | \Delta t = y) \\ &= 1 - \frac{(\lambda y)^0}{0!} e^{-\lambda y} = 1 - e^{-\lambda y} \ , \end{aligned}$$

weil die Zeit zwischen zwei Ereigniseintritten genau dann größer als y ist, wenn innerhalb des Zeitraumes y das betrachtete Ereignis nicht eintritt. △

3.6.3 Normalverteilung $N(\mu, \sigma^2)$

Die Normalverteilung (manchmal auch *Gaußverteilung* genannt) ist sicherlich die am häufigsten in Theorie und Anwendung anzutreffende Wahrscheinlichkeitsverteilung. Dies ist überwiegend auch gerechtfertigt, wenngleich in manchen Situationen (hauptsächlich in der Praxis) die Verwendung eigentlich nicht zu vertreten wäre.

Der Wertebereich umfasst ganz $\boldsymbol{R}$, obwohl eine normalverteilte *ZG* Werte fast ausschließlich (mit Wahrscheinlichkeit 99.7 %) nur im 3σ–Bereich um den Mittelwert annimmt. Die *DF* ist durch

$$f_X(x) = \frac{1}{\sqrt{2\pi}\,\sigma} e^{-(x-\mu)^2/2\sigma^2} \quad \text{für } -\infty < x < \infty \tag{3.51}$$

definiert. Mittelwert und Varianz sind durch die beiden Parameter gegeben. Es gilt $\mu_X = \mu$ und $\sigma_X^2 = \sigma^2$. Die Abb. 3.10 zeigt die *DF* einiger Normalverteilungen. Der *Standardnormalverteilung* $N(0,1)$ mit Mittelwert 0 und Varianz 1 kommt zentrale Bedeutung in der Statistik zu. Ihre *DF* lautet

$$\varphi(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2} \quad \text{für } -\infty < x < \infty \tag{3.52}$$

und ihre *VF* ergibt sich zu

$$\Phi(x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}}\, e^{-u^2/2}\, du \quad \text{für } -\infty < x < \infty \ . \tag{3.53}$$

Diese liegt auch in tabellierter Form vor (Tab. A1), wobei wegen der offensichtlichen Symmetrie der Normalverteilung die Tabellen üblicherweise nur positive Argumentwerte enthalten. Es gilt nämlich für eine standardnormalverteilte *ZG* Z (siehe Abb.3.11)

$$\Phi(-z) = P(Z \le -z) = P(Z \ge z) = 1 - \Phi(z) \tag{3.54}$$

und damit beispielweise

$$P(-z \le Z \le z) = \Phi(z) - \Phi(-z) = \Phi(z) - (1 - \Phi(z)) = 2\Phi(z) - 1 \ . \tag{3.55}$$

Abbildung 3.10: *Dichtefunktion der Normalverteilung* $N(\mu, \sigma^2)$

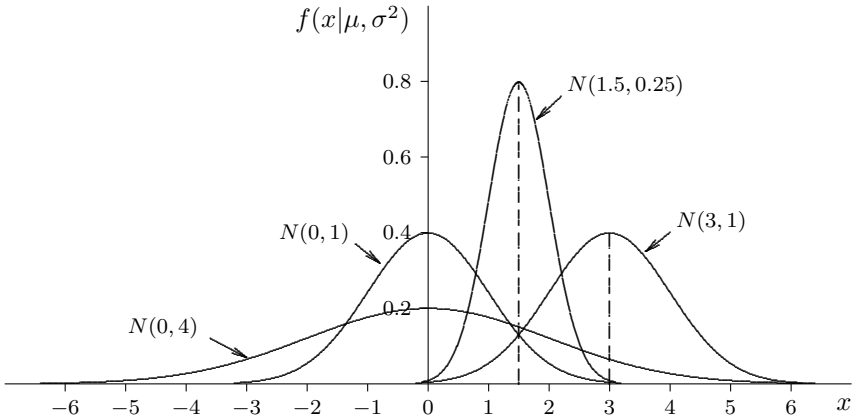
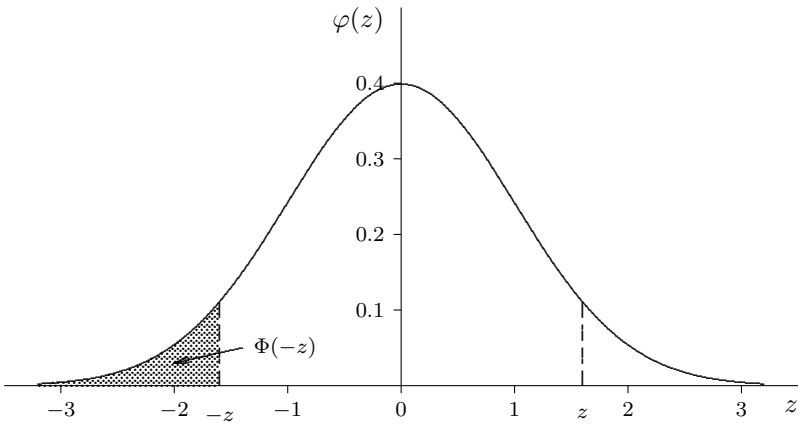


Abbildung 3.11: *Dichtefunktion der Standardnormalverteilung* $N(0, 1)$



Es gilt folgende wichtige Beziehung für normalverteilte *ZGen*:

$$X \text{ vt } N(\mu, \sigma^2) \;\Rightarrow\; aX + b \text{ vt } N(a\mu + b, a^2\sigma^2) \quad \text{für } a, b \in \boldsymbol{R} \;, \tag{3.56}$$

also Lineartransformationen normalverteilter *ZGen* führen wieder zu Normalverteilungen.

Mit obiger Beziehung lässt sich nun die *VF* einer normalverteilten *ZG* X einfach durch die der Standardnormalverteilung angeben. Es gilt wegen (3.56)

$$\frac{X - \mu}{\sigma} \text{ vt } N(0, 1) \tag{3.57}$$

und damit ist

$$F_X(x) = P(X \le x) = P\left(\frac{X - \mu}{\sigma} \le \frac{x - \mu}{\sigma}\right) = \Phi\left(\frac{x - \mu}{\sigma}\right) \;. \tag{3.58}$$

Somit lassen sich Wahrscheinlichkeitsangaben für eine $N(\mu, \sigma^2)$–verteilte *ZG* X stets mit Hilfe der Tab. A1 ableiten. Die Tab. 3.1 enthält einige wichtige Wahrscheinlichkeitswerte für typische Bereiche einer $N(\mu, \sigma^2)$–verteilten Messgröße X.

Tabelle 3.1: *Wahrscheinlichkeiten für normalverteilte Messgrößen*

k	$\Pr(X \le \mu + k\sigma)$ $= \Phi(k)$	$\Pr(\mu - k\sigma \le X \le \mu + k\sigma)$ $= \Phi(k) - \Phi(-k) = 2\Phi(k) - 1$
0	0.5	0
0.67	0.75	0.5
1	0.8413	0.6826
1.28	0.9	0.8
1.64	0.95	0.9
1.96	0.975	0.95
2	0.9772	0.9544
2.33	0.99	0.98
2.58	0.995	0.99
3	0.9987	0.9974

3.6.4 Logarithmische Normalverteilung $LN(\mu, \sigma^2)$

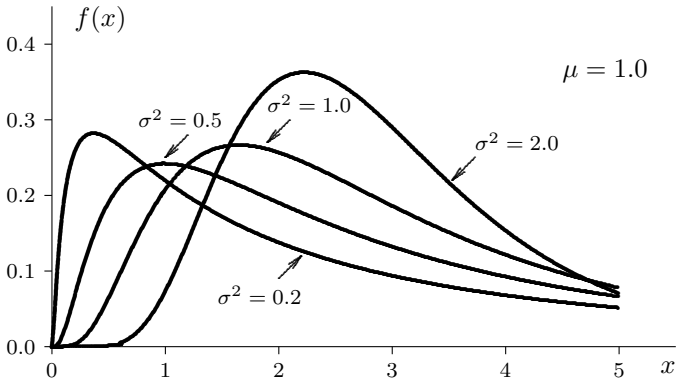
Diese Verteilung hängt – wie der Name schon andeutet – eng mit der Normalverteilung zusammen. Eine *ZG* X ist logarithmisch normalverteilt mit den Parametern μ und σ^2, kurz auch lognormal verteilt genannt, wenn ihr Logarithmus $Y = \ln X$ $N(\mu, \sigma^2)$–verteilt ist. Über diesen Zusammenhang stehen für Wahrscheinlichkeitsangaben alle Möglichkeiten der Normalverteilung offen.

Als Wertebereich erhält man auf Grund obiger Beziehung $M_X = (0, \infty)$ und für die *DF* ergibt sich

$$f_X(x) = \frac{1}{\sqrt{2\pi}\,\sigma x} e^{-(\ln x - \mu)^2/2\sigma^2} \quad \text{für } x > 0 \;, \tag{3.59}$$

Die Abb. 3.12 zeigt die *DF* der Log-Normalverteilung für einige Parameterkombinationen.

Abbildung 3.12: *Dichtefunktion der Lognormalverteilung* $LN(\mu, \sigma^2)$



Für die *VF* folgt

$$F_X(x) = P(X \le x) = P\left(\frac{\ln X - \mu}{\sigma} \le \frac{\ln x - \mu}{\sigma}\right) = \Phi\left(\frac{\ln x - \mu}{\sigma}\right) , \tag{3.60}$$

und für die ersten beiden Momente erhält man

$$\mathrm{E}(X) = e^{\mu+\sigma^2/2} \qquad \text{und} \qquad \mathrm{Var}(X) = e^{\mu+\sigma^2}\left(e^{\sigma^2} - 1\right) \tag{3.61}$$

3.6.5 Chiquadrat–Verteilung χ^2_f

Mit dieser Verteilung beginnt eine Reihe sogenannter *Prüfverteilungen*, deren Bedeutung in der schließenden Statistik begründet ist. Eine χ^2–Verteilung mit f *Freiheitsgraden* ergibt sich als Verteilung von

$$Y = X_1^2 + \cdots + X_f^2 , \tag{3.62}$$

wobei die X_i unabhängige standardnormalverteilte *ZGen* darstellen $(i = 1, \ldots, f)$.

Als Wertebereich erhält man damit wieder die positiven reellen Zahlen $M_X = (0, \infty)$ und als *DF*

$$f_X(x) = \frac{x^{f/2-1}\, e^{-x/2}}{\Gamma(f/2)\, 2^{f/2}} \quad \text{für } x > 0 . \tag{3.63}$$

Abb. 3.13 zeigt die *DF* der χ^2–Verteilung für einige Kombinationen von Freiheitsgraden. Ihre ersten Momente lauten

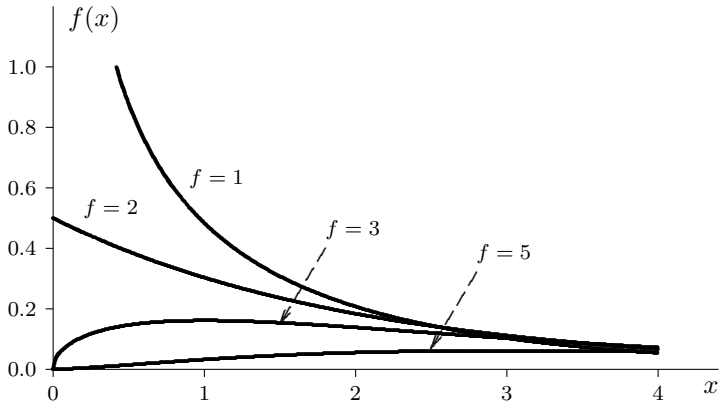
$$\mu_X = f \quad \text{und} \quad \sigma_X^2 = 2f . \tag{3.64}$$

3.6.6 t–Verteilung t_f

Eine t–Verteilung (manchmal auch als *Student*–Verteilung bezeichnet) ergibt sich als Verteilung von

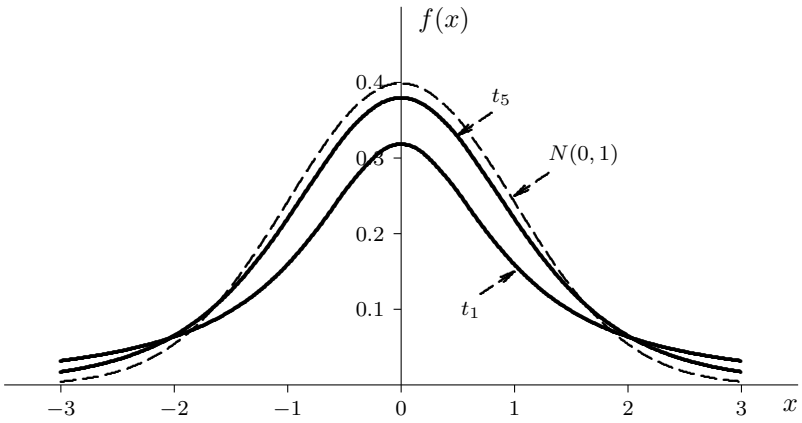
$$Y = \frac{Z}{\sqrt{V/f}} , \tag{3.65}$$

Abbildung 3.13: *Dichtefunktion der χ^2_f–Verteilung*



wobei Z eine standardnormalverteilte ZG und V eine mit f Freiheitsgraden χ^2–verteilte ZG darstellen, wobei die beiden unabhängig sind. Die Abb. 3.14 zeigt die DF der t–Verteilung für einige Freiheitsgrade. Man erkennt dabei, dass für eine wachsende Zahl von Freiheitsgraden die t–Verteilung gegen die $N(0,1)$–Standardnormalverteilung strebt.

Abbildung 3.14: *Dichtefunktion der t_f–Verteilung*



Da die t–Verteilung offensichtlich symmetrisch ist, gilt $\mu_X = 0$ (falls $f > 1$) und für die Varianz erhält man $\sigma_X^2 = f/(f-2)$ (falls $f > 2$).

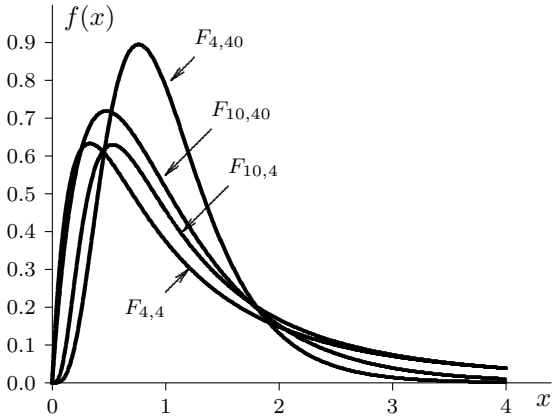
3.6.7 F–Verteilung F_{f_1,f_2}

Eine F–Verteilung, benannt nach dem bekannten Statistiker Sir R.A. Fisher, ergibt sich als Verteilung von

$$Y = \frac{V/f_1}{W/f_2} \ , \tag{3.66}$$

wobei V und W unabhängige χ^2–verteilte *ZGen* mit f_1 bzw. f_2 Freiheitsgraden darstellen. Die Abb. 3.15 zeigt die *DF* der F–Verteilung für einige Kombinationen von Freiheitsgraden.

Abbildung 3.15: F_{f_1,f_2}*–Verteilung*



Für die Momente der F–Verteilung erhält man schließlich

$$\mu_X = \frac{f_2}{f_2 - 2} \quad (\text{falls } f_2 > 2) \tag{3.67}$$

$$\sigma_X^2 = \frac{2\, f_2^2 (f_1 + f_2 - 2)}{f_1 (f_2 - 2)^2 (f_2 - 4)} \quad (\text{falls } f_2 > 4)\ . \tag{3.68}$$

3.7 QQ-Diagramme

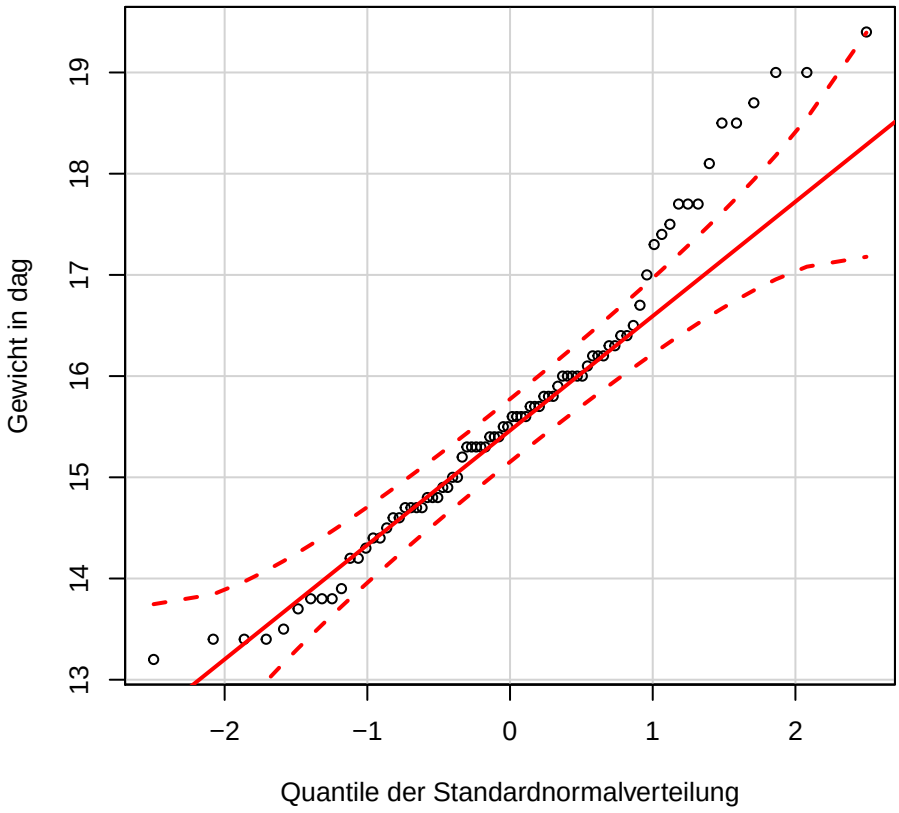
Durch die vordergründig ähnliche Form von Verteilungsfunktionen ist es häufig nicht leicht, (empirische) Verteilungen mit ihnen zu vergleichen. Die Verteilungsfunktionen F verschiedener parametrischer Verteilungen haben unterschiedliche Definitionsbereiche (ganz $\boldsymbol{R}$, $\boldsymbol{R}^+$, ...), wegen $0 \le F(x) \le 1$ ist der Definitionsbereich ihrer Umkehrabbildung F^{-1} (der Quantilsfunktion) jedoch immer das Intervall $[0, 1]$. Unter anderem aus diesem Grund hat es sich bewährt, nicht empirische Verteilungen von Stichproben mit theoretischen Verteilungen zu vergleichen, sondern ihre Umkehrabbildung, also die Quantile.

Zieht man eine Stichprobe aus einer Standardnormalverteilung, dann sollten die empirischen Quantile der Stichprobe in der Nähe der theoretischen Quantile der Standardnormalverteilung liegen. Stellt sich noch die Frage, welche Quantile der Stichprobe mit ihren theoretischen Gegenstücken verglichen werden. Für $\alpha = i/n$, $i = 1, \ldots, n-1$ ist $n\alpha$ aus der Formel zur Quantilsberechnung auf Seite 17 trivialerweise immer ganzzahlig, sodass für die Quantile q_α jeder Wert aus den Intervallen $[x_{(n\alpha)}, x_{(n\alpha+1)}]$ zulässig ist. Nimmt man jeweils die linke Intervallgrenze ergeben sich als Quantile einfach die geordnete Stichprobe $x_{(1)}, \ldots, x_{(n-1)}$.

Folgt die Stichprobe einer vorgegebenen Verteilung F, sollten die Rangstatistiken $x_{(i)}$ in der Nähe der theoretischen Quantile $F^{-1}(i/n)$ liegen. Zur optischen Überprüfung bildet man die Wertepaare

$$\big(F^{-1}(i/n), x_{(i)}\big)$$

Abbildung 3.16: *QQ-Diagramm*



und zeichet damit ein Streudiagramm, das sogenannte Quantil-Quantil-Diagramm, kurz *QQ-Diagramm*. Je näher die abgetragenen Punkte an der Winkelhalbierenden (45 Grad-Gerade durch den Ursprung) liegen, desto besser passt die Stichprobe zur Verteilung.

Folgt die Stichprobe einer *Normalverteilung* mit beliebigem μ und σ, dann ist

$$z_i = \frac{x_i - \mu}{\sigma}$$

standardnormalverteilt. Es gibt nun zwei Möglichkeiten ein QQ-Diagramm zu zeichnen:

1. Vergleich der Originaldaten x_i mit den theoretischen Quantilen der $N(\mu, \sigma^2)$.
2. Vergleich der standardisierten Daten z_i mit den theoretischen Quantilen der Standardnormalverteilung $N(0, 1)$.

In beiden Fällen sollten die Punkte im QQ-Diagramm entlang der Winkelhalbierenden liegen, falls die Stichprobe normalverteilt ist. Jedoch muss in beiden Fällen für die unbekannten Parameter μ und σ ein Schätzwert verwendet werden. Da die Standardisierung $z_i = (x_i - \mu)/\sigma$ eine Lineartransformation ist, bleiben Geraden erhalten. Daher kann man für *beliebige* Normalverteilungen die empirischen Quantile einer Stichprobe mit jenen der Standardnormalverteilung

vergleichen und bekommt wieder eine Gerade (allerdings i. Allg. nicht die Winkelhabierende). Dies funktioniert auch bei vielen anderen Verteilungsfamilien, z.B. können beliebige stetige Gleichverteilungen mit der Gleichverteilung auf $[0, 1]$ verglichen werden.

Beispiel 3.14 Die Abb. 3.16 zeigt ein QQ-Diagramm der Daten aus Bsp. 1.1. Im unteren und mittleren Bereich liegen die Punkte annähernd auf einer Geraden, im oberen Bereich weichen sie systematisch davon ab. Dies passt zu der von uns bereits in Abschnitt 2.4.4 diagnostizierten Schiefe der Daten.

◇◇◇

3.8 Mehrdimensionale Zufallsgrößen

Die meisten in Natur, Technik und Gesellschaft zu beobachtenden Phänomene sind *multivariater* Art (vgl. Beispiel 1.2: Studentenbefragung). Je mehr man beobachtet, desto größer ist in der Regel der Informationsgewinn. Das gilt zwar in erster Linie für die *Anzahl* von Beobachtungen (Stichprobenumfang), aber in gewissem Sinn auch für die *"Breite"* der Beobachtung, d.h. für die Zahl gleichzeitig beobachteter Merkmale. Ein wesentlicher Gesichtspunkt bei der Modellierung derartiger Phänomene ist die Beschreibung des Zusammenhanges von Merkmalen.

Aus der Definition der k–dimensionalen *ZG* $\mathbf{X} = (X_1, X_2, \ldots, X_k)$ in Abschnitt 3.3 folgt zunächst, dass die einzelnen Komponenten X_i $(i = 1, \ldots, k)$ eindimensionale *ZGen* darstellen. Außerdem ist man mit dieser Definition in der Lage, Wahrscheinlichkeitsaussagen für $\mathbf{X}$ zu treffen. Am wichtigsten ist dabei die Beschreibung der Wahrscheinlichkeitsverteilung von $\mathbf{X}$ durch die Verteilungsfunktion *(VF)*:

$$F_{\mathbf{X}}(x_1, x_2, \ldots, x_k) = P(X_1 \leq x_1 \wedge X_2 \leq x_2 \wedge \cdots \wedge X_k \leq x_k) \tag{3.69}$$

mit $x_i \in \boldsymbol{R}$ für $i = 1, \ldots, k$. Analog dem eindimensionalen Fall lässt sie sich für *jeden* Typ von Zufallsgrößen angeben.

Definition: Unter der *Randverteilung* von X_i der gemeinsamen Verteilung einer k–dimensionalen *ZG* $\mathbf{X} = (X_1, X_2, \ldots, X_k)$ versteht man die sich unter Nichtbeachtung aller anderen Komponenten ergebende *W-Vt.* Ihre *VF* ist durch

$$F_{X_i}(x_i) = P(X_i \leq x_i) = \lim_{\substack{x_l \to \infty \\ l = 1, \ldots, k \\ l \neq i}} F_{\mathbf{X}}(x_1, \ldots, x_i, \ldots, x_k) \tag{3.70}$$

gegeben.

Die in den beiden Folgeabschnitten beschriebenen Verteilungstypen stellen Sonderfälle dar, die in der Praxis dadurch erreicht werden, dass man nicht dazu passende Komponenten einfach nicht berücksichtigt. Im übrigen stellen Mischtypen die am weitesten verbreitete Form multivariater *ZGen* dar. Ihre Beschreibung ist aber aufwendiger und im allgemeinen auch nicht notwendig.

3.8.1 Diskrete mehrdimensionale *ZGen*

Analog dem eindimensionalen Fall lässt sich bei diskreten multivariaten *ZGen* die Wahrscheinlichkeitsverteilung sehr einfach beschreiben (vgl. Abschnitt 3.3).

Definition: Eine k–dimensionale *ZG* $\mathbf{X}$ heißt *diskret*, wenn es höchstens abzählbar unendlich viele $\mathbf{x} \in \boldsymbol{R}^k$ gibt, die $\mathbf{X}$ annehmen kann, wenn also mit $M_{\mathbf{X}} = \{\mathbf{x}_1, \mathbf{x}_2, \ldots\} \subset \boldsymbol{R}^k$ und

$$p_{\mathbf{X}}(\mathbf{x}_l) = P(\mathbf{X} = \mathbf{x}_l) = P((X_1 = x_{l1}) \wedge (X_2 = x_{l2}) \wedge \cdots \wedge (X_k = x_{lk}))$$

gilt

$$\sum_{\mathbf{x}_l \in M_{\mathbf{X}}} p_{\mathbf{X}}(\mathbf{x}_l) = 1 \ .$$

Die Funktion $p_{\mathbf{X}}$ heißt *Wahrscheinlichkeitsfunktion (W-Fkt).*

Damit lässt sich die Randverteilung einer Komponente, etwa X_i, einfach durch deren *W-Fkt*

$$p_{X_i}(x_i) = \sum_{\substack{(u_1, \ldots, u_k) \in M_{\mathbf{X}} \\ u_i = x_i}} p_{\mathbf{X}}(u_1, \ldots, u_k) \tag{3.71}$$

beschreiben.

Die beiden folgenden Verteilungen stellen häufig anzutreffende Beispiele mehrdimensionaler diskreter *ZGen* bzw. Verteilungen dar.

Multivariate diskrete Gleichverteilung:

Wie im eindimensionalen Fall liegt diese Verteilung dann vor, wenn es

1) *endlich* viele Werte $\mathbf{x} \in M_{\mathbf{X}}$ gibt, etwa $M_{\mathbf{X}} = \{\mathbf{x}_1, \ldots, \mathbf{x}_m\}$, und
2) jedem dieser Werte dieselbe Wahrscheinlichkeit zukommt, also

$$p_{\mathbf{X}}(\mathbf{x}_l) = P(\mathbf{X} = \mathbf{x}_l) = \frac{1}{m} \quad \text{für } l = 1, \ldots, m$$

gilt.

Beispiel 3.15 Beim Würfeln mit zwei Würfeln steht X_1 für die Augenzahl des ersten, X_2 für die des zweiten Würfels. Man erhält $M_{\mathbf{X}} = \{(1,1), (1,2), \ldots, (6,5), (6,6)\}$ mit $m = |M_{\mathbf{X}}| = 36$. Es gilt beispielsweise

$$p_{\mathbf{X}}(3,5) = P(X_1 = 3, X_2 = 5) = \frac{1}{36} \ .$$

◇◇◇

Multinomialverteilung:

Diese Verteilung stellt die Verallgemeinerung der Binomialverteilung dar, wenn man sie als

zweidimensionale *W-Vt* auffasst. Eine binomialverteilte *ZG* X_A beschreibt die Häufigkeit, mit der ein bestimmtes Ereignis A während n unabhängig durchgeführten Versuchen eintritt. Eine zweite – hier direkt abhängige – *ZG* X_{A^c} gibt die Anzahl von dabei beobachteten Fehlversuchen an, also $X_{A^c} = n - X_A$. Für die zweidimensionale *ZG* $\mathbf{X} = (X_A, X_{A^c})$ ergibt sich die folgende *W-Fkt* (vgl. Abschnitt 3.6)

$$p_{\mathbf{X}}(l, n-l) = \frac{n!}{l!(n-l)!}\, p^l (1-p)^{n-l} \quad \text{für } 0 \le l \le n$$

mit $p = P(A)$ und demnach $1 - p = P(A^c)$.

Betrachtet man für einen Versuch die einander paarweise ausschließenden Ereignisse $A_1, A_2, \ldots, A_k$ mit

$$A_1 \cup A_2 \cup \cdots \cup A_k = M$$

(diese k Ereignisse bilden also eine *Zerlegung* des Merkmalraumes M) und

$$p_i = P(A_i) \quad \text{für } i = 1, \ldots, k\ ,$$

sodass also

$$\sum_{i=1}^{k} p_i = 1$$

gilt, und bezeichnet X_i die Häufigkeit des Ereignisses A_i bei n unabhängigen Versuchen $(i = 1, \ldots, k)$, dann wird die *W-Vt* von $\mathbf{X} = (X_1, X_2, \ldots, X_k)$ durch die *W-Fkt*

$$p_{\mathbf{X}}(x_1, x_2, \ldots, x_k) = \frac{n!}{x_1! x_2! \cdots x_k!}\, p_1^{x_1} p_2^{x_2} \cdots p_k^{x_k} \tag{3.72}$$

für $0 \le x_i \le n$ $(i = 1, \ldots, k)$ und $x_1 + x_2 + \cdots + x_k = n$ beschrieben. Sie wird als k-dimensionale *Multinomialverteilung* $M_{n;p_1,p_2,\ldots,p_k}$ bezeichnet.

Für die Momente dieser Verteilung (siehe Unterabschnitt 4) gilt unter Beachtung, dass jede Komponente für sich binomialverteilt ist:

$$\begin{aligned} \mu_i &= \mathrm{E}(X_i) = np_i \\ \sigma_i^2 &= \mathrm{Var}(X_i) = np_i(1-p_i) \\ \sigma_{ij} &= \mathrm{Cov}(X_i, X_j) = -np_ip_j \end{aligned}$$

für $1 \le i \ne j \le k$.

3.8.2 Stetige mehrdimensionale *ZGen*

Analog zum eindimensionalen Fall lassen sich auch mehrdimensionale *ZGen* durch ihre Dichtefunktion charakterisieren.

Definition: Eine k–dimensionale *ZG* $\mathbf{X} = (X_1, X_2, \ldots, X_k)$ heißt *stetig* (oder *kontinuierlich*), falls es eine *nichtnegative* Funktion $f_{\mathbf{X}} : \boldsymbol{R}^k \to \boldsymbol{R}_+$ gibt, sodass für beliebige $-\infty < a_i \le b_i < \infty$ $(i = 1, \ldots, k)$ gilt

$$P(a_i < X_i \le b_i, \quad i = 1, \ldots, k) = \int_{a_1}^{b_1} \cdots \int_{a_k}^{b_k} f_{\mathbf{X}}(x_1, x_2, \ldots, x_k)\, dx_1 \ldots dx_k \tag{3.73}$$

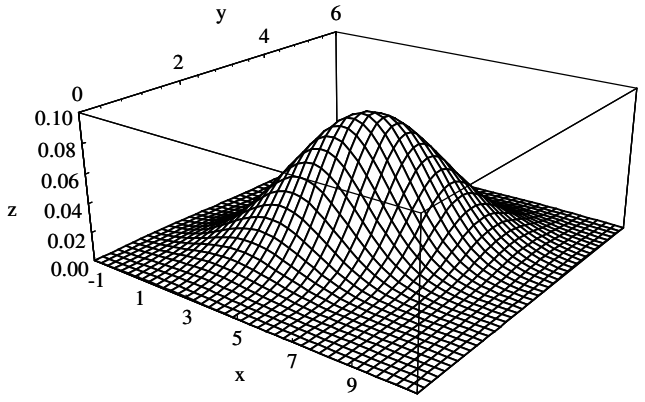
Die Randverteilung einer Komponente X_i lässt sich durch ihre Dichtefunktion

$$f_{X_i}(x_i) = \underbrace{\int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty}}_{k-1} f_{\mathbf{X}}(u_1, \ldots, u_{i-1}, x_i, u_{i+1}, \ldots, u_k)\, du_1 \ldots du_{i-1} du_{i+1} \ldots du_k \quad (3.74)$$

beschreiben. Diese wird demnach auch *Randdichte* genannt.

Die wichtigste mehrdimensionale stetige *W-Vt* ist wohl die *multivariate Normalverteilung*, die durch ihre ersten und zweiten (auch gemischte) Momente eindeutig beschrieben ist. In diesem Fall sind die einzelnen Komponenten eindimensional normalverteilt mit den sich aus den oben erwähnten Momenten ergebenden Parametern.

Abbildung 3.17: *2–dimensionale Normalverteilung*



3.8.3 Unabhängigkeit

In manchen Fällen hat eine *ZG* überhaupt keinen Einfluss auf eine andere und sie werden dann auch als unabhängig bezeichnet. Die formale Definition greift auf den Unabhängigkeitsbegriff bei Ereignissen zurück.

Definition: Die *ZGen* $X_1, X_2, \ldots, X_k$ heißen *unabhängig*, wenn beliebige durch die *ZGen* X_i beschreibbare Ereignisse unabhängig sind. Das ist gleichbedeutend damit, dass für die gemeinsame *VF* von $\mathbf{X}$ gilt:

$$F_{\mathbf{X}}(x_1, x_2, \ldots, x_k) = \prod_{i=1}^{k} F_{X_i}(x_i) \quad \text{für } (x_1, x_2, \ldots, x_k) \in \boldsymbol{R}^k, \quad (3.75)$$

wobei F_{X_i} für die *VF* der *ZG* X_i steht $(i = 1, \ldots, k)$.

Bei unabhängigen diskreten *ZGen* gilt daher für die gemeinsame *W-Fkt*

$$p_{\mathbf{X}}(x_1, x_2, \ldots, x_k) = \prod_{i=1}^{k} p_{X_i}(x_i) \quad \text{für } (x_1, x_2, \ldots, x_k) \in M_{\mathbf{X}} \quad (3.76)$$

mit den eindimensionalen *W-Fkten* p_{X_i} $(i = 1, \ldots, k)$ und im stetigen Fall ergibt sich die gemeinsame Dichtefunktion als

$$f_{\mathbf{X}}(x_1, x_2, \ldots, x_k) = \prod_{i=1}^{k} f_{X_i}(x_i) \quad \text{für } (x_1, x_2, \ldots, x_k) \in \boldsymbol{R}^k, \tag{3.77}$$

mit den eindimensionalen *DFen* f_{X_i}. Diese Eigenschaften charakterisieren darüberhinaus unabhängige diskrete bzw. stetige *ZGen*.

Beispiel 3.16 Es bezeichnet X_1 die Summe und X_2 die (absolute) Differenz der Augenzahlen beim Würfeln mit zwei Würfeln. Offensichtlich erhält man unter Ausnützung der klassischen

Tabelle 3.2: *Augensumme* X_1 *und Augendifferenz* X_2

x_2	x_1 2	3	4	5	6	7	8	9	10	11	12	$p_{X_2}(x_2)$
0	$\frac{1}{36}$	0	$\frac{1}{36}$	0	$\frac{1}{36}$	0	$\frac{1}{36}$	0	$\frac{1}{36}$	0	$\frac{1}{36}$	$\frac{6}{36}$
1	0	$\frac{2}{36}$	0	$\frac{2}{36}$	0	$\frac{2}{36}$	0	$\frac{2}{36}$	0	$\frac{2}{36}$	0	$\frac{10}{36}$
2	0	0	$\frac{2}{36}$	0	$\frac{2}{36}$	0	$\frac{2}{36}$	0	$\frac{2}{36}$	0	0	$\frac{8}{36}$
3	0	0	0	$\frac{2}{36}$	0	$\frac{2}{36}$	0	$\frac{2}{36}$	0	0	0	$\frac{6}{36}$
4	0	0	0	0	$\frac{2}{36}$	0	$\frac{2}{36}$	0	0	0	0	$\frac{4}{36}$
5	0	0	0	0	0	$\frac{2}{36}$	0	0	0	0	0	$\frac{2}{36}$
$p_{X_1}(x_1)$	$\frac{1}{36}$	$\frac{2}{36}$	$\frac{3}{36}$	$\frac{4}{36}$	$\frac{5}{36}$	$\frac{6}{36}$	$\frac{5}{36}$	$\frac{4}{36}$	$\frac{3}{36}$	$\frac{2}{36}$	$\frac{1}{36}$	$\sum = 1$

Wahrscheinlichkeitsdefinition die in Tab. 3.2 enthaltene gemeinsame *W-Fkt*, sowie die ebenfalls angegebenen *W-Fkten* für die Randverteilungen. Nun gilt aber beispielsweise

$$p_{(X_1,X_2)}(7,3) = \frac{2}{36} \neq \frac{1}{6} \times \frac{1}{6} = p_{X_1}(7) \times p_{X_2}(3) \ ,$$

und damit sind die beiden *ZGen* abhängig.

◇◇◇

3.8.4 Momente mehrdimensionaler *ZGen*

Zunächst lassen sich für alle Komponenten X_i einer multivariaten *ZG* $\mathbf{X} = (X_1, X_2, \ldots, X_k)$ die üblichen Momente, wie sie in Abschnitt 3.4 beschrieben sind, definieren. Damit erhält man u.a. den *Mittelwertsvektor* $\mu = (\mu_1, \mu_2, \ldots, \mu_k)$ mit $\mu_i = \mathrm{E}(X_i)$ für $i = 1, \ldots, k$.

Eine echte Neuerung im Vergleich zu eindimensionalen *ZGen* sind *gemischte Momente* der Form

$$\mathrm{E}(X_1^{m_1} X_2^{m_2} \cdots X_k^{m_k}) \ ,$$

wobei $m = m_1 + m_2 + \cdots + m_k$ die *Ordnung* des Momentes definiert.

Unabhängigkeit und gemischte Momente:

Für gemischte Momente unabhängiger *ZGen* $X_1, X_2, \ldots, X_k$ gilt

$$\mathrm{E}(X_1^{m_1} X_2^{m_2} \cdots X_k^{m_k}) = \mathrm{E}(X_1^{m_1}) \times \mathrm{E}(X_2^{m_2}) \times \cdots \times \mathrm{E}(X_k^{m_k}) \ . \tag{3.78}$$

Beweis:
Für stetige *ZGen* erhält man wegen (3.77)

$$\begin{aligned}
&\mathrm{E}(X_1^{m_1} X_2^{m_2} \cdots X_k^{m_k}) \\
&= \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} x_1^{m_1} x_2^{m_2} \cdots x_k^{m_k} f_{\mathbf{X}}(x_1, x_2, \ldots, x_k) dx_1 dx_2 \cdots dx_k \\
&= \int_{-\infty}^{\infty} x_1^{m_1} f_{X_1}(x_1) dx_1 \int_{-\infty}^{\infty} x_2^{m_2} f_{X_2}(x_2) dx_2 \ \cdots \ \int_{-\infty}^{\infty} x_k^{m_k} f_{X_k}(x_k) dx_k \\
&= \mathrm{E}(X_1^{m_1}) \times \mathrm{E}(X_2^{m_2}) \times \cdots \times \mathrm{E}(X_k^{m_k}) \ .
\end{aligned}$$

Für diskrete *ZGen* verläuft der Nachweis analog. △

Zu den wichtigsten (gemischten) Momenten zählen die der Ordnung zwei. Für eindimensionale *ZGen* ist dies die Varianz, für den echt gemischten Fall erklärt dies folgende

Definition: Unter der *Kovarianz* $\sigma_{X,Y}$ zweier *ZGen* X und Y oder zweier Komponenten einer höherdimensionalen *ZG* versteht man das *zentrierte, zweite gemischte Moment*

$$\sigma_{X,Y} = \mathrm{Cov}(X, Y) = \mathrm{E}((X - \mu_X)(Y - \mu_Y)) \ . \tag{3.79}$$

Es gilt offensichtlich der folgende, oft nützlich einzusetzende

Verschiebungssatz für Kovarianzen:

$$\sigma_{X,Y} = \mathrm{E}(XY) - \mu_X \mu_Y \ . \tag{3.80}$$

Beweis:
Aus der Definition der Kovarianz folgt

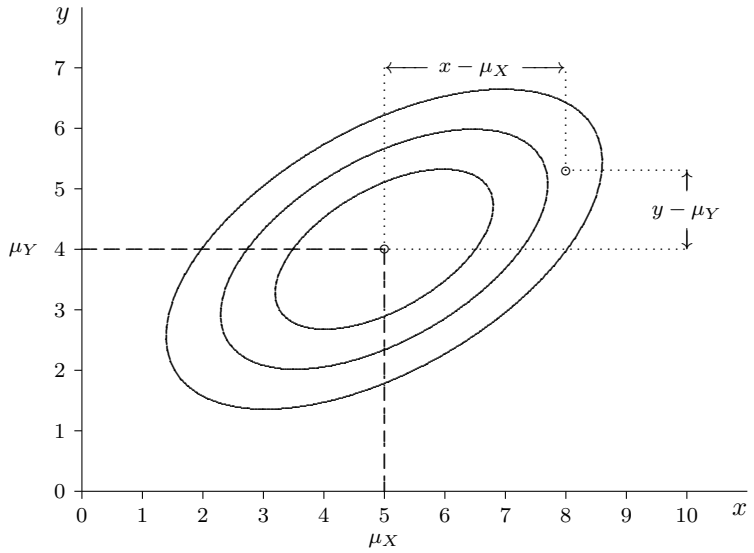
$$\begin{aligned}
\sigma_{X,Y} &= \mathrm{E}((X - \mu_X)(Y - \mu_Y)) \\
&= \mathrm{E}((XY - \mu_X Y - \mu_Y X + \mu_X \mu_Y) \\
&= \mathrm{E}(XY) - \mu_X \mathrm{E}(Y) - \mu_Y \mathrm{E}(X) + \mu_X \mu_Y \\
&= \mathrm{E}(XY) - \mu_X \mu_Y \ .
\end{aligned}$$

△

Die Interpretation der Kovarianz wird mit Hilfe der Abb. 3.18 verständlich. Verläuft die *gemeinsame* Verteilung von X und Y grob in positiver Richtung, so entsprechen relativ häufig größere x–Werte auch größeren y–Werte bzw. kleinere x–Werte eher kleineren y–Werten. Seltener treffen entgegengesetzte Konstellationen zu. Daher sind Produkte der Form $(x-\mu_X)(y-\mu_Y)$ häufiger positiv ("plus mal plus" oder "minus mal minus") als negativ ("plus mal minus" oder umgekehrt) und somit im Durchschnitt (Erwartungswert!) *positiv*. Umkehrt verläuft die Argumentation, wenn die gemeinsame Verteilung einen negativ orientierten Verlauf aufweist.

Zusammenhang Kovarianz/Varianz:

Abbildung 3.18: *Zusammenhang und Kovarianz*



Es gilt

$$|\sigma_{X,Y}| \leq \sigma_X \sigma_Y \tag{3.81}$$

als Folgerung der sogenannten *Schwarz'schen Ungleichung.*

Offensichtlich wird der Betrag der Kovarianz von den Standardabweichungen der beteiligten *ZGen* beeinflusst, wodurch die Bedeutung der Kovarianz für die Beurteilung des Zusammenhanges von zwei Merkmalen gemindert wird. Einen Ausweg bietet folgende

Definition: Die durch Normierung der Kovarianz gewonnene Größe

$$\rho_{X,Y} = \mathrm{E}\left(\left(\frac{X-\mu_X}{\sigma_X}\right)\left(\frac{Y-\mu_Y}{\sigma_Y}\right)\right) = \frac{\sigma_{X,Y}}{\sigma_X \sigma_Y} \tag{3.82}$$

nennt man *Korrelationskoeffizient* zwischen X und Y.

Korrelationskoeffizient:

Für den Korrelationskoeffizienten zweier *ZGen* X und Y gelten folgende Aussagen:

KK 1) Er stellt ein Maß für den *linearen* Zusammenhang zwischen X und Y dar;

KK 2) $|\rho_{X,Y}| \leq 1$;

KK 3) Falls $\rho_{X,Y} = 1$ gilt, existiert ein gesicherter linearer Zusammenhang der Form $Y = a + bX$ mit $b > 0$; mit der Kenntnis von X weiß man somit über Y sicher Bescheid. Falls $\rho_{X,Y} = -1$ gilt, ergibt sich der analoge Sachverhalt, bloß mit $b < 0$;

KK 4) Falls $\rho_{X,Y} = 0$ besteht sicher *kein linearer* Zusammenhang. Die beiden *ZGen* können aber trotzdem einen möglicherweise sehr starken (nichtlinearen) Zusammenhang aufweisen.

KK 5) Falls X und Y unabhängig sind, gilt wegen

$$\sigma_{X,Y} = \mathrm{E}((X - \mu_X)(Y - \mu_Y)) = \mathrm{E}(X - \mu_X)\mathrm{E}(Y - \mu_Y) = 0$$

auch

$$\rho_{X,Y} = 0 \ ,$$

also sind X und Y unkorreliert. Die Unabhängigkeit ist somit eine stärkere Eigenschaft als die Unkorreliertheit.

KK 6) Falls X und Y gemeinsam normalverteilt sind, folgt aus der Unkorreliertheit auch die Unabhängigkeit der beiden *ZGen*. Es sind also in diesem Fall beide Eigenschaften äquivalent.

Zwei spezielle Formen des Korrelationskoeffizienten stellen der *partielle* und der *multiple Korrelationskoeffizient* dar.

Der partielle Korrelationskoeffizient

Mitunter kommt es vor, dass die Korrelation zwischen zwei Merkmalen X und Y deshalb groß ist, weil beide durch eine dritte Zufallsgröße U beeinflusst werden. Man spricht dann von einer *Scheinkorrelation* zwischen X und Y. Bekannte Beispiele sind etwa der "Zusammenhang" zwischen der Geburtenrate in einem Gebiet und der Häufigkeit von Störchen (guter, wenig feuchter Boden bedeutet in der Regel Wohlstand; dieser drückt meist auf die Geburtenrate) oder die "Korrelation" zwischen der Häufigkeit von Waldbränden und dem Ernteertrag (heiße Witterung erhöht das Risiko von Waldbränden und drückt auf den Ernteertrag).

Definition: Unter dem *partiellen Korrelationskoeffizienten* $\rho_{(X,Y)|U}$ (unter *Konstanthaltung* von U) versteht man die durch

$$\rho_{(X,Y)|U} = \frac{\rho_{X,Y} - \rho_{X,U}\rho_{Y,U}}{\sqrt{1 - \rho_{X,U}^2}\sqrt{1 - \rho_{Y,U}^2}} \tag{3.83}$$

definierte Größe.

Bemerkung: Der partielle Korrelationskoeffizient zwischen X und Y unter Konstanthaltung von U stellt nichts anderes als den gewöhnlichen Korrelationskoeffizienten zwischen den beiden Restgrößen dar, die entstehen, wenn man einerseits X und andererseits Y bestmöglich *linear* durch U zu erklären versucht.

Der multiple Korrelationskoeffizient

Zur Beschreibung der (linearen) Abhängigkeit eines Merkmals Y von einer Reihe anderer Zufallsgrößen $X_1, \ldots, X_k$ dient der Korrelationskoeffizient zwischen Y und der am besten passenden Linearkombination $Z = a_1 X_1 + \cdots + a_k X_k$.

Dieser *multiple Korrelationskoeffizient* ergibt sich im einfachsten Fall mit bloß zwei erklärenden Merkmalen X_1 und X_2 als

$$\rho_{Y;(X_1,X_2)} = \sqrt{\frac{\rho^2_{Y,X_1} + \rho^2_{Y,X_2} - 2\,\rho_{Y,X_1}\rho_{Y,X_2}\rho_{X_1,X_2}}{1 - \rho^2_{X_1,X_2}}} \; . \tag{3.84}$$

Die Größe

$$B_{Y,(X_1,X_2)} = \rho^2_{Y,(X_1,X_2)} \tag{3.85}$$

wird als (theoretisches) *Bestimmtheitsmaß* bezeichnet. Es gibt die Güte der (linearen) Beschreibbarkeit von Y durch die beiden Merkmale X_1 und X_2 an.

Der multiple Korrelationskoeffizient und das Bestimmtheitsmaß lassen sich in analoger Form für jede beliebige Anzahl erklärender Merkmale angeben.

Definition: Unter der *Kovarianzmatrix* einer gegebenen k–dimensionalen *ZG* $\mathbf{X} = (X_1, X_2, \ldots, X_k)$ versteht man die Zusammenstellung der Varianzen der einzelnen Komponenten und der Kovarianzen aller möglichen Paare von Komponenten in Matrixform:

$$\mathbf{\Sigma_X} = \begin{pmatrix} \sigma_1^2 & \sigma_{12} & \cdots & \sigma_{1k} \\ \sigma_{21} & \sigma_2^2 & \cdots & \sigma_{2k} \\ \vdots & \vdots & \vdots & \vdots \\ \sigma_{k1} & \sigma_{k2} & \cdots & \sigma_k^2 \end{pmatrix} \; . \tag{3.86}$$

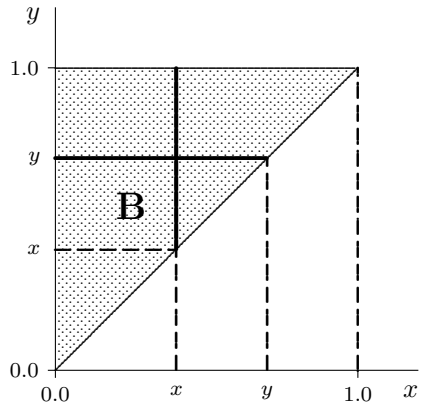
Wählt man anstelle der Kovarianzen die Korrelationskoeffizienten, erhält man in analoger Form die *Korrelationsmatrix*:

$$\mathbf{R_X} = \begin{pmatrix} 1 & \rho_{12} & \cdots & \rho_{1k} \\ \rho_{21} & 1 & \cdots & \rho_{2k} \\ \vdots & \vdots & \vdots & \vdots \\ \rho_{k1} & \rho_{k2} & \cdots & 1 \end{pmatrix} \; . \tag{3.87}$$

Da offensichtlich $\sigma_{X,Y} = \sigma_{Y,X}$ und damit auch $\rho_{X,Y} = \rho_{Y,X}$ gilt, handelt es sich in beiden Fällen um *symmetrische* Matrizen, die darüberhinaus wegen (3.81) sogar *positiv semidefinit* sind.

Beispiel 3.17 In einer Wagenladung Schüttgut finden sich Verunreinigungen durch Stoff A im Ausmaß von x % und durch Stoff B mit einem Anteil von y %. Die gemeinsame Verteilung der Mengen ist in Abb. 3.19 dargestellt. Sie soll im schraffierten Bereich konzentriert und dort

Abbildung 3.19: *Dichtefunktion von X und Y*



gleichmäßig sein. Wie lauten die Kenngrößen für X und Y? Wie hängen X und Y zusammen?

Lösung:
Zunächst benötigt man die gemeinsame *DF* $f(x,y)$. Da

$$1 = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(x,y)\,dx\,dy = \iint_B c\,dx\,dy$$

gelten muss, ergibt sich $c = 2$. Für die Randdichten erhält man

$$f_X(x) = \int_{-\infty}^{\infty} f(x,v)\,dv = \int_x^1 2\,dv = 2(1-x) \quad \text{für } 0 < x < 1$$

und

$$f_Y(y) = \int_{-\infty}^{\infty} f(u,y)\,du = \int_0^y 2\,du = 2y \quad \text{für } 0 < y < 1\ .$$

Damit lassen sich Mittelwert und Varianz für X und Y berechnen:

$$\begin{aligned}
\mu_X &= \int_{-\infty}^{\infty} x\,f_X(x)\,dx = \int_0^1 x\,2(1-x)\,dx = 2\left(\frac{x^2}{2} - \frac{x^3}{3}\right)\Big|_0^1 = \frac{2}{6} = \frac{1}{3} \\
\sigma_X^2 &= \int_{-\infty}^{\infty} x^2\,f_X(x)\,dx - \mu_X^2 = \int_0^1 x^2\,2(1-x)\,dx - \left(\frac{1}{3}\right)^2 \\
&= 2\left(\frac{x^3}{3} - \frac{x^4}{4}\right)\Big|_0^1 - \frac{1}{9} = \frac{2}{12} - \frac{1}{9} = \frac{1}{18}
\end{aligned}$$

und analog

$$\begin{aligned}
\mu_Y &= \int_{-\infty}^{\infty} y\,f_Y(y)\,dy = \int_0^1 y\,2y\,dy = \frac{2y^3}{3}\Big|_0^1 = \frac{2}{3} \\
\sigma_Y^2 &= \int_{-\infty}^{\infty} y^2\,f_Y(y)\,dy - \mu_Y^2 = \int_0^1 y^2\,2y\,dy - \left(\frac{2}{3}\right)^2 \\
&= \frac{2y^4}{4}\Big|_0^1 - \frac{4}{9} = \frac{1}{2} - \frac{4}{9} = \frac{1}{18}\ .
\end{aligned}$$

Zur Berechnung der Kovarianz verwendet man den Verschiebungssatz:

$$\begin{aligned}
\sigma_{X,Y} &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} xy\ f(x,y)\,dx\,dy - \mu_X\mu_Y \\
&= \int_0^1 \left(\int_x^1 xy\ 2\ dy\right) dx - \frac{1}{3}\times\frac{2}{3} \\
&= \int_0^1 x\ \left(y^2\big|_x^1\right) dx - \frac{2}{9} \\
&= \int_0^1 x\,(1-x^2)\,dx - \frac{2}{9} \\
&= \left(\frac{x^2}{2} - \frac{x^4}{4}\right)\Big|_0^1 - \frac{2}{9} = \frac{1}{4} - \frac{2}{9} = \frac{1}{36}\ .
\end{aligned}$$

Damit erhält man als Korrelationskoeffizient

$$\rho_{X,Y} = \frac{\sigma_{X,Y}}{\sigma_X\sigma_Y} = \frac{1/36}{\sqrt{1/18}\sqrt{1/18}} = \frac{1/36}{1/18} = \frac{1}{2}\ .$$

Es besteht damit ein durchschnittlich ausgeprägter, positiv orientierter linearer Zusammenhang zwischen X und Y.

◇◇◇

Momente von Linearkombinationen:

Bildet man zu den *ZGen* $X_1, X_2, \ldots, X_k$ mit $\mu_i = \mathrm{E}(X_i)$, $\sigma_i^2 = \mathrm{Var}(X_i)$ und $\sigma_{ij} = \mathrm{Cov}(X_i, X_j)$ die Linearkombination

$$Y = a_1X_1 + a_2X_2 + \cdots + a_kX_k$$

mit den (reellen) Koeffizienten $a_1, a_2, \ldots, a_k$, so gilt

$$\mathrm{E}(Y) = a_1\mu_1 + a_2\mu_2 + \cdots + a_k\mu_k \tag{3.88}$$

und

$$\mathrm{Var}(Y) = \sum_{i=1}^{k}\sum_{j=1}^{k} a_ia_j\sigma_{ij}\ , \tag{3.89}$$

wobei $\sigma_{ii} = \sigma_i^2$ bezeichnet ($1 \le i, j \le k$). Falls die X_i unabhängig sind, gilt für die Varianz von Y die einfachere Beziehung

$$\mathrm{Var}(Y) = \sum_{i=1}^{k} a_i^2\sigma_i^2\ . \tag{3.90}$$

Beweis:
Die Formel (3.88) für den Erwartungswert ergibt sich unmittelbar aus der Linearität von Summe bzw. Integral (Erwartungswertbildung). Für die allgemeine Varianzbeziehung (3.89) ergibt sich unter

Ausnützung der Linearität der Erwartungswertbildung

$$\begin{aligned}
\mathrm{Var}(Y) &= \mathrm{E}((Y-\mu_Y)^2) \\
&= \mathrm{E}\left(\left(\sum_{i=1}^{k} a_i X_i - \sum_{i=1}^{k} a_i\mu_i\right)^2\right) = \mathrm{E}\left(\left(\sum_{i=1}^{k} a_i (X_i-\mu_i)\right)^2\right) \\
&= \mathrm{E}\left(\sum_{i=1}^{k}\sum_{j=1}^{k} a_i a_j (X_i-\mu_i)(X_j-\mu_j)\right) \\
&= \sum_{i=1}^{k}\sum_{j=1}^{k} a_i a_j \mathrm{E}((X_i-\mu_i)(X_j-\mu_j)) = \sum_{i=1}^{k}\sum_{j=1}^{k} a_i a_j \sigma_{ij} \ .
\end{aligned}$$

Im Falle der Unabhängigkeit der X_i verschwinden alle Kovarianzen und man erhält die Beziehung (3.90). △

Momente des Stichprobenmittels:

Bilden $X_1, X_2, \ldots, X_n$ eine (unabhängige) Stichprobe vom Umfang n für ein Merkmal X (d.h. die Verteilungen der X_i stimmen überein und gleichen der von X), so gilt für das Stichprobenmittel $\bar{X} = (X_1 + X_2 + \cdots + X_n)/n$

$$\mathrm{E}(\bar{X}) = \mu_X \tag{3.91}$$

und

$$\mathrm{Var}(\bar{X}) = \frac{\sigma_X^2}{n} \ , \tag{3.92}$$

wobei μ_X und σ_X^2 Mittelwert und Varianz von X darstellen.

Beweis:
Mit $a_1 = a_2 = \cdots = a_n = \frac{1}{n}$ ergibt sich aus (3.88) die Formel (3.91) für den Mittelwert und wegen der Unabhängigkeit der X_i folgt bei Anwendung von (3.90) die Beziehung (3.92) für die Varianz des Stichprobenmittels. △

3.8.5 Bedingte Verteilung, bedingte Erwartung

Bei der Betrachtung multivariater Phänomene $\mathbf{X} = (X_1, X_2, \ldots, X_k)$ tritt häufig der Fall ein, dass die Modellierung einer Komponente X_i nur bei bestimmten Konstellationen für andere Merkmale X_j $(j \neq i)$ von Interesse ist. Man betrachtet dann etwa die Verteilung von X_i unter der Voraussetzung, dass für X_j ein spezielles Ereignis, etwa der Form "$X_j = x_j$" oder "$a_j < X \leq b_j$", eintritt.

Beispiel 3.18 a) In Bsp. 1.2 kann man die Verteilung des Berufs des Vaters (Merkmal X_3) bei HTL–Maturanten mit der bei Absolventen anderer Schultypen vergleichen. Im ersten Fall lautet das bedingende Ereignis $(X_4 = 2)$ und im zweiten Fall $(X_4 \in \{1, 3, 4\})$.

b) In Bsp. 3.18 kann die Verteilung der Verunreinigung durch Stoff A (Merkmal X) bei einem Verunreinigungsgrad $Y < 0.5\,\%$ durch Stoff B verglichen werden mit der bei $Y > 0.5\,\%$. ◇◇◇

Diskrete *ZGen*:

Die durch $Y = y$ *bedingte* Wahrscheinlichkeitsfunktion $p_X(x_l|Y = y)$ der *ZG* X ergibt sich auf Grund der Definition der bedingten Wahrscheinlichkeit als

$$p_X(x_l|Y = y) = P(X = x_l|Y = y) = \frac{P((X = x_l) \wedge (Y = y))}{P(Y = y)} = \frac{p_{X,Y}(x_l, y)}{p_Y(y)} \tag{3.93}$$

für $x_l \in M_X$ und $y \in M_Y$, wobei $p_{X,Y}$ die gemeinsame *W-Fkt* von X und Y und p_Y die Randwahrscheinlichkeitsfunktion von Y darstellen.

Stetige *ZGen*:

Die durch $Y = y$ bedingte *DF* $f_X(x|Y = y)$ der *ZG* X ergibt sich formal analog zum obigen Fall als

$$f_X(x|Y = y) = \frac{f_{X,Y}(x, y)}{f_Y(y)} \tag{3.94}$$

mit der gemeinsamen *DF* $f_{X,Y}$ und der Randdichte $f_Y(y)$ von Y.

Definition: Der durch das Ereignis $Y = y$ *bedingte Erwartungswert* $\mathrm{E}(X|Y = y)$ der *ZG* X ist der Mittelwert der durch $Y = y$ bedingten Verteilung von X. Da Y eine *ZG* darstellt, wird durch die Zuordnung

$$y \to \mathrm{E}(X|Y = y)$$

eine neue *ZG* definiert, die man als die *durch die* ZG Y *bedingte Erwartung von* X, kurz $\mathrm{E}(X|Y)$, bezeichnet.

Bedingte und nichtbedingte Erwartung:

Es gilt für die *ZGen* X und Y

$$\mathrm{E}(\mathrm{E}(X|Y)) = \mathrm{E}(X) \tag{3.95}$$

also "im Durchschnitt" decken sich die durch $Y = y$ bedingten Erwartungswerte von X mit dem unabhängig von Y betrachteten Erwartungswert von X.

Beispiel 3.19 In Fortsetzung von Bsp. 3.18 ergibt sich die durch $Y = y$ bedingte *DF* $f_X(x|Y = y)$ von X zu

$$f_X(x|Y = y) = \frac{f_{X,Y}(x, y)}{f_Y(y)} = \frac{2}{2y} = \frac{1}{y} \quad \text{für } 0 < x < y \ .$$

Die durch $Y = y$ bedingte Verteilung von X ist somit eine Rechteckverteilung $S(0, y)$ im Intervall $(0, y)$.
Der durch $Y = y$ bedingte Erwartungswert ist demnach

$$\mathrm{E}(X|Y = y) = \frac{y}{2}$$

und die durch Y bedingte Erwartung von X ergibt sich als

$$\mathrm{E}(X|Y) = \frac{Y}{2} \ .$$

Offensichtlich gilt

$$\mathrm{E}(\mathrm{E}(X|Y)) = \mathrm{E}\left(\frac{Y}{2}\right) = \frac{1}{2}\,\mathrm{E}(Y) = \frac{1}{2}\,\frac{2}{3} = \frac{1}{3} = \mathrm{E}(X) \ .$$

◇◇◇

Beispiel 3.20 Sind die beiden *ZGen* X und Y 2–dimensional normalverteilt, so besitzen sie die (2–dimensionale) *DF*

$$f_{X,Y}(x,y) = \frac{1}{2\pi\,\sigma_X\sigma_Y\sqrt{1-\rho^2}}\,\exp\left(-\frac{1}{2(1-\rho^2)}\left[\left(\frac{x-\mu_X}{\sigma_X}\right)^2 - 2\rho\left(\frac{x-\mu_X}{\sigma_X}\right)\left(\frac{y-\mu_Y}{\sigma_Y}\right)+\left(\frac{y-\mu_Y}{\sigma_Y}\right)^2\right]\right) \tag{3.96}$$

◇◇◇

3.8.6 Fehlerfortpflanzungsgesetz

Eine nützliche *Näherungsformel* für Erwartung und Varianz einer Funktion $Y = h(X_1,\ldots,X_k)$ von *ZGen* $X_1,\ldots,X_k$, die in der Praxis gerne angewandt wird, ergibt sich im sogenannten

Fehlerfortpflanzungsgesetz:

Ist $h : \boldsymbol{R}^k \to \boldsymbol{R}$ eine differenzierbare Funktion und stellen $X_1,\ldots,X_k$ *ZGen* dar mit $\mu_i = \mathrm{E}(X_i)$, $\sigma_i^2 = \mathrm{Var}(X_i)$ und $\sigma_{ij} = \mathrm{Cov}(X_i,X_j)$, so gilt für die Erwartung von $Y = h(X_1,\ldots,X_k)$ die *Näherungsformel*

$$\mathrm{E}(Y) \approx h(\mu_1,\ldots,\mu_k) \tag{3.97}$$

und für die Varianz von Y erhält man die Approximation

$$\mathrm{Var}(Y) \approx \sum_{i=1}^{k}\sum_{j=1}^{k}\frac{\partial h}{\partial x_i}(\mu_1,\ldots,\mu_k)\,\frac{\partial h}{\partial x_j}(\mu_1,\ldots,\mu_k)\,\sigma_{ij}\ . \tag{3.98}$$

Beweis:
Zunächst lässt sich für Y die Taylorreihenentwicklung

$$Y = h(X_1,\ldots,X_k) = h(\mu_1,\ldots,\mu_k) + \sum_{i=1}^{k}\frac{\partial h}{\partial x_i}(\mu_1,\ldots,\mu_k)(X_i-\mu_i) + R(X_1,\ldots,X_k)$$

mit dem Restterm R angeben, sodass bei Nichtberücksichtigung dieses Terms

$$Y \approx Y_{\mathrm{lin}} = h(\mu_1,\ldots,\mu_k) + \sum_{i=1}^{k}\frac{\partial h}{\partial x_i}(\mu_1,\ldots,\mu_k)(X_i-\mu_i)$$

gilt. Daraus erhält man

$$\mathrm{E}(Y) \approx \mathrm{E}(Y_{\mathrm{lin}}) = h(\mu_1,\ldots,\mu_k) + \sum_{i=1}^{k}\frac{\partial h}{\partial x_i}(\mu_1,\ldots,\mu_k)\underbrace{\mathrm{E}(X_i-\mu_i)}_{0}$$

bzw.

$$\begin{aligned}\mathrm{Var}(Y) &\approx \mathrm{Var}(Y_{\mathrm{lin}}) = \mathrm{E}((Y_{\mathrm{lin}}-\mathrm{E}(Y_{\mathrm{lin}}))^2)\\ &= \mathrm{E}\left(\left(\sum_{i=1}^{k}\frac{\partial h}{\partial x_i}(\mu_1,\ldots,\mu_k)(X_i-\mu_i)\right)^2\right)\\ &= \sum_{i=1}^{k}\sum_{j=1}^{k}\frac{\partial h}{\partial x_i}(\mu_1,\ldots,\mu_k)\frac{\partial h}{\partial x_j}(\mu_1,\ldots,\mu_k)\underbrace{\mathrm{E}((X_i-\mu_i)(X_j-\mu_j))}_{\sigma_{ij}}\ .\end{aligned}$$

△

Beispiel 3.21 Wie lautet das Verhältnis $V = X/Y$ der Verunreinigungen durch Stoff A bzw. B im Durchschnitt und wie streut es?

Lösung: Für $h(x, y) = x/y$ gilt

$$\frac{\partial h}{\partial x}(x,y) = \frac{1}{y} \qquad \text{bzw.} \qquad \frac{\partial h}{\partial y}(x,y) = -\frac{x}{y^2}$$

und damit wegen $\mu_X = \frac{1}{3}$ bzw. $\mu_Y = \frac{2}{3}$

$$\frac{\partial h}{\partial x}(\mu_X,\mu_Y) = \frac{1}{2/3} = \frac{3}{2} \qquad \text{bzw.} \qquad \frac{\partial h}{\partial y}(\mu_X,\mu_Y) = \frac{-1/3}{(2/3)^2} = -\frac{3}{4}\ .$$

Daher gilt

$$V \approx \frac{1}{2} + \frac{3}{2}(X - \frac{1}{3}) - \frac{3}{4}(Y - \frac{2}{3})\ .$$

Man erhält nun

$$\mu_V = \mathrm{E}(V) \approx \frac{1}{2}$$

und

$$\begin{aligned} \sigma_V^2 &= \mathrm{Var}(V) \approx \left(\frac{3}{2}\right)^2 \sigma_X^2 - 2\,\frac{3}{2}\,\frac{3}{4}\sigma_{X,Y} + \left(\frac{3}{4}\right)^2 \sigma_Y^2 \\ &= \left(\frac{9}{4} - \frac{9}{4} + \frac{9}{16}\right)\,\frac{1}{18} = \frac{1}{32} \end{aligned}$$

bzw. $\sigma_V \approx \sqrt{2}/8 = 0.177$.

◇◇◇

3.9 Zentraler Grenzwertsatz

Zentrale Grenzwertsätze *(ZGS)* geben Auskunft über das Konvergenzverhalten der Verteilung von Summen $Z_n = X_1 + X_2 + \cdots + X_n$ in der Regel unabhängiger *ZGen* $X_1, X_2, \ldots$. Meist handelt es sich bei diesen Grenzverteilungen um eine Normalverteilung, wodurch die überragende Bedeutung dieser Verteilung unterstrichen wird.

Sind $X_1, \ldots X_n$ stochastisch unabhängige Zufallsvariablen mit Mittelwerten μ_i und Varianzen σ_i^2, dann konvergiert die Summe (unter gewissen technischen Bedingungen)

$$Z_i = \sum_{i=1}^{n} X_i$$

für $n \to \infty$ gegen eine normalverteilte Zufallsvariable mit Mittelwert $\mu_Z = \sum \mu_i$ und Varianz $\sigma_Z^2 = \sum \sigma_i^2$.

In sehr einfacher Form gibt ein *ZGS* die Grenzverteilung des Stichprobenmittels $\bar{X}$ an, dieser *ZGS* heißt auch *Gesetz der großen Zahlen*:

Zentraler Grenzwertsatz für $\bar{X}$:

Bei einer Stichprobe $X_1, X_2, \ldots, X_n$ unabhängiger Beobachtungen einer *ZG* X mit $\mathrm{E}(X) = \mu$ und $\mathrm{Var}(X) = \sigma^2$ gilt für das Stichprobenmittel

$$\bar{X}_n = \frac{1}{n} \sum_{i=1}^{n} X_i$$

zunächst

$$\mathrm{E}(\bar{X}_n) = \mu \qquad \text{und} \qquad \mathrm{Var}(\bar{X}_n) = \frac{\sigma^2}{n}$$

und für seine Verteilung

$$\lim_{n\to\infty} P\left(\frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} \le z\right) = \Phi(z) \tag{3.99}$$

bzw.

$$F_{\bar{X}_n}(x) = P\left(\frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} \le \frac{x-\mu}{\sigma/\sqrt{n}}\right) \approx \Phi\left(\frac{x-\mu}{\sigma/\sqrt{n}}\right) \ . \tag{3.100}$$

Bemerkung: Der Typ der Verteilung von X ist offensichtlich nicht ausschließlich auf stetige Verteilungen beschränkt. Für die Aussagen reichen die Existenz von $\mathrm{E}(X)$ und $\mathrm{Var}(X)$. Die Gesetze der großen Zahlen bzw. zentralen Grenzwertsätze sind ein sehr wichtiges Kapitel der Wahrscheinlichkeitstheorie und existieren in einer Vielzahl von Varianten, die in der Praxis jedoch meist nicht relevant sind. Die Kernaussagen sind immer ident, Summen konvergieren in der Regel mit wachsender Anzahl von Summanden gegen eine Normalverteilung.

Das folgende Beispiel zeigt für einen sehr extremen Fall eines nichtnormalverteilten Merkmales (Exponentialverteilung) die Wirksamkeit dieser einfachen Form eines Zentralen Grenzwertsatzes.

Beispiel 3.22 Die Lebensdauer eines bestimmten Leistungstransistors in der Schaltzentrale eines EVU sei exponentialverteilt mit der mittleren Lebensdauer 10.000 h, d.h.

$$\Pr(T > t) = e^{-t/10000} \qquad \text{für } t > 0 \quad .$$

Betrachtet man eine Reihe derartiger Transistoren mit voneinander unabhängigen Ausfallzeitpunkten $T_1, T_2, T_3, \ldots, T_n, \ldots$, so gilt auf Grund des Zentralen Grenzwertsatzes für das Stichprobenmittel (3.99) $\bar{T}_n = (T_1 + T_2 + \cdots + T_n)/n$

$$\lim_{n\to\infty} \Pr((\bar{T}_n - \mathrm{E}(\bar{T}_n))/\sqrt{\mathrm{Var}(\bar{T}_n)} \le u) = \Phi(u) \qquad \text{für } u \in \boldsymbol{R} \ ,$$

d.h.

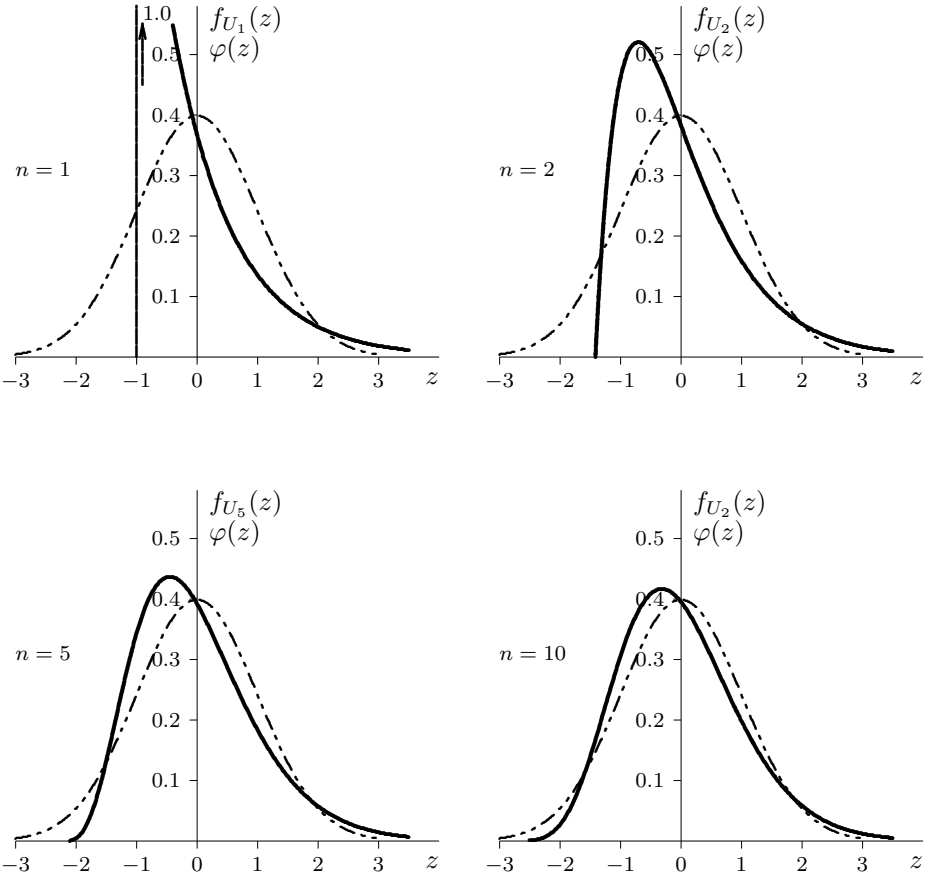
$$U_n = \frac{\bar{T}_n - \mathrm{E}(\bar{T}_n)}{\sqrt{\mathrm{Var}(\bar{T}_n)}}$$

ist asymptotisch normalverteilt und damit $\bar{T}_n$ selbst ungefähr $N(\mathrm{E}(\bar{T}_n), \mathrm{Var}(\bar{T}_n))$–verteilt. Anhand der Dichtefunktion des *standardisierten* Stichprobenmittels U_n für $n = 1$, 2, 5 und 10 und ihrer Annäherung an die Dichtefunktion der $N(0,1)$–Standardnormalverteilung wird die Wirksamkeit des *ZGS* sehr deutlich. Die Abb. 3.20 zeigt diesen Sachverhalt grafisch.

◇◇◇

Auf dem *ZGS* beruhen auch die für die Praxis wichtigen Approximationsformeln

Abbildung 3.20: *ZGS für* $\bar{T}_n$



$$
\begin{aligned}
Bi(n,p) &\approx N(np, np(1-p)) && \text{für } np \geq 5 \text{ und } n(1-p) \geq 0,\\
Po(\mu) &\approx N(\mu, \mu) && \text{für } \mu \geq 10,\\
\chi^2_n &\approx N(n, 2n) && \text{für } n \geq 30\ .
\end{aligned}
$$

Zur *Stetigkeitskorrektur* verwendet man aber für die Approximation diskreter Verteilung in der Regel eine Korrektur des Argumentwertes um eine halbe Einheit, also etwa im Fall einer binomialverteilten *ZG* X gilt dann

$$P(X \leq k) \approx \Phi\left(\frac{k + 0.5 - np}{\sqrt{np(1-p)}}\right) \tag{3.101}$$

und Analoges gilt für die Poisson–Verteilung.

Kapitel 4

Grundlagen der schließenden Statistik

Die Wahrscheinlichkeitsrechnung stellt eine Reihe von Modellen und Regeln zur Verfügung, mit deren Hilfe sich die Phänomene *Unsicherheit* oder *Zufälligkeit* sehr gut beschreiben lassen. In der Welt dieser Modelle lassen sich eine Reihe von Aussagen ableiten (z.B. über die Verteilung der Summe zweier unabhängiger rechteckverteilter *ZGen* oder über die asymptotische Verteilung des Stichprobenmittels), die zunächst einmal nur von theoretischem Wert erscheinen.

Zur Festlegung der betrachteten Modelle dienen häufig *Parameter* (z.B. Mittelwert oder Varianz einer Verteilung). Ohne Kenntnis dieser Parameter haben die Modelle aber wenig praktischen Wert. So trägt beispielsweise die Aussage "die Anzahl von Fichtensämlingen in einem Normquadrat mit $10\,m$ Seitenlänge in einem bestimmten Forst ist poissonverteilt" ohne Angabe eines Wertes (oder zumindest eines Wertebereiches) für den Parameter μ wenig zur praktischen Verwertbarkeit bei.

Die Anwendbarkeit dieser *stochastischen* Modelle steht und fällt somit sehr häufig mit der Beschreibung der Modellparameter. Dazu dient die *schließende Statistik*. Die Grundlage stellt eine Stichprobe von

- identisch verteilten und
- in der Regel unabhängigen

Beobachtungen dar. Formal geht man also von der Verteilung einer (ein– oder auch mehrdimensionalen) *ZG* X mit der *VF* F_X aus (z.B. Normalverteilung, Poissonverteilung) und beobachtet diese Größe n–mal, in der Regel *unabhängig* voneinander. Das Ergebnis $x_1, x_2, \ldots, x_n$ bildet eine *Stichprobe* vom *Umfang* n für die *ZG* X. Für die Verteilung jeder einzelnen Beobachtung gilt offensichtlich:

$$X_i \sim F_X \quad (i = 1, \ldots, n)$$

(die Großschreibung von X_i soll darauf hindeuten, dass jede Stichprobenbeobachtung natürlich auch als Wert einer *ZG* aufgefasst werden kann).

Zwei zentrale Aufgaben stellen sich nun bei der Analyse von Stichproben. Zum einen kann die Schätzung von Parametern im Vordergrund stehen, zum anderen die Beurteilung von Aussagen über diese Parameter oder die Verteilung von X. Im ersten Fall ist man z.B.

am *Wert* des Mittelwertparameters oder an einem *Bereich* für den Varianzparameter einer normalverteilten *ZG* interessiert. Im zweiten Fall geht es um die Überprüfung von Aussagen der Form "der Mittelwertparameter ist positiv" oder "die *ZG X* ist normalverteilt". Der erste Aufgabenbereich bildet das Thema *Schätzung von Parametern* und der zweite die Frage des *Testens von Hypothesen* (Aussagen).

4.1 Parameterschätzung

Geht man von einer Stichprobe $x_1, x_2, \ldots, x_n$ für die *ZG X* aus, so versucht man, für einen unbekannten Parameter θ der Verteilung von X in der Regel zunächst *einen* Wert $\hat{\theta}$ aus der Stichprobe abzuleiten, d.h. zu schätzen. Zur Schätzung des Mittelwertparameters μ kann man etwa den arithmetischen Mittelwert (das Stichprobenmittel) $\bar{x}$ heranziehen, also $\hat{\mu} = \bar{x}$. Das $\hat{}$–Zeichen über dem Parameter θ dient als Hinweis dafür, dass es sich bei dem abgeleiteten Wert um eine *Schätzung* für den Parameter handelt.

Man nennt diese Form der Schätzung auch *Punktschätzung* ("1–Wert–Schätzung") im Gegensatz zur später zu behandelnden *Bereichsschätzung*, bei der für den unbekannten Parameter ein ganzer Bereich (z.B. Intervall) angeboten wird.

Es ist aber einleuchtend, dass mit einer neuen Stichprobe vermutlich ein anderer Schätzwert für den Parameter abgeleitet würde. Die Schätzung $\hat{\theta} = \hat{\theta}(X_1, \ldots, X_n)$ hängt also von der Stichprobe ab und stellt damit selbst eine *Zufallsgröße* dar. Um das auch sprachlich herauszustreichen, spricht man von einem *Schätzer* oder *Schätzverfahren*, wenn man die Schätzung als *ZG* auffasst. Die Verteilung von $\hat{\theta}$ hängt nur vom Stichprobenumfang n und von der Verteilung von X ab. Je nach Art der Verteilung von $\hat{\theta}$ spricht man von *kontinuierlicher* oder *diskreter* Parameterschätzung.

Eigenschaften und Kriterien für Schätzungen $\hat{\theta}$ lassen sich in der Regel anhand der Verteilung von $\hat{\theta}$ formulieren und überprüfen.

4.1.1 Eigenschaften von Schätzern

Im Folgenden werden einige wichtige Eigenschaften zusammengefasst, die man von brauchbaren Schätzungen in der Regel erwarten wird. Der Folgeabschnitt behandelt dann eine Möglichkeit zur Konstruktion geeigneter Schätzer, die diesen Eigenschaften weitestgehend entsprechen.

Erwartungstreue (Verzerrungsfreiheit, *engl.* unbiasedness):

Es ist klar, dass ein Schätzwert $\hat{\theta}$ in der Regel kaum mit dem theoretischen Parameter θ übereinstimmen wird. Andererseits kann man verlangen, dass im Zuge vieler Stichprobenbeobachtungen die Schätzwerte *im Durchschnitt* mit dem zu schätzenden Parameter übereinstimmen. Formal verlangt man also, dass der *Erwartungswert* des Schätzers $\hat{\theta}$ mit dem zugrundeliegenden Parameter θ zusammenfällt:

$$\mathrm{E}(\hat{\theta}|\theta) = \theta \ . \tag{4.1}$$

Beispiel 4.1 Der Stichprobenmittelwert ist ein erwartungstreuer Schätzer für den Mittelwertparameter (z.B. Normalverteilung, Exponentialverteilung, Alternativverteilung, Poissonverteilung). Es gilt nämlich für $\mathrm{E}(X) = \mu$ auf Grund der Linearität der Erwartung (siehe

(3.88))

$$\mathrm{E}(\bar{X}|\mu) = \mathrm{E}\left(\frac{1}{n}\sum_{i=1}^{n} X_i\big|\mu\right) = \frac{1}{n}\sum_{i=1}^{n}\mathrm{E}(X_i|\mu) = \frac{1}{n}\,n\mu = \mu\ .$$

◇◇◇

Beispiel 4.2 Die Stichprobenvarianz ist erwartungstreu für den Varianzparameter. Mit $\mathrm{E}(X) = \mu$ und $\mathrm{Var}(X) = \sigma^2$ gilt nämlich unter Verwendung der Regel (2.4)

$$\begin{aligned}
\mathrm{E}(S^2|\sigma^2) &= \mathrm{E}\left(\frac{1}{n-1}\left(\sum_{i=1}^{n} X_i^2 - n\bar{X}^2\right)|\sigma^2\right) \\
&= \frac{1}{n-1}\sum_{i=1}^{n}\mathrm{E}(X_i^2|\sigma^2) - \frac{n}{n-1}\mathrm{E}(\bar{X}^2|\sigma^2) \\
&= \frac{n}{n-1}(\sigma^2+\mu^2) - \frac{n}{n-1}\left(\frac{\sigma^2}{n}+\mu^2\right) \\
&= \sigma^2\ ,
\end{aligned}$$

wobei (siehe (3.20))

$$\sigma^2 = \mathrm{Var}(X_i) = \mathrm{E}(X_i^2) - \mu^2$$

und daher

$$\mathrm{E}(X_i^2) = \sigma^2 + \mu^2\ ,$$

sowie (siehe (3.92))

$$\frac{\sigma^2}{n} = \mathrm{Var}(\bar{X}) = \mathrm{E}(\bar{X}^2) - \mu^2$$

und somit

$$\mathrm{E}(\bar{X}^2) = \frac{\sigma^2}{n} + \mu^2$$

verwendet wird. Damit erhält die anfangs eigentümlich anmutende Division der Summe der Abweichungsquadrate

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

durch $n-1$ anstelle von n ihre Rechtfertigung.

◇◇◇

Falls die Beziehung (4.1) für endliche Stichprobenumfänge nicht gilt, aber wenigstens mit wachsendem n immer besser erfüllt wird:

$$\lim_{n\to\infty}\ \mathrm{E}(\hat{\theta}_n|\theta) = \theta\ , \tag{4.2}$$

spricht man von *asymptotischer Erwartungstreue*. Der Index n bei der Schätzung $\hat{\theta}_n$ deutet dabei auf den der jeweiligen Schätzung zugrundeliegenden Stichprobenumfang hin.

Effizienz:

Ein erwartungstreues Schätzverfahren wird umso geeigneter erscheinen, je genauer der zu

schätzende Parameter getroffen wird. Ausgedrückt über die Verteilung des Schätzers, bedeutet dies, dass die Varianz möglichst klein ausfallen soll. Ein Schätzer heißt nun *effizient*, wenn

$$\mathrm{Var}(\hat{\theta}|\theta) \ \ldots \ \min! \tag{4.3}$$

für alle θ gilt.

Es ist Sache des Statistikers zu beurteilen, wann ein Schätzer effizient ist. In den (vielen) Fällen sogenannter regulärer Schätzer vereinfacht sich diese Überprüfung nachhaltig, weil die Varianz eines in Frage kommenden Schätzers eine angebbare Schranke (die sogenannte *Frechet–Rao–Cramer–* oder kurz *FRC–Schranke*) *nicht unterschreiten* kann. Ist nun für einen Schätzer die Varianz gleich dieser Schranke, so muss er effizient sein.

Eine schwächere Bedingung ist die *asymptotische Effizienz*, bei der sich die Varianz eines asymptotisch erwartungstreuen Schätzers mit wachsendem Stichprobenumfang der *FRC–*Schranke immer besser annähert.

Beispiel 4.3 Das Stichprobenmittel $\bar{X}$ ist im Falle einer normalverteilten *ZG* X effizient für den Parameter μ, weil dessen Varianz $\sigma^2_{\bar{X}} = \sigma^2/n$ gleich der hier geltenden *FRC–*Schranke ist. Ein aus anderen Gründen (Robustheit, einfache Berechnung) ebenfalls verbreiteter Schätzer für μ ist der Stichprobenmedian. Er ist zwar erwartungstreu, aber mit einer um ca. 25 % größeren Standardabweichung als $\bar{X}$ sicher nicht effizient. Andererseits ist der Effizienzverlust von 25 % möglicherweise nicht stark genug, um die anderen Vorteile von $\tilde{X}$ aufzuheben.

◇◇◇

Konsistenz:

Eine relativ schwache Eigenschaft, die das Verhalten eines Schätzers bei größerem Stichprobenumfang beschreibt, ist die Konsistenz. Strebt ein Schätzer mit wachsendem n gegen den zu schätzenden Parameter, dann heißt er *konsistent*. So strebt beispielsweise das Stichprobenmittel für $n \to \infty$ gegen den theoretischen Mittelwert und ist somit für diesen konsistent.

Unterschiedliche Konvergenzbegriffe für diese Grenzverhalten führen auch zu unterschiedlichen Konsistenzbegriffen.

4.1.2 Maximum–Likelihood–Schätzung

Die Idee dieser auch als Plausibilitätsmethode bekannten Art, Schätzer zu konstruieren, ist bestrickend, ihre Umsetzung zumeist aber mühsam. Man schätzt einen Parameter θ durch denjenigen Wert $\hat{\theta}_{ML}$, für den die beobachtete Stichprobe am plausibelsten oder "am wahrscheinlichsten" erscheint.

Beispiel 4.4 Das Gewicht X eines Golddelicious–Apfels sei als normalverteilt vorausgesetzt. Ein willkürlich herausgegriffener Apfel wiegt 175 g. Unter allen Werten für den Mittelwertparameter μ (bei unbekannter, aber gleicher Varianz σ^2) ist die Wahrscheinlichkeit, den Wert 175 oder einen Wert knapp davon zu beobachten, für $\mu = 175$ am größten. Also wäre in diesem Fall $\hat{\mu}_{ML} = 175$ der Maximum–Likelihood–(ML)–Schätzwert für μ .

◇◇◇

Grundlage für die Herleitung des *Maximum–Likelihood*–Schätzwertes $\hat{\theta}_{ML}$ (kurz: *ML*–Schätzwert) eines Parameters θ ist die Wahrscheinlichkeit, mit der – in Abhängigkeit vom betrachteten Parameter – die vorliegende Stichprobe beobachtet werden kann. Im Falle eines *stetigen* Merkmals X mit der (von θ abhängigen) *DF* $f_X(x|\theta)$ ergibt sich die Wahrscheinlichkeit, die vorliegende Stichprobe oder eine knapp herumliegende zu beobachten als

$$\begin{aligned} & P(x_1 \le X_1 \le x_1 + \Delta x, x_2 \le X_2 \le x_2 + \Delta x, \ldots, x_n \le X_n \le x_n + \Delta x|\theta) \\ \approx & \prod_{i=1}^{n} (f_X(x_i|\theta) \cdot \Delta x) = (\Delta x)^n \prod_{i=1}^{n} f_X(x_i|\theta) \ . \end{aligned} \tag{4.4}$$

Es ist zu beachten, dass das Volumen $(\Delta x)^n$ des beliebig kleinen n–dimensionalen Würfels mit der Kantenlänge Δx nur als Proportionalitätsfaktor dient und für die Maximierung der Wahrscheinlichkeit bloß das Produkt der Werte der *DF* an den Stichprobenwerten relevant ist.

Ähnlich beschreibt man im Falle einer *diskreten ZG* X mit der (von θ abhängigen) *W-Fkt* $p_X(x|\theta)$ die Wahrscheinlichkeit dafür, die vorliegende Stichprobe auch tatsächlich beobachten zu können, durch

$$P(X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n|\theta) = \prod_{i=1}^{n} p_X(x_i|\theta) \ . \tag{4.5}$$

Definiert man die sogenannnte *Likelihood–Funktion* (kurz: *Lik–Fkt*) für den Parameter θ durch

$$L(x_1, x_2, \ldots, x_n; \theta) \ := \ \begin{cases} \prod_{i=1}^{n} f_X(x_i|\theta) & \text{falls } X \text{ stetig} \\ \prod_{i=1}^{n} p_X(x_i|\theta) & \text{falls } X \text{ diskret,} \end{cases} \tag{4.6}$$

so läuft die Suche nach dem *ML*–Schätzwert $\hat{\theta}_{ML}$ für θ offensichtlich auf die Maximierung der *Lik–Fkt* hinaus. Als *ML–Schätzwert* erhält man sodann

$$\hat{\theta}_{ML} \ := \ L(x_1, x_2, \ldots, x_n; \hat{\theta}_{ML}) = \max_{\theta} L(x_1, x_2, \ldots, x_n; \theta) \ . \tag{4.7}$$

Da es nur um das Aufsuchen der Maximalstelle von $L(x_1, x_2, \ldots, x_n|\theta)$ geht und nicht um das Maximum selbst, kann man vorher eine *monotone Transformation* (z.B. Logarithmus, Potenz) anwenden, um den Rechenaufwand zu reduzieren. Eine effiziente Methode ist üblicherweise der Übergang zum (natürlichen) Logarithmus

$$l(x_1, x_2, \ldots, x_n; \theta) = \ln \, L(x_1, x_2, \ldots, x_n; \theta) \ ; \tag{4.8}$$

man spricht dann auch von der *Log–Likelihood–Funktion* (kurz: *Log–Lik–Fkt*).

Beispiel 4.5 Wie lautet der *ML*–Schätzwert für den Parameter p (Anteil) einer alternativverteilten *ZG* X ?

Lösung:

Die *W–Fkt* lautet hier (siehe (3.30))

$$p_X(x|p) = \begin{cases} p & \text{falls } x = 1 \\ 1 - p & \text{falls } x = 0 \end{cases}$$

oder in einer geschlossenen Form

$$p_X(x|p) = p^x(1-p)^{1-x}$$

für $x \in \{0,1\}$. Damit ergibt sich die *Lik–Fkt* als

$$L(x_1, x_2, \ldots, x_n; p) = \prod_{i=1}^{n} p_X(x_i|p) = p^{\sum_{i=1}^n x_i}(1-p)^{n-\sum_{i=1}^n x_i}$$

bzw. die *Log–Lik–Fkt* als

$$\begin{aligned} l(x_1, x_2, \ldots, x_n; p) &= \ln L(x_1, x_2, \ldots, x_n; p) \\ &= \left(\sum_{i=1}^{n} x_i\right) \ln p + \left(n - \sum_{i=1}^{n} x_i\right) \ln(1-p) \ . \end{aligned}$$

Zur Maximierung von $l(x_1, x_2, \ldots, x_n|p)$ bildet man zunächst die Ableitung nach dem Parameter p

$$\frac{\partial}{\partial p} l(x_1, x_2, \ldots, x_n; p) = \left(\sum_{i=1}^{n} x_i\right) \cdot \frac{1}{p} + \left(n - \sum_{i=1}^{n} x_i\right) \cdot \frac{(-1)}{1-p}$$

und setzt diese dann null:

$$\begin{aligned} \left(\sum_{i=1}^{n} x_i\right) \cdot \frac{1}{p} + \left(n - \sum_{i=1}^{n} x_i\right) \cdot \frac{(-1)}{1-p} &= 0 \qquad | \times p(1-p) \\ \sum_{i=1}^{n} x_i - p\sum_{i=1}^{n} x_i - np + p\sum_{i=1}^{n} x_i &= 0 \end{aligned}$$

Das ergibt mit

$$\hat{p}_{ML} = \frac{\sum_{i=1}^n x_i}{n} = \bar{x}$$

das Stichprobenmittel (d.h. hier die realtive Häufigkeit) als *ML*–Schätzwert für p .

◇◇◇

Beispiel 4.6 Wie lauten die *ML*–Schätzer für die Parameter μ und σ^2 einer normalverteilten *ZG*?

Lösung:

Die *Lik–Fkt* für die Parameter μ und $\nu = \sigma^2$ (diese Umbenennung erfolgt nur zur Vermeidung von Schwierigkeiten bei der folgenden Ableitung) ergibt sich als

$$\begin{aligned} L(x_1, x_2, \ldots, x_n; \mu, \nu) &= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\nu}} \, e^{-\frac{(x_i-\mu)^2}{2\nu}} \\ &= \frac{1}{(2\pi)^{n/2}\nu^{n/2}} \, e^{-\frac{1}{2\nu}\sum_{i=1}^n (x_i-\mu)^2} \ . \end{aligned}$$

Damit erhält man als *Log–Lik–Fkt*

$$l(x_1, x_2, \ldots, x_n; \mu, \nu) = -\frac{n}{2} \ln(2\pi) - \frac{n}{2} \ln \nu - \frac{1}{2\nu} \sum_{i=1}^{n} (x_i - \mu)^2$$

und deren partiellen Ableitungen

$$\begin{aligned}\frac{\partial}{\partial\mu}l(x_1,\ldots,x_n;\mu,\nu) &= -\frac{1}{2\nu}\sum_{i=1}^{n}2(x_i-\mu)(-1)\\ \frac{\partial}{\partial\nu}l(x_1,\ldots,x_n;\mu,\nu) &= -\frac{n}{2\nu}-\frac{(-1)}{2\nu^2}\sum_{i=1}^{n}(x_i-\mu)^2\ .\end{aligned}$$

Durch Nullsetzen und kleine Vereinfachungen ergibt sich das Gleichungssystem

$$\begin{aligned}\sum_{i=1}^{n}(x_i-\mu) &= 0\\ n\nu-\sum_{i=1}^{n}(x_i-\mu)^2 &= 0\ .\end{aligned}$$

Aus der ersten Gleichung ergibt sich als *ML*–Schätzwert für μ wieder das Stichprobenmittel: $\hat{\mu}_{ML}=\bar{x}$. Aus der zweiten Beziehung erhält man als *ML*–Schätzer für $\sigma^2=\nu$ schließlich

$$\hat{\sigma}^2_{ML}=\hat{\nu}_{ML}=\frac{1}{n}\sum_{i=1}^{n}(x_i-\hat{\mu}_{ML})^2=\frac{n-1}{n}\,s^2\ .$$

◇◇◇

Aus der letzten Formel erkennt man unmittelbar, dass *ML*–Schätzer nicht erwartungstreu sein müssen. Nichtsdestoweniger ist aber $\hat{\sigma}^2_{ML}$ asymptotisch erwartungstreu.

Eigenschaften von *ML*–Schätzern:

Unter relativ milden Voraussetzungen über die Verteilung der zugrundeliegenden *ZG* X gilt für die *ML*–Schätzung $\hat{\theta}_{ML}$ des Parameters θ

- $\hat{\theta}_{ML}$ ist zumindest asymptotisch erwartungstreu,
- $\hat{\theta}_{ML}$ ist zumindest asymptotisch effizient,
- $\hat{\theta}_{ML}$ ist konsistent

und vor allem

- nähert sich die Verteilung von $\hat{\theta}_{ML}$ mit wachsendem Stichprobenumfang einer Normalverteilung mit Mittelwertparameter θ und bekannter, von der Verteilung von X abhängiger Varianz–/Kovarianzstruktur (vgl. *Fisher–Information*).

Bemerkung 1: Speziell mit Hilfe der letzten Aussage über die asymptotische Normalverteilung von $\hat{\theta}_{ML}$ ist man in der Lage, *quantitative* Aussagen über die Schätzung zu machen.

Bemerkung 2: Ein großer Nachteil der *ML*–Methode besteht darin, dass das entstehende *ML*–Gleichungssystem der nullgesetzten partiellen Ableitungen der *(Log)–Lik–Fkt* selten geschlossen lösbar ist. Meist kann es nur numerisch mit Hilfe geeigneter Iterationsverfahren näherungsweise gelöst werden. Nichtsdestoweniger bleiben die oben angeführten Eigenschaften der Schätzung erhalten.

4.1.3 Konfidenzintervalle

Da die Chance, mit einer Schätzung den unbekannten Parameter tatsächlich zu "erraten", entweder überhaupt null (bei stetigen Schätzern) oder nur sehr klein ist (etwa bei diskreten Schätzungen), hat ein solcher Schätzwert alleine nur geringe Bedeutung. Durch Angabe eines *Intervalles*, in dem der Parameter mit relativ hoher Wahrscheinlichkeit liegen muss, wird die Schätzung mit einer Genauigkeitsschranke versehen.

Definition: Unter einem *Konfidenzintervall* (kurz: *K–Int*) für den Parameter θ mit der *Überdeckungswahrscheinlichkeit* (oder auch: *Sicherheit*) $1-\alpha$ (für α wählt man oft 1 %, 5 % oder 10 %) versteht man ein Intervall

$$(\theta_u(x_1,\ldots,x_n),\ \theta_o(x_1,\ldots,x_n))$$

mit *zufälligen* Grenzen, die von der beobachteten Stichprobe abhängen, sodass

$$P(\theta \in (\theta_u(X_1,\ldots,X_n),\ \theta_o(X_1,\ldots,X_n))) = 1-\alpha \tag{4.9}$$

gilt.

Bemerkung 1: Eine der beiden Grenzen kann auch $\pm\infty$ sein. Man spricht dann von *einseitigen* Konfidenzintervallen; im anderen Fall liegen *zweiseitige* Konfidenzintervalle vor.

Bemerkung 2: Je enger derartige *K–Int* ausfallen, umso präziser ist die Schätzung. Dies gelingt entweder bei einem wenig streuenden Merkmal X und/oder bei einem großen Stichprobenumfang.

Bemerkung 3: Die anschauliche Bedeutung eines *K–Int* besteht darin, dass die Situation, in der durch das angegebenene Intervall der unbekannte zugrundeliegende Parameter *nicht* überdeckt wird, nur mit der kleinen Irrtumswahrscheinlichkeit α eintreten kann.

Beispiel 4.7 Sei X ein $N(\mu,\sigma_0^2)$–normalverteiltes Merkmal mit *bekannter* Varianz σ_0^2. Für ein *K–Int* zu μ geht man von

$$P\left(-z_{1-\alpha/2} \le \frac{\bar{X}-\mu}{\sigma_0/\sqrt{n}} \le z_{1-\alpha/2}\right) = 1-\alpha$$

aus. Durch einfache Umformungen, die die Gültigkeit der Doppelungleichung nicht stören, erhält man

$$P\left(\bar{X} - z_{1-\alpha/2}\frac{\sigma_0}{\sqrt{n}} \le \mu \le \bar{X} + z_{1-\alpha/2}\frac{\sigma_0}{\sqrt{n}}\right) = 1-\alpha\ ,$$

sodass die Grenzen eines zweiseitigen *K–Int* für μ mit

$$\left.\begin{matrix}\mu_o\\ \mu_u\end{matrix}\right\} = \bar{X} \pm z_{1-\alpha/2}\frac{\sigma_0}{\sqrt{n}}$$

gegeben ist. Ein einseitig (nach unten begrenztes) *K–Int* erhält man beispielsweise aus

$$P\left(\frac{\bar{X}-\mu}{\sigma_0/\sqrt{n}} \le z_{1-\alpha}\right) = 1-\alpha$$

durch

$$P\left(\bar{X} - z_{1-\alpha}\frac{\sigma_0}{\sqrt{n}} \leq \mu\right) = 1 - \alpha$$

als

$$\begin{array}{rcl} \mu_o & = & \infty \\ \mu_u & = & \bar{X} - z_{1-\alpha}\frac{\sigma_0}{\sqrt{n}} \end{array} .$$

Analog dazu ergibt sich ein einseitiges, nach oben begrenztes *K–Int* als

$$\begin{array}{rcl} \mu_o & = & \bar{X} + z_{1-\alpha}\frac{\sigma_0}{\sqrt{n}} \\ \mu_u & = & -\infty \end{array} .$$

◇◇◇

Notwendiger Stichprobenumfang

In der Situation des obigen Beispiels lässt sich der für eine geforderte Genauigkeit *notwendige Stichprobenumfang* angeben. Verlangt man bei einer Überdeckungswahrscheinlichkeit $1 - \alpha$ (d.h. Irrtumswahrscheinlichkeit α) eine maximale Länge d für ein zweiseitiges Konfidenzintervall zu μ, so muss offensichtlich gelten

$$d \geq \mu_o - \mu_u = \left(\bar{X} + z_{1-\alpha/2}\frac{\sigma_0}{\sqrt{n}}\right) - \left(\bar{X} - z_{1-\alpha/2}\frac{\sigma_0}{\sqrt{n}}\right) = 2z_{1-\alpha/2}\frac{\sigma_0}{\sqrt{n}}$$

bzw.

$$n \geq \left(\frac{2z_{1-\alpha/2}\sigma_0}{d}\right)^2 .$$

Beispiel 4.8 Das Gewicht eines Golddelicious–Apfels ist normalverteilt mit der Standardabweichung $\sigma_0 = 15\,g$. Wie lautet ein 95 %–*K–Int* zu μ mit der Maximalbreite $d = 5\,g$?

Lösung:

$$n \geq \left(\frac{2 \times 1.96 \times 15}{5}\right)^2 = 11.76^2 \approx 138.39 ,$$

also wählt man $n = 139$.

◇◇◇

4.2 Testen von Hypothesen

Ziel ist die Überprüfung von Aussagen (Hypothesen) über Parameter oder allgemein über die Verteilung einer zugrundeliegenden *ZG* X. Beispiele für derartige Hypothesen – auch *Nullhypothese* H_0 genannt – sind etwa:

a) $H_0 : \mu = \mu_0$ $\qquad (X \sim N(\mu, \sigma_0^2))$

b) $H_0 : \mu \leq \mu_0$

c) $H_0 : \sigma^2 \geq \sigma_0^2$

d) $H_0 : X \sim S(0,1)$

e) $H_0 : X \sim Ex(\tau) \qquad (\tau > 0)$

f) $H_0 : X$ und Y sind unabhängig; $\qquad (X \ldots \text{Größe}, Y \ldots \text{Gewicht})$.

Im Falle a) bezieht sich die Aussage auf einen einzigen Parameterwert μ_0; man spricht daher von einer *einfachen* Nullhypothese. Bei b) bezieht man sich auf denselben Parameter (Mittelwert), bloß behauptet man nun, dass dieser nicht größer als μ_0 ausfällt. Unendlich viele Parameterwerte (kleiner oder gleich μ_0) sind hier zulässig und man nennt solche Hypothesen *zusammengesetzt*. Dasselbe gilt für den Fall c). In allen drei bisher genannten Fällen bezieht sich die Aussage auf Parameterwerte, und man spricht daher von *Parameterhypothesen*.

Bei d) wird behauptet, dass X im Intervall $(0,1)$ stetig gleichverteilt ist. Es wird somit eine einzige Verteilung angesprochen. Daher handelt es sich wieder um eine einfache Nullhypothese. Im Fall e) sagt die Nullhypothese aus, dass X exponentialverteilt ist, ohne auf den Parameter τ einschränkend Bezug zu nehmen. Es liegt daher eine zusammengesetzte Hypothese vor. In beiden Fällen werden Verteilungen angesprochen, und so nennt man die Nullhypothesen auch *Verteilungshypothesen*.

Schließlich wird im Fall f) eine Aussage über das Verhalten zweier zugrundeliegender Merkmale X und Y getroffen (es liegt somit ein zweidimensionales Merkmal (X,Y) vor).

Das Gegenteil der Nullhypothese wird als *Alternative* H_1 (oder *Alternativhypothese, Gegenhypothese*) bezeichnet. Eine der beiden Möglichkeiten trifft daher sicher zu.

Auf Grund einer Stichprobe soll nun eine *Entscheidung* über diese Hypothesen erfolgen:

"H_0 gilt"		*"H_0 gilt nicht"*
"H_0 wird akzeptiert"	oder	*"H_0 wird verworfen"*
"H_0 wird angenommen"		*"H_0 wird abgelehnt"* .

Die Anführungszeichen deuten darauf hin, dass es sich um *Urteile* handelt, die *fehlerhaft* sein können, weil man auf Grund der Endlichkeit der Stichprobe unmöglich ein vollständiges Bild über die zugrundeliegende Situation gewinnen kann. Damit gibt es vier denkbare Entscheidungssituationen:

Tatsache	Entscheidung	
	"H_0 annehmen"	"H_0 verwerfen"
H_0	o.k.	Fehler 1. Art
H_1	Fehler 2. Art	o.k.

Zwei korrekten stehen zwei Fehlentscheidungssituationen gegenüber:

1. Ein *irrtümliches* Verwerfen einer *tatsächlich* zutreffenden Nullhypothese, wo das Stichprobenergebnis so unglücklich liegt, dass die Entscheidung gegen die Nullhypothese getroffen werden "muss".

   Die Nullhypothese behauptet etwa, dass die durchschnittliche Körpergröße μ von Männern größer als 160 *cm* ist. Diese Aussage scheint auf Grund der allgemeinen Erfahrung auch tatsächlich zuzutreffen. Wie soll man sich entscheiden, wenn man in einer

Stichprobe von fünf Männern die Körpergrößen 153 *cm*, 159 *cm*, 145 *cm*, 170 *cm* und 155 *cm* vorfindet, weil man unglücklicherweise auf die Liliputanergruppe eines in der Stadt weilenden Zirkus gestoßen ist ? Die *Stichprobe* spricht *gegen* die Nullhypothese, daher ist sie – offensichtlich fälschlicherweise – zu verwerfen.

2. Die *irrtümliche* Annahme der Nullhypothese, obwohl *tatsächlich* die Gegenhypothese zutrifft. Bloß liegt hier eben die Stichprobe (für die Nullhypothese) so glücklich, dass die Nullhypothese angenommen wird.

   Lautet etwa im obigen Beispiel die Nullhypothese $H_0 : \ \mu \geq 190$ *cm*, so sagt die Erfahrung, dass diese Hypothese tatsächlich *nicht* gilt. Nun kann aber eine Stichprobe durchaus die Körpergrößen 195 *cm*, 210 *cm*, 198 *cm*, 190 *cm* und 189 *cm* ergeben (zufällig ein halbes Basketball–Team). Es deutet dann die Stichprobe *irrtümlich* auf die Gültigkeit von H_0 hin.

Definition: Unter einem *statistischen Test* τ versteht man eine Entscheidungsregel, die für jedes Stichprobenergebnis $x_1, \ldots, x_n$ eine Entscheidung für (0) oder gegen (1) die Nullhypothese H_0 angibt:

$$\tau \ : \ (x_1, \ldots, x_n) \ \rightarrow \ \tau(x_1, \ldots, x_n) \in \{0, 1\} \ .$$

Die Stichprobenergebnisse, die zur Ablehnung von H_0 führen, bilden den *kritischen Bereich* C des Tests

$$C = \{(x_1, \ldots, x_n) : \ \tau(x_1, \ldots, x_n) = 1\} \ .$$

Dadurch wird der Test charakterisiert.

Ein Test besitzt das *Signifikanzniveau* α $(0 < \alpha < 1)$, falls die Wahrscheinlichkeit eines Fehlers 1. Art höchstens gleich α ausfällt:

$$\sup_{\theta \in H_0} P(\tau(X_1, \ldots, X_n) = 1 | \theta) = \alpha$$

oder

$$\sup_{F_X \in H_0} P(\tau(X_1, \ldots, X_n) = 1 | F_X) = \alpha$$

je nachdem ob es sich bei H_0 um eine Parameter– oder eine Verteilungshypothese handelt. Üblicherweise wird für die Konstruktion eines Tests ein kleiner Wert für α vorgeschrieben, wobei gerne $\alpha = 0.01$ oder $\alpha = 0.05$ verwendet wird.

Die Beachtung eines kleinen Signifikanzniveaus sagt nichts aus über die Wahrscheinlichkeit einer Fehlentscheidungssituation 2. Art. Diese wird durch die sogenannte *Annahmefunktion* (auch: *Operationscharakteristik* oder *OC–Kurve*)

$$\beta(\theta) = P(\tau(X_1, \ldots, X_n) = 0 | \theta)$$

bzw.

$$\beta(F_X) = P(\tau(X_1, \ldots, X_n) = 0 | F_X)$$

beschrieben, soweit sie den Bereich $\theta \in H_1$ oder $F_X \in H_1$ betrifft, und manchmal als β–Fehler bezeichnet. Meistens wird zur Charakterisierung der Güte eines Tests das Gegenteil

$$g(\theta) = 1 - \beta(\theta) = P(\tau(X_1, \ldots, X_n) = 1 | \theta) \tag{4.10}$$

bzw.

$$g(F_X) = 1 - \beta(F_X) = P(\tau(X_1, \ldots, X_n) = 1 | F_X) \tag{4.11}$$

betrachtet und als *Gütefunktion* (*Schärfe, Macht*, engl. *power function*) dieses Tests bezeichnet.

Bemerkung: So unverständlich es im ersten Moment auch aussehen möge, so wird in der Regel eine Nullhypothese eigentlich zu dem Zweck formuliert, um sie *ablehnen* zu können. Durch das Konstruktionsprinzip des Tests hat man dann die Gewähr einer geringen Irrtumswahrscheinlichkeit, nämlich α.

Beispiel 4.9 Das Gewicht X eines Golddelicious–Apfels sei normalverteilt mit bekannter Varianz σ_0^2. Es wird behauptet, dass der Mittelwert gleich $\mu_0 = 150\,g$ beträgt. Wie lautet ein Test für diese Nullhypothese mit dem Signifikanzniveau $\alpha = 5\,\%$?

Lösung:

Die Entscheidung und somit der Test wird wohl auf dem Stichprobenmittelwert aufbauen. Als *Teststatistik*, anhand derer die Entscheidung formuliert wird, kann man

$$u = \frac{\bar{x} - \mu_0}{\sigma_0/\sqrt{n}}$$

heranziehen, weil

1) der Zusammenhang mit $\bar{x}$ unmittelbar gegeben ist und
2) die Verteilung von U (als ZG aufgefasst) bekannt ist. Es gilt nämlich

$$\bar{X} \sim N(\mu, \sigma_0^2/n)$$

und damit

$$\frac{\bar{X} - \mu_0}{\sigma_0/\sqrt{n}} \sim N(0,1)\ ,$$

falls $\mu \in H_0$. also $\mu = \mu_0$.

Nun wird man wohl umso eher für H_0 entscheiden, je näher $\bar{x}$ dem Wert $\mu_0 = 150\,g$ liegt, bzw. je kleiner, absolut gesehen, die Teststatistik ausfällt. Wählt man als Entscheidungsgrenzen

$$\left.\begin{matrix} c_o \\ c_u \end{matrix}\right\} = \pm z_{1-\alpha/2}$$

mit der Entscheidungsregel

$$u \left\{\begin{matrix} \in \\ \notin \end{matrix}\right\} (c_u, c_o) \Rightarrow \left\{\begin{matrix} \text{Annahme} \\ \text{Ablehnung} \end{matrix}\right\} \text{ von } H_0, \tag{4.12}$$

so erhält man wegen

$$\begin{aligned} P(\tau(X_1, \ldots, X_n) = 1 | \mu_0) &= P(U \notin (c_u, c_o) | \mu_0) \\ &= 1 - P\left(-z_{1-\alpha/2} \le \frac{\bar{X} - \mu_0}{\sigma_0/\sqrt{n}} \le z_{1-\alpha/2}\right) \\ &= 1 - (1-\alpha) \ = \ \alpha \end{aligned}$$

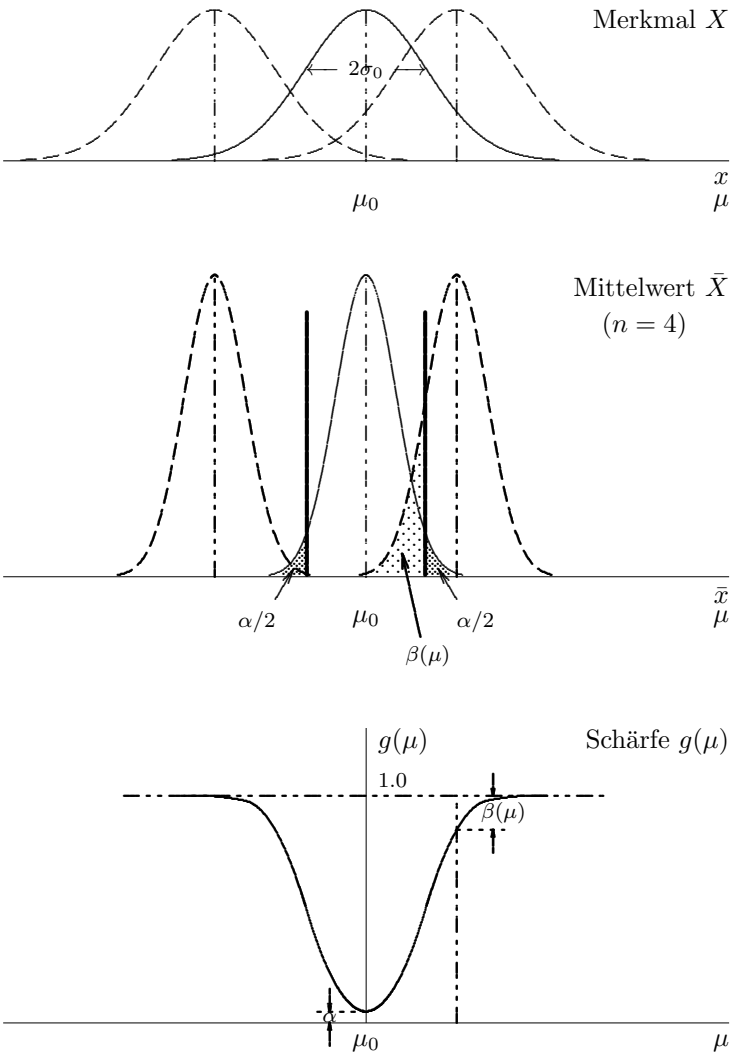
Der kritische Bereich wird somit durch $\bar{x}$–Werte unter $\mu_0 - z_{1-\alpha/2}\,\sigma_0/\sqrt{n}$ bzw. über $\mu_0 + z_{1-\alpha/2}\,\sigma_0/\sqrt{n}$ gebildet.

Wie man aus der Abb. 4.1 erkennt, sinkt die Wahrscheinlichkeit eines Fehlers 2. Art, je weiter μ von μ_0 wegliegt, und wächst beliebig gegen $1-\alpha$ (!), je mehr es sich μ_0 nähert:

$$\begin{aligned}
\beta(\mu) &= P(\tau(X_1,\ldots,X_n)=0|\mu) \\
&= P(U\in(c_u,c_o)|\mu) \\
&= P\left(-z_{1-\alpha/2} \le \frac{\bar{X}-\mu_0}{\sigma_0/\sqrt{n}} \le z_{1-\alpha/2}|\mu\right) \\
&= P\left(-z_{1-\alpha/2} \le \frac{(\bar{X}-\mu)+(\mu-\mu_0)}{\sigma_0/\sqrt{n}} \le z_{1-\alpha/2}\right) \\
&= P\left(-z_{1-\alpha/2} - \frac{\mu-\mu_0}{\sigma_0/\sqrt{n}} \le \frac{\bar{X}-\mu}{\sigma_0/\sqrt{n}} \le z_{1-\alpha/2} - \frac{\mu-\mu_0}{\sigma_0/\sqrt{n}}\right) \\
&= \Phi\left(z_{1-\alpha/2} - \frac{\mu-\mu_0}{\sigma_0/\sqrt{n}}\right) - \Phi\left(-z_{1-\alpha/2} - \frac{\mu-\mu_0}{\sigma_0/\sqrt{n}}\right)\ .
\end{aligned}$$

◇◇◇

Abbildung 4.1: *Test für den Parameter μ einer $N(\mu, \sigma_0^2)$–Verteilung*



Kapitel 5

Normalverteilungsverfahren

In diesem Kapitel werden Schätz– und Testverfahren behandelt, die sich auf die Parameter der *Normalverteilung* beziehen. Zusätzlich lassen sich diese Methoden auch in Situationen verwenden, bei denen die Approximationsmöglichkeit durch eine Normalverteilung ausgenützt werden kann.

5.1 Der Mittelwert μ

Die Verfahren zur Schätzung des Mittelwertparameters μ einer normalverteilten *ZG* X unterscheiden sich dadurch, ob die Standardabweichung als bekannt vorausgesetzt werden kann oder nicht. Ist $\sigma = \sigma_0$ bekannt, so verwendet man die Tatsache, dass

$$\frac{\bar{X} - \mu}{\sigma_0/\sqrt{n}} \tag{5.1}$$

standardnormalverteilt ist. Falls σ unbekannt ist, nützt man aus, dass *unabhängig* von $\bar{X}$ und μ

$$\frac{(n-1)S^2}{\sigma^2}$$

χ^2–verteilt ist mit $n-1$ Freiheitsgraden. Zusammen mit (5.1) und der Definition der t–Verteilung ist daher

$$\frac{\frac{\bar{X}-\mu}{\sigma/\sqrt{n}}}{\sqrt{\frac{(n-1)S^2}{\sigma^2}/(n-1)}} = \frac{\bar{X}-\mu}{S/\sqrt{n}} \tag{5.2}$$

t–verteilt mit $n-1$ Freiheitsgraden. Den darauf basierenden Test nennt man häufig (Einstichproben–) t–Test.

Die Tabelle 5.1 zeigt nun zu diesen beiden Situationen Konfidenzintervalle für μ, die nach dem in Kapitel 4 erläuterten Prinzip konstruiert sind.

Beispiel 5.1 In einer Stichprobe von Äpfeln aus einer Lieferung der Sorte Golddelicious wird das Apfelgewicht (in g) gemessen (vergl. Beispiel 1.1):

147 160 158 156 174 162 160 158 139 135

Tabelle 5.1: *Konfidenzintervalle für* μ

σ	Grenzen
bekannt $\sigma = \sigma_0$	$\left.\begin{matrix}\mu_o \\ \mu_u\end{matrix}\right\} = \bar{x} \pm z_{1-\alpha/2}\frac{\sigma_0}{\sqrt{n}}$
	$\begin{matrix}\mu_o & = & \bar{x} + z_{1-\alpha}\frac{\sigma_0}{\sqrt{n}} \\ \mu_u & = & -\infty\end{matrix}$
	$\begin{matrix}\mu_o & = & \infty \\ \mu_u & = & \bar{x} - z_{1-\alpha}\frac{\sigma_0}{\sqrt{n}}\end{matrix}$
unbekannt	$\left.\begin{matrix}\mu_o \\ \mu_u\end{matrix}\right\} = \bar{x} \pm t_{n-1;1-\alpha/2}\frac{s}{\sqrt{n}}$
	$\begin{matrix}\mu_o & = & \bar{x} + t_{n-1;1-\alpha}\frac{s}{\sqrt{n}} \\ \mu_u & = & -\infty\end{matrix}$
	$\begin{matrix}\mu_o & = & \infty \\ \mu_u & = & \bar{x} - t_{n-1;1-\alpha}\frac{s}{\sqrt{n}}\end{matrix}$

Daraus ergeben sich

$$\begin{matrix}\bar{x} & = & 154.9 \\ s^2 & = & 133.21 \\ s & = & 11.5\end{matrix}$$

als Schätzwerte für μ und σ^2 bzw. σ. Falls die (theoretische) Standardabweichung mit $\sigma_0 = 10\ g$ als bekannt vorausgesetzt werden kann, erhält man mit

$$\left.\begin{matrix}\mu_o \\ \mu_u\end{matrix}\right\} = 154.9 \pm 1.96 \times \frac{10}{\sqrt{10}} = \left\{\begin{matrix}161.1 \\ 148.7\end{matrix}\right.$$

ein zweiseitiges 95 %–Konfidenzintervall für μ.

Im anderen Fall ergibt sich etwa ein einseitig nach unten begrenztes *K–Int* mit $t_{n-1;1-\alpha} = t_{9;0.95} = 1.833$ als

$$\mu \geq \mu_u = 154.9 - 1.833 \times \frac{11.5}{\sqrt{10}} = 148.2$$

Die Tabelle 5.2 enthält die Grenzen des jeweiligen Annahmebereiches für die Tests zu den angegebenen Nullhypothesen. Analog den Konfidenzintervallen ist der Unterschied im Lösungsansatz zu beachten, wenn die Standardabweichung bekannt ist oder durch s geschätzt werden muss.

Tabelle 5.2: *Tests für* μ

σ	Teststatistik t	
	H_0	Annahmebereich (c_u, c_o)
bekannt $\sigma = \sigma_0$	$t = \frac{\bar{x}-\mu_0}{\sigma/\sqrt{n}}$	
	$\mu = \mu_0$	$\left.\begin{matrix} c_o \\ c_u \end{matrix}\right\} = \pm z_{1-\alpha/2}$
	$\mu \leq \mu_0$	$c_o = z_{1-\alpha}$, $c_u = -\infty$
	$\mu \geq \mu_0$	$c_o = \infty$, $c_u = -z_{1-\alpha}$
unbekannt	$t = \frac{\bar{x}-\mu_0}{s/\sqrt{n}}$	
	$\mu = \mu_0$	$\left.\begin{matrix} c_o \\ c_u \end{matrix}\right\} = \pm t_{n-1;1-\alpha/2}$
	$\mu \leq \mu_0$	$c_o = t_{n-1;1-\alpha}$, $c_u = -\infty$
	$\mu \geq \mu_0$	$c_o = \infty$, $c_u = -t_{n-1;1-\alpha}$

Beispiel 5.2 Im Anschluss an Bsp. 5.1 ist die Hypothese $H_0 : \ \mu \geq \mu_0 = 155$ mit der Sicherheit $1 - \alpha = 0.95$ zu testen, wobei die Standardabweichung als bekannt vorausgesetzt werden kann. Die Teststatistik ergibt

$$t = \frac{\bar{x} - \mu_0}{\sigma_0/\sqrt{n}} = \frac{154.9 - 155}{10/\sqrt{10}} = -0.032$$

und liegt damit über dem (unteren) kritischen Wert

$$c_u = -z_{1-\alpha} = -1.645 \ ,$$

sodass die Nullhypothese nicht verworfen werden kann. Es gibt damit *keinen* signifikanten Einwand gegen die Behauptung, dass $\mu \geq 155$ gilt ("Der Stichprobenmittelwert 154.9 g ist nur *zufällig* kleiner als 155 g").

Für einen Test der Nullhypothese $H_0 : \ \mu = \mu_0 = 160$ zum Signifikanzniveau $\alpha = 0.05$

ohne Kenntnis über die (theoretische) Standardabweichung bildet man die Testgröße

$$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} = \frac{154.9 - 160}{11.5/\sqrt{10}} = -1.402 \ ;$$

auch diese liegt im Annahmebereich

$$\left.\begin{matrix} c_o \\ c_u \end{matrix}\right\} = \pm t_{n-1;1-\alpha/2} = \pm t_{9;0.975} = \pm 2.262 \ ,$$

sodass auch hier die Nullhypothese beibehalten wird.

5.2 Die Varianz σ^2

Die Herleitung von Konfidenzintervallen für σ^2 (bzw. σ) und von Tests für entsprechende Hypothesen beruht auf der Tatsache, dass

$$\frac{(n-1)S^2}{\sigma^2} \ \sim \ \chi^2_{n-1} \tag{5.3}$$

gilt. In Tabelle 5.3 sind Konfidenzintervalle für σ^2 zusammengestellt, aus denen man durch Wurzelziehen sofort welche für σ erhält. Die Tabelle 5.4 enthält die Annahmebereiche für Tests von Hypothesen über σ^2 bzw. σ.

Tabelle 5.3: *Konfidenzintervalle für* σ^2

Grenzen	
σ^2_u	σ^2_o
$(n-1)s^2/\chi^2_{n-1;1-\alpha/2}$	$(n-1)s^2/\chi^2_{n-1;\alpha/2}$
0	$(n-1)s^2/\chi^2_{n-1;\alpha}$
$(n-1)s^2/\chi^2_{n-1;1-\alpha}$	∞

Ein zweiseitiges $(1-\alpha)$–*K–Int* ergibt sich dabei wegen (5.3)

$$P\left(\chi^2_{n-1;\alpha/2} \le \frac{(n-1)S^2}{\sigma^2} \le \chi^2_{n-1;1-\alpha/2}\right) = 1-\alpha$$

bzw. nach Umformung der Doppelungleichung

$$P\left(\frac{(n-1)S^2}{\chi^2_{n-1;1-\alpha/2}} \le \sigma^2 \le \frac{(n-1)S^2}{\chi^2_{n-1;\alpha/2}}\right) = 1-\alpha$$

als

$$\left(\frac{(n-1)S^2}{\chi^2_{n-1;1-\alpha/2}} \ , \ \frac{(n-1)S^2}{\chi^2_{n-1;\alpha/2}}\right) \tag{5.4}$$

Tabelle 5.4: *Tests für* σ^2

H_0	Teststatistik t	
	Annahmebereich (c_u, c_o)	
	c_u	c_o
	$t = \frac{(n-1)\,s^2}{\sigma_0^2}$	
$\sigma^2 = \sigma_0^2$	$\chi^2_{n-1;\alpha/2}$	$\chi^2_{n-1;1-\alpha/2}$
$\sigma^2 \leq \sigma_0^2$	0	$\chi^2_{n-1;1-\alpha}$
$\sigma^2 \geq \sigma_0^2$	$\chi^2_{n-1;\alpha}$	∞

bzw. eines für σ durch einfaches Wurzelziehen:

$$\left(\sqrt{\frac{(n-1)S^2}{\chi^2_{n-1;1-\alpha/2}}} \, , \, \sqrt{\frac{(n-1)S^2}{\chi^2_{n-1;\alpha/2}}} \, . \right) \tag{5.5}$$

Beispiel 5.3 Zu den Daten aus Beispiel 5.1 erhält man ein zweiseitiges 95 %–*K–Int* für σ^2 als

$$\begin{aligned} \sigma_o^2 &= \frac{(n-1)s^2}{\chi^2_{n-1;\alpha/2}} = \frac{9 \times s^2}{\chi^2_{9;0.025}} = \frac{9 \times 133.21}{2.70} = 444.03 \\ \sigma_u^2 &= \frac{(n-1)s^2}{\chi^2_{n-1;1-\alpha/2}} = \frac{9 \times s^2}{\chi^2_{9;0.975}} = \frac{9 \times 133.21}{19.02} = 62.44 \end{aligned}$$

bzw. durch

$$\begin{aligned} \sigma_o &= \sqrt{444.03} = 21.1 \\ \sigma_u &= \sqrt{62.44} = 7.9 \end{aligned}$$

eines für σ.

Zur Überprüfung der Nullhypothese $H_0: \ \sigma \leq 10\ g$ auf dem Signifikanzniveau $\alpha = 0.05$ (also Sicherheit $1-\alpha = 0.95$) beachtet man zunächst, dass diese Hypothese der Nullhypothese $H_0': \ \sigma^2 \leq 100\ [g^2]$ für σ^2 entspricht. Zunächst berechnet man die Testgröße

$$t = \frac{(n-1)s^2}{\sigma_0^2} = \frac{9 \times 133.21}{10^2} = 11.989$$

und vergleicht diese mit dem (oberen) kritischen Wert

$$c_o = \chi^2_{n-1;1-\alpha} = \chi^2_{9;0.95} = 16.92 \ ;$$

Offensichtlich ist $t < c_o$, sodass die Nullhypothese beibehalten wird. Es besteht also *kein signifikanter Einwand gegen* die Hypothese $\sigma \leq 10\ g$.

5.3 Vergleich zweier Mittelwerte

Für zwei Merkmale X und Y, die nach $N(\mu_X, \sigma_X^2)$ und $N(\mu_Y, \sigma_Y^2)$ verteilt vorausgesetzt werden, interessiert der Vergleich ihrer Mittelwerte μ_X und μ_Y.

5.3.1 Unabhängige Stichproben

Liegen zwei unabhängig gewonnene Stichproben $x_1, \ldots, x_{n_X}$ und $y_1, \ldots, y_{n_Y}$ für diese Merkmale vor, so wird der Unterschied naheliegenderweise durch die Differenz $\bar{x} - \bar{y}$ geschätzt. Diese Schätzung ist offensichtlich erwartungstreu.

Ein praktisches Indiz für die Unabhängigkeit ist der häufig unterschiedlich große Stichprobenumfang der Proben. Darüberhinaus sprechen aber in der Regel sachliche Gründe, wie die Art der Probenahme, unterschiedliche Kollektive und Ähnliches für eine allfällige Unabhängigkeit der Proben. In diesem Fall lässt sich die Varianz der Differenz $\bar{X} - \bar{Y}$ mit Hilfe der Einzelvarianzen angeben.

a) σ_X^2 und σ_Y^2 bekannt:

In diesem Fall gilt

$$\bar{X} - \bar{Y} \;\sim\; N(\mu_X - \mu_Y, \frac{\sigma_X^2}{n_X} + \frac{\sigma_Y^2}{n_Y}) \quad , \tag{5.6}$$

da wegen der Unabhängigkeit der beiden Stichproben dasselbe für die Stichprobenmittelwerte $\bar{X}$ und $\bar{Y}$ zutrifft und $\bar{X} \sim N(\mu_X, \sigma_X^2/n_X)$ bzw. $\bar{Y} \sim N(\mu_Y, \sigma_Y^2/n_Y)$ gilt.

b) $\sigma_X^2 = \sigma_Y^2 = \sigma^2$ unbekannt:

Lässt sich die Gleichheit der beiden (unbekannten) Varianzen σ_X^2 und σ_Y^2 tatsächlich voraussetzen, dann kann die gemeinsame Varianz σ^2 durch den gewichteten Mittelwert der beiden Stichprobenvarianzen s_X^2 und s_Y^2

$$s^2 = \frac{(n_X - 1)s_X^2 + (n_Y - 1)s_Y^2}{n_X + n_Y - 2} \tag{5.7}$$

(engl. *pooled variance*) erwartungstreu geschätzt werden. Da $(n_X + n_Y - 2)S^2/\sigma^2$ dann nach $\chi^2_{n_X+n_Y-2}$ verteilt ist, gilt wegen (5.6) und 3.6.6 (t–Verteilung)

$$\frac{\frac{(\bar{X}-\bar{Y})-(\mu_X-\mu_Y)}{\sigma\sqrt{1/n_X+1/n_Y}}}{\sqrt{\frac{S^2}{\sigma^2}}} = \frac{(\bar{X} - \bar{Y}) - (\mu_X - \mu_Y)}{S\sqrt{\frac{1}{n_X} + \frac{1}{n_Y}}} \;\sim\; t_{n_X+n_Y-2} \; . \tag{5.8}$$

Der darauf aufbauende Test wird daher auch als Zweistichproben–t–Test bezeichnet.

c) σ_X^2 und σ_Y^2 unbekannt und beliebig:

Lässt sich bei unbekannter Varianz die Gleichheit von σ_X^2 und σ_Y^2 nicht gewährleisten, so kann die Tatsache ausgenützt werden, dass bei wachsenden Stichprobenumfängen

$$\frac{(\bar{X} - \bar{Y}) - (\mu_X - \mu_Y)}{\bar{S}_\Delta} \tag{5.9}$$

Tabelle 5.5: *Konfidenzintervall* $(\Delta\mu_u, \Delta\mu_o)$ *für* $\Delta\mu = \mu_X - \mu_Y$

σ_X und σ_Y	Grenzen
bekannt	$\left.\begin{matrix}\Delta\mu_o \\ \Delta\mu_u\end{matrix}\right\} = (\bar{x}-\bar{y}) \pm z_{1-\alpha/2}\sqrt{\frac{\sigma_X^2}{n_X}+\frac{\sigma_Y^2}{n_Y}}$
	$\begin{aligned}\Delta\mu_o &= (\bar{x}-\bar{y}) + z_{1-\alpha}\sqrt{\frac{\sigma_X^2}{n_X}+\frac{\sigma_Y^2}{n_Y}} \\ \Delta\mu_u &= -\infty\end{aligned}$
	$\begin{aligned}\Delta\mu_o &= \infty \\ \Delta\mu_u &= (\bar{x}-\bar{y}) - z_{1-\alpha}\sqrt{\frac{\sigma_X^2}{n_X}+\frac{\sigma_Y^2}{n_Y}}\end{aligned}$
unbekannt $\sigma_X = \sigma_Y = \sigma$	$\left.\begin{matrix}\Delta\mu_o \\ \Delta\mu_u\end{matrix}\right\} = (\bar{x}-\bar{y}) \pm s\cdot t_{f;1-\alpha/2}\sqrt{\frac{1}{n_X}+\frac{1}{n_Y}}$ $\begin{aligned}s &= \sqrt{\frac{(n_X-1)s_X^2+(n_Y-1)s_Y^2}{n_X+n_Y-2}} \\ f &= n_X+n_Y-2\end{aligned}$
	$\begin{aligned}\Delta\mu_o &= (\bar{x}-\bar{y}) + s\cdot t_{f;1-\alpha}\sqrt{\frac{1}{n_X}+\frac{1}{n_Y}} \\ \Delta\mu_u &= -\infty\end{aligned}$
	$\begin{aligned}\Delta\mu_o &= \infty \\ \Delta\mu_u &= (\bar{x}-\bar{y}) - s\cdot t_{f;1-\alpha}\sqrt{\frac{1}{n_X}+\frac{1}{n_Y}}\end{aligned}$
unbekannt σ_X, σ_Y beliebig	$\left.\begin{matrix}\Delta\mu_o \\ \Delta\mu_u\end{matrix}\right\} = (\bar{x}-\bar{y}) \pm s\cdot t_{f;1-\alpha/2}$ $\begin{aligned}s &= \sqrt{\frac{s_X^2}{n_X}+\frac{s_Y^2}{n_Y}} \\ f &= \frac{\left(\frac{s_X^2}{n_X}+\frac{s_Y^2}{n_Y}\right)^2}{\left(\frac{s_X^2}{n_X}\right)^2/(n_X-1)+\left(\frac{s_Y^2}{n_Y}\right)^2/(n_Y-1)}\end{aligned}$
	$\begin{aligned}\Delta\mu_o &= (\bar{x}-\bar{y}) + s\cdot t_{f;1-\alpha} \\ \Delta\mu_u &= -\infty\end{aligned}$
	$\begin{aligned}\Delta\mu_o &= \infty \\ \Delta\mu_u &= (\bar{x}-\bar{y}) - s\cdot t_{f;1-\alpha}\end{aligned}$

Tabelle 5.6: *Tests für die Differenz* $\Delta\mu = \mu_X - \mu_Y$

σ_X und σ_Y	Teststatistik t	
	H_0	Annahmebereich (c_u, c_o)
bekannt	$t = \dfrac{\bar{x} - \bar{y}}{\sqrt{\frac{\sigma_X^2}{n_X} + \frac{\sigma_Y^2}{n_Y}}}$	
	$\mu_X = \mu_Y$ $(\mu_X - \mu_Y = 0)$	$\left.\begin{array}{l} c_o \\ c_u \end{array}\right\} = \pm z_{1-\alpha/2}$
	$\mu_X \leq \mu_Y$ $(\mu_X - \mu_Y \leq 0)$	$c_o = z_{1-\alpha}$ $c_u = -\infty$
	$\mu_X \geq \mu_Y$ $(\mu_X - \mu_Y \geq 0)$	$c_o = \infty$ $c_u = -z_{1-\alpha}$
unbekannt $\sigma_X = \sigma_Y = \sigma$	$t = \dfrac{\bar{x} - \bar{y}}{s \times \sqrt{\frac{1}{n_X} + \frac{1}{n_Y}}}$ $s = \sqrt{\frac{(n_X-1)s_X^2 + (n_Y-1)s_Y^2}{n_X+n_Y-2}}$ $f = n_X + n_Y - 2$	
	$\mu_X = \mu_Y$ $(\mu_X - \mu_Y = 0)$	$\left.\begin{array}{l} c_o \\ c_u \end{array}\right\} = \pm t_{f;1-\alpha/2}$
	$\mu_X \leq \mu_Y$ $(\mu_X - \mu_Y \leq 0)$	$c_o = t_{f;1-\alpha}$ $c_u = -\infty$
	$\mu_X \geq \mu_Y$ $(\mu_X - \mu_Y \geq 0)$	$c_o = \infty$ $c_u = -t_{f;1-\alpha}$
unbekannt σ_X, σ_Y beliebig	$t = \dfrac{\bar{x} - \bar{y}}{\sqrt{\frac{s_X^2}{n_X} + \frac{s_Y^2}{n_Y}}}$ $f = \dfrac{\left(\frac{s_X^2}{n_X} + \frac{s_Y^2}{n_Y}\right)^2}{\left(\frac{s_X^2}{n_X}\right)^2 / (n_X - 1) + \left(\frac{s_Y^2}{n_Y}\right)^2 / (n_Y - 1)}$	
	$\mu_X = \mu_Y$ $(\mu_X - \mu_Y = 0)$	$\left.\begin{array}{l} c_o \\ c_u \end{array}\right\} = \pm t_{f;1-\alpha/2}$
	$\mu_X \leq \mu_Y$ $(\mu_X - \mu_Y \leq 0)$	$c_o = t_{f;1-\alpha}$ $c_u = -\infty$
	$\mu_X \geq \mu_Y$ $(\mu_X - \mu_Y \geq 0)$	$c_o = \infty$ $c_u = -t_{f;1-\alpha}$

mit

$$\bar{S}_\Delta = \sqrt{\frac{S_X^2}{n_X} + \frac{S_Y^2}{n_Y}}$$

immer besser durch eine t–Verteilung mit f Freiheitsgraden beschrieben wird, wobei

$$f = \frac{\left(\frac{s_X^2}{n_X} + \frac{s_Y^2}{n_Y}\right)^2}{\left(\frac{s_X^2}{n_X}\right)^2 \Big/ (n_X - 1) + \left(\frac{s_Y^2}{n_Y}\right)^2 \Big/ (n_Y - 1)} \tag{5.10}$$

gilt (siehe Hartung, 1990). Diese Aufgabe ist in der Literatur als Behrens–Fisher–Problem bekannt (siehe Scheffe, 1970).

Damit ergeben sich für die Schätzung der Differenz $\mu_X - \mu_Y$ je nach Kenntnis der theoretischen Varianzen die in Tabelle 5.5 zusammengestellten Konfidenzintervalle zur Überdeckungswahrscheinlichkeit $1 - \alpha$. Für die über μ_X und μ_Y zu formulierenden Hypothesen bieten sich die in Tabelle 5.6 aufgelisteten Tests zum Signifikanznivau α an.

Beispiel 5.4 Golddelicious–Äpfel werden einer Handelskette von zwei Äpfelplantagen angeboten. Als Entscheidungsgrundlage für einen Liefervertrag werden von jeder Plantage einige Äpfel geprüft und ihr Gewicht (in g) gemessen (vergl. Beispiel 1.1):

Plantage A	147	160	158	156	174
	162	160	158	139	135
Plantage B	157	146	135	126	134
	172	170	155		

Daraus ergeben sich

$$\begin{array}{rclrcl} \bar{x} &=& 154.9 & \bar{y} &=& 149.4 \\ s_X^2 &=& 133.21 & s_Y^2 &=& 289.70 \\ s_X &=& 11.5 & s_Y &=& 17.0 \end{array}$$

Angenommen, dass $\sigma_X = 10\,g$ und $\sigma_Y = 15\,g$ als bekannt vorausgesetzt werden können. Dann erhält man mit $z_{0.975} = 1.96$ durch

$$\begin{aligned} \left.\begin{array}{l}\Delta\mu_o \\ \Delta\mu_u\end{array}\right\} &= (\bar{x} - \bar{y}) \pm z_{1-\alpha/2}\sqrt{\frac{\sigma_X^2}{n_X} + \frac{\sigma_Y^2}{n_Y}} \\ &= (154.9 - 149.4) \pm 1.96 \times \sqrt{\frac{10^2}{10} + \frac{15^2}{8}} \\ &= 5.5 \pm 1.96 \times 6.2 = 5.5 \pm 12.2 = \left\{\begin{array}{r} 17.7 \\ -6.7 \end{array}\right. \end{aligned}$$

ein zweiseitiges 95 %–Konfidenzintervall für die Differenz $\Delta\mu = \mu_X - \mu_Y$ der Mittelwerte.

Sind die (theoretischen) Standardabweichungen unbekannt, aber können sie als gleich vorausgesetzt werden, so ergibt sich ein nach oben hin beschränktes einseitiges *K–Int* mit $t_{16;0.95} = 1.746$ und

$$s = \sqrt{\frac{(n_X - 1)s_X^2 + (n_Y - 1)s_Y^2}{n_X + n_Y - 2}} = \sqrt{\frac{9 \times 133.21 + 7 \times 289.70}{16}} = 14.2$$

als

$$\begin{aligned}\Delta\mu &\leq (\bar{x}-\bar{y}) + s \cdot t_{n_X+n_Y-2;1-\alpha}\sqrt{\frac{1}{n_X}+\frac{1}{n_Y}} \\ &= (154.9-149.4) + 14.2 \times 1.746 \times 0.474 = 17.3\ .\end{aligned}$$

Schließlich soll ohne jede Kenntnis über die Standardverteilungen von X und Y die Hypothese $H_0:\ \mu_X < \mu_Y$ untersucht werden. Dazu benötigt man die Testgröße

$$t = \frac{\bar{x}-\bar{y}}{\sqrt{\frac{s_X^2}{n_X}+\frac{s_Y^2}{n_Y}}} = \frac{154.9-149.4}{\sqrt{\frac{133.21}{10}+\frac{289.70}{8}}} = 0.785\ ,$$

die mit dem (oberen) kritischen Wert

$$c_o = t_{f;1-\alpha} = t_{11.85;0.95} = 1.783$$

zu vergleichen ist, wobei sich die Freiheitsgrade aus

$$f = \frac{\left(\frac{133.21}{10}+\frac{289.70}{8}\right)^2}{\left(\frac{133.21}{10}\right)^2/9 + \left(\frac{289.70}{8}\right)^2/7} = 11.85$$

ergeben (Achtung: meist nicht ganzzahlige Freiheitsgrade bedingen eine Interpolation in den Tabellen oder eine Rundung der Freiheitsgrade). Da der Wert der Teststatistik unter dem kritischen Wert liegt, kann die Nullhypothese *nicht verworfen* werden, es besteht also *kein signifikanter Einwand* gegen die Behauptung $\mu_X < \mu_Y$.

Nichtparametrische Alternativen für die hier verwendeten Tests sind der Kolmogorov–Smirnov–Zweistichprobentest und der Wilcoxon–Test, die in Kapitel 8 beschrieben sind.

5.3.2 Abhängige Stichproben

Im Falle abhängiger Merkmale X und Y liegen *paarweise* Stichproben (x_1, y_1), (x_2, y_2), ..., (x_n, y_n) vor (z.B. Körpergröße am Morgen und am Abend an *einer* Person, Umsatz 1991 und 1992 *einer* Firma, durchschnittliche Milchleistung 1991 bzw. 1992 bei *einem* Landwirt). Für Fragen nach dem Verhältnis der Mittelwerte μ_X und μ_Y zueinander betrachtet man hier die Differenzen $x_i - y_i$ $(i = 1, \ldots, n)$. Daher wird die Methode gerne als *Differenzenmethode* bezeichnet.

a) $X - Y$ normalverteilt:

Falls $X - Y$ normalverteilt vorausgesetzt werden kann (Achtung: das ist nicht selbstverständlich!), wird die (normalverteilte) Differenz $D = X - Y$ anhand der Stichprobe der Differenzen $d_i = x_i - y_i$ $(i = 1, \ldots, n)$ untersucht. Da die Standardabweichung σ_D nicht bekannt ist, sind die Verfahren zur Beurteilung eines normalverteilten Merkmals ohne Kenntnis der Standardabweichung anzuwenden (siehe 5.1).

Für die erwartungstreue Schätzung von $\Delta\mu = \mu_X - \mu_Y$ verwendet man auch hier $\bar{x} - \bar{y}$. Die Konfidenzintervalle für $\Delta\mu$ finden sich in Tabelle 5.7, Tests über μ_X und μ_Y sind in Tabelle 5.8 zusammengestellt und

$$s_D^2 = \frac{1}{n-1}\sum_{i=1}^{n}(d_i - \bar{d})^2 \tag{5.11}$$

Tabelle 5.7: *Konfidenzintervall* $(\Delta\mu_u, \Delta\mu_o)$ *für* $\Delta\mu = \mu_X - \mu_Y$

$\Delta\mu_u$	$\Delta\mu_o$
$(\bar{x} - \bar{y}) - t_{n-1;1-\alpha/2}\frac{s_D}{\sqrt{n}}$	$(\bar{x} - \bar{y}) + t_{n-1;1-\alpha/2}\frac{s_D}{\sqrt{n}}$
$-\infty$	$(\bar{x} - \bar{y}) + t_{n-1;1-\alpha}\frac{s_D}{\sqrt{n}}$
$(\bar{x} - \bar{y}) - t_{n-1;1-\alpha}\frac{s_D}{\sqrt{n}}$	∞

Tabelle 5.8: *Tests für die Differenz* $\Delta\mu = \mu_X - \mu_Y$

Teststatistik t		
H_0	Annahmebereich (c_u, c_o)	
	c_u	c_o
$t = \frac{\bar{x}-\bar{y}}{s_D/\sqrt{n}}$		
$\mu_X = \mu_Y$ $(\mu_X - \mu_Y = 0)$	$-t_{n-1;1-\alpha/2}$	$t_{n-1;1-\alpha/2}$
$\mu_X \leq \mu_Y$ $(\mu_X - \mu_Y \leq 0)$	$-\infty$	$t_{n-1;1-\alpha}$
$\mu_X \geq \mu_Y$ $(\mu_X - \mu_Y \geq 0)$	$-t_{n-1;1-\alpha}$	∞

steht für die Stichprobenvarianz der d_i–Werte, mit deren Hilfe die unbekannte Varianz σ_D^2 der Differenz $X - Y$ geschätzt wird.

Beispiel 5.5 Für die Agrarstatistik wird im Rahmen einer Befragung das bäuerliche Jahreseinkommen (in 1.000 €) für die Jahre 2000 und 2001 erhoben:

Landwirt / Jahr	1	2	3	4	5	6
2000	18.47	19.48	11.94	16.39	8.45	32.50
2001	20.54	20.09	13.01	17.33	9.10	32.97
d_{01-00}	2.07	0.61	1.07	0.94	0.65	0.47

Hat sich das Jahreseinkommen von 2000 auf 2001 erhöht (Signifikanzniveau $\alpha = 0.05$)? Wie lautet ein zweiseitiges 95 %–Konfidenzintervall für den durchschnittlichen Einkommenszuwachs? Es scheint nichts gegen eine normalverteilte Einkommensdifferenz zwischen den Jahren 2000 und 2001 zu sprechen (das könnte man übrigens mittels QQ-Diagramm überprüfen!)

Zunächst benötigt man die Einkommensdifferenzen d_{01-00}, die in der Angabe bereits berechnet sind. Daraus erhält man

$$\bar{x}_{01} - \bar{x}_{00} = \bar{d}_{01\text{-}00} = 0.968 \qquad \text{und} \qquad s_{D_{01\text{-}00}} = 0.584$$

Die Grenzen des geforderten Konfidenzintervalles ergeben sich nun unter Verwendung von Tabelle 5.7 zu

$$\begin{aligned}\Delta\mu_o &= (\bar{x}_{01} - \bar{x}_{00}) + s_{D_{01\text{-}00}} \cdot t_{5;0.975}/\sqrt{6} \\ &= (18.840 - 17.872) + 0.584 \times 2.571/2.449 = 1.581\end{aligned}$$

und analog $\Delta\mu_u = 0.356$. Zur Überprüfung der Nullhypothese H_0: $\mu_{01} \leq \mu_{00}$ (Gegenteil!), oder gleichbedeutend H_0': $\Delta\mu \leq 0$, benötigt man die Testgröße

$$t = \frac{\bar{x}_{01} - \bar{x}_{00}}{s_{D_{01\text{-}00}}/\sqrt{n}} = \frac{0.968}{0.584/2.449} = 4.059 \ .$$

Große Werte für diese Testgröße sprechen gegen die Nullhypothese (mittlere Zeile in Tabelle 5.8); daher benötigt man einen *oberen* kritischen Wert $t_{5;0.95} = 2.015$. Wegen $4.059 > 2.015$ kann die Nullhypothese verworfen werden, es gab also im Jahr 2001 einen signifikanten Einkommenszuwachs.

b) $X - Y$ beliebig verteilt:

Für eine beliebig verteilte Differenz $X - Y$ kann man bei großen Stichprobenumfängen (Faustregel: $n \geq 30$) den Zentralen Grenzverteilungssatz anwenden und $\bar{D}$ als annähernd normalverteilt ansehen. Für die unbekannte Varianz σ_D^2/n von $\bar{D}$ wählt man einfach s_D^2/n. Da in diesem Fall unter der Nullhypothese *keine* t–Verteilung mehr vorliegt, kann man gleich den nächsten Approximationsschritt setzen und

$$\frac{\bar{D} - \Delta\mu}{s_D/\sqrt{n}}$$

als annähernd $N(0,1)$–standardnormalverteilt auffassen, woraus sich entsprechende Schätz– und Testformeln herleiten lassen.

Für die erwartungstreue Schätzung von $\Delta\mu = \mu_X - \mu_Y$ verwendet man wieder $\bar{x} - \bar{y}$. Die (näherungsweisen) Konfidenzintervalle für $\Delta\mu$ finden sich in Tabelle 5.9 und entsprechende Tests über μ_X und μ_Y sind in Tabelle 5.10 zusammengestellt.

Beispiel 5.6 In der Erhebung aus dem vorigen Beispiel wurden natürlich mehr Landwirte befragt; die folgende Tabelle zeigt einen Ausschnitt davon und bringt das Ergebnis von 30 Befragten.

Die Fragen stellen sich analog dem vorigen Beispiel: hat sich das Jahreseinkommen von 2000 auf 2001 erhöht (Signifikanzniveau $\alpha = 0.05$)? Wie lautet ein zweiseitiges 95 %–Konfidenzintervall für den durchschnittlichen Einkommenszuwachs? Über eine allfällige Normalität der Einkommensdifferenz zwischen den Jahren 2000 und 2001 ist hier nichts bekannt.

Tabelle 5.9: *Konfidenzintervall* $(\Delta\mu_u, \Delta\mu_o)$ *für* $\Delta\mu = \mu_X - \mu_Y$

$\Delta\mu_u$	$\Delta\mu_o$
$(\bar{x} - \bar{y}) - z_{1-\alpha/2}\frac{s_D}{\sqrt{n}}$	$(\bar{x} - \bar{y}) + z_{1-\alpha/2}\frac{s_D}{\sqrt{n}}$
$-\infty$	$(\bar{x} - \bar{y}) + z_{1-\alpha}\frac{s_D}{\sqrt{n}}$
$(\bar{x} - \bar{y}) - z_{1-\alpha}\frac{s_D}{\sqrt{n}}$	∞

Tabelle 5.10: *Tests für die Differenz* $\Delta\mu = \mu_X - \mu_Y$

Teststatistik t		
H_0	Annahmebereich (c_u, c_o)	
	c_u	c_o
$t = \frac{\bar{x}-\bar{y}}{s_D/\sqrt{n}}$		
$\mu_X = \mu_Y$ $(\mu_X - \mu_Y = 0)$	$-z_{1-\alpha/2}$	$z_{1-\alpha/2}$
$\mu_X \leq \mu_Y$ $(\mu_X - \mu_Y \leq 0)$	$-\infty$	$z_{1-\alpha}$
$\mu_X \geq \mu_Y$ $(\mu_X - \mu_Y \geq 0)$	$-z_{1-\alpha}$	∞

LW-Nr.	2000	2001	$d_{01\text{-}00}$	LW-Nr.	2000	2001	$d_{01\text{-}00}$
1	10.14	11.35	1.21	16	21.02	21.48	0.46
2	20.33	20.78	0.45	17	16.63	17.50	0.87
3	13.64	14.71	1.07	18	12.85	13.53	0.68
4	13.92	13.79	-0.13	19	17.21	18.13	0.92
5	7.98	8.77	0.79	20	12.55	12.86	0.31
6	23.47	23.93	0.46	21	9.43	10.30	0.87
7	16.94	17.85	0.91	22	24.53	25.00	0.47
8	11.73	11.90	0.17	23	18.56	19.64	1.08
9	29.20	29.89	0.69	24	32.55	33.36	0.81
10	16.77	17.14	0.37	25	18.47	20.54	2.07
11	10.04	11.08	1.04	26	19.48	20.09	0.61
12	7.28	7.08	-0.20	27	11.94	13.01	1.07
13	9.32	9.98	0.66	28	16.39	17.33	0.94
14	15.86	16.13	0.27	29	8.45	9.10	0.65
15	12.22	13.28	1.06	30	32.50	32.97	0.47

Zunächst benötigt man die Einkommensdifferenzen d_{01-00}, die wiederum in der Angabe bereits berechnet sind. Daraus erhält man

$$\bar{x}_{01} - \bar{x}_{00} = \bar{d}_{01\text{-}00} = 0.703 \qquad \text{und} \qquad s_{D_{01\text{-}00}} = 0.437$$

Die Grenzen des geforderten Konfidenzintervalles ergeben sich nun unter Verwendung von Tabelle 5.9 zu

$$\begin{aligned}\Delta\mu_o &= (\bar{x}_{01} - \bar{x}_{00}) + s_{D_{01\text{-}00}} \cdot z_{0.975}/\sqrt{30} \\ &= (17.083 - 16.380) + 0.437 \times 1.960/5.477 = 0.859\end{aligned}$$

und analog $\Delta\mu_u = 0.547$. Zur Überprüfung der Nullhypothese H_0: $\mu_{01} \leq \mu_{00}$ (Gegenteil!), oder gleichbedeutend H_0': $\Delta\mu \leq 0$, benötigt man die dieselbe Testgröße wie im vorigen Beispiel

$$t = \frac{\bar{x}_{01} - \bar{x}_{00}}{s_{D_{01\text{-}00}}/\sqrt{n}} = \frac{0.703}{0.437/5.477} = 8.811 \ .$$

Das Entscheidungskriterium bleibt natürlich dasselbe (mittlere Zeile in Tabelle 5.10); man benötigt wieder einen *oberen* kritischen Wert $z_{0.95} = 1.645$. Wegen $8.811 > 1.645$ kann die Nullhypothese verworfen werden, es gab also im Jahr 2001 einen signifikanten Einkommenszuwachs.

Eine nichtparametrische Alternative für diese Tests ist der Vorzeichentest, der ebenfalls in Kapitel 8 beschrieben ist.

5.4 Vergleich zweier Varianzen

Zwei betrachtete Merkmale X und Y werden als unabhängig $N(\mu_X, \sigma_X^2)$– bzw. $N(\mu_Y, \sigma_Y^2)$–verteilt vorausgesetzt. Der Vergleich der beiden (theoretischen) Varianzen erfolgt anhand des Quotienten s_X^2/s_Y^2 der beiden Stichprobenvarianzen. Die Tests basieren auf der Tatsache, dass im Falle $\sigma_X^2 = \sigma_Y^2$

$$\frac{S_X^2}{S_Y^2} = \frac{\frac{(n_X-1)S_X^2}{\sigma_X^2}\Big/(n_X-1)}{\frac{(n_Y-1)S_Y^2}{\sigma_Y^2}\Big/(n_Y-1)} \sim F_{n_X-1,n_Y-1} \tag{5.12}$$

gilt *(F–Test)*.

Einen nützlichen Sachverhalt für die Tabellierung von Quantilen der F–Verteilung beschreibt die folgende Aussage:

Es gilt:

$$F_{n_X-1,n_Y-1;\gamma} = \frac{1}{F_{n_Y-1,n_X-1;1-\gamma}} \tag{5.13}$$

Beweis:
Sind Y_1 und Y_2 zwei unabhängige, χ^2–verteilte *ZGen* mit f_1 bzw. f_2 Freiheitsgraden, so gilt bekanntlich

$$\frac{Y_1/f_1}{Y_2/f_2} \sim F_{f_1,f_2} \ .$$

Also gilt

$$P\left(\frac{Y_1/f_1}{Y_2/f_2} \leq F_{f_1,f_2;\gamma}\right) = \gamma$$

und damit auch

$$P\left(\frac{Y_2/f_2}{Y_1/f_1} \geq \frac{1}{F_{f_1,f_2;\gamma}}\right) = \gamma$$

bzw.

$$P\left(\frac{Y_2/f_2}{Y_1/f_1} < \frac{1}{F_{f_1,f_2;\gamma}}\right) = 1-\gamma \ .$$

Nun gilt aber

$$\frac{Y_2/f_2}{Y_1/f_1} \ \sim \ F_{f_2,f_1}$$

und damit

$$\frac{1}{F_{f_1,f_2;\gamma}} = F_{f_2,f_1;1-\gamma} \ .$$

△

Bemerkung: Tabellen für die F–Verteilung enthalten daher in der Regel nur Quantile mit Wahrscheinlichkeiten über 0.5 , aus denen man nach (5.13) auch Quantile zu Wahrscheinlichkeiten unter 0.5 erhält.

Sinnvolle Hypothesen betreffen den unmittelbaren Vergleich von σ_X und σ_Y . Bei Gleichheit dieser Varianzen rechnet man mit einem Stichprobenvarianzquotienten in der Nähe von 1, weit abliegende Werte gelten als signifikant. In Tabelle 5.11 finden sich die Annahmebereiche von Tests für diese Hypothesen.

Tabelle 5.11: *Tests für Varianzvergleich*

Teststatistik t		
H_0	Annahmebereich (c_u, c_o)	
	c_u	c_o
$t = \frac{s_X^2}{s_Y^2}$		
$\sigma_X = \sigma_Y$	$F_{n_X-1,n_Y-1;\alpha/2}$	$F_{n_X-1,n_Y-1;1-\alpha/2}$
$\sigma_X \leq \sigma_Y$	0	$F_{n_X-1,n_Y-1;1-\alpha}$
$\sigma_X \geq \sigma_Y$	$F_{n_X-1,n_Y-1;\alpha}$	∞

Bei den *zweiseitigen* Tests gibt es eine Vereinfachung, wenn man bedenkt, dass der obere kritische Wert stets größer als eins und der untere stets kleiner als eins ausfällt. Damit reicht es, die Testgröße mit c_o zu vergleichen, wenn sie größer als eins ist, und nur mit c_u zu vergleichen, wenn sie kleiner als eins ausfällt. In letzterem Fall ist aber

$$F_{n_X-1,n_Y-1;\alpha/2} \leq \frac{s_X^2}{s_Y^2}$$

gleichbedeutend mit

$$F_{n_Y-1,n_X-1;1-\alpha/2} = \frac{1}{F_{n_X-1,n_Y-1;\alpha/2}} \geq \frac{s_Y^2}{s_X^2} \ ,$$

wobei der rechte Term nun größer als eins ist. Für die Durchführung eines *zweiseitigen* Tests zweier Varianzen gilt daher folgende

Regel:

Die Reihung der Stichproben ist so festzusetzen, dass die zu berechnende Testgröße *größer als eins* ausfällt. Dieser Wert ist dann bloß mit dem entsprechenden *oberen* kritischen Wert (des zweiseitigen Tests, also mit dem $1-\alpha/2$–Quantil!) zu vergleichen, wobei zu beachten ist, dass die Freiheitsgrade, dem Zähler und Nenner der Testgröße entsprechend, gewählt werden.

Beispiel 5.7 In Fortsetzung von Beispiel 5.4 soll die Hypothese $\sigma_X = \sigma_Y$ überprüft werden (zweiseitige Fragestellung!). In Ausnützung der obigen Regel reiht man die beiden Stichproben zunächst so, dass die Testgröße größer als eins ausfällt. Damit erhält die Stichprobe für Y die Position 1 und die Testgröße ergibt sich zu

$$t = \frac{289.70}{133.21} = 2.175 \ .$$

Da dieser Wert unter

$$c_o = F_{n_Y-1,n_X-1;1-\alpha/2} = F_{7,9;0.975} = 4.197$$

liegt, wird die Nullhypothese beibehalten.

5.5 Analyse von Anteilen

Schätz– und Testverfahren für die Wahrscheinlichkeit $p = P(A)$ eines betrachteten Ereignisses A (z.B. "Antwort JA") beruhen in der Regel auf der *relativen Häufigkeit* $\bar{x}_A$ des Ereignisses A. Sie ergibt sich als (Stichproben–) Mittelwert einer Stichprobe $x_1, x_2, \ldots, x_n$ zu einer *alternativverteilten ZG* mit dem Parameter p.

Kann die Unabhängigkeit für die Stichprobenbeobachtungen vorausgesetzt werden, so ist die absolute Häufigkeit

$$y_A = \sum_{i=1}^{n} x_i = n\,\bar{x}_A$$

$Bi(n,p)$–binomialverteilt. Für exakte Konfidenzintervalle und Tests benötigt man also Quantile der Binomialverteilung. Unter Ausnützung des Zusammenhangs von Binomial– und F–Verteilung erhält man die Grenzen von Konfidenzintervallen auch über Quantile der F–Verteilung ($\rightarrow$ *Pearson–Clopper*–Grenzwerte).

Bei großen Stichprobenumfängen (Faustregel: $np(1-p) > 9$) nützt man die Approximationsmöglichkeit der Binomialverteilung durch eine Normalverteilung mit gleichem Mittelwert und gleicher Varianz aus (siehe Abschnitt 3.9). Dann gilt etwa

$$P\big(-z_{1-\alpha/2} \leq \frac{Y_A - np}{\sqrt{np(1-p)}} \leq z_{1-\alpha/2}\big) \approx 1-\alpha \ .$$

Die Varianz jedes x_i ist $p(1-p)$, damit gilt approximativ

$$p_u \leq p \leq p_o$$

mit

$$\left.\begin{matrix} p_o \\ p_u \end{matrix}\right\} = \bar{x}_A \pm z_{1-\alpha/2}\sqrt{\frac{\bar{x}_A(1-\bar{x}_A)}{n}}\ ,$$

abgeleitet aus Tab. 5.1, Konfidenzintervall für normalverteilte Größen mit bekannter Varianz. Die einzelnen x_i sind natürlich nicht normalverteilt, unser Schätzer $\bar{x}_A$ für den Anteil p ist aber asymptotisch normalverteilt.

Tabelle 5.12: *Konfidenzintervalle für p*

p_u	p_o
$\bar{x}_A - z_{1-\alpha/2}\sqrt{\frac{\bar{x}_A(1-\bar{x}_A)}{n}}$	$\bar{x}_A + z_{1-\alpha/2}\sqrt{\frac{\bar{x}_A(1-\bar{x}_A)}{n}}$
0	$\bar{x}_A + z_{1-\alpha}\sqrt{\frac{\bar{x}_A(1-\bar{x}_A)}{n}}$
$\bar{x}_A - z_{1-\alpha}\sqrt{\frac{\bar{x}_A(1-\bar{x}_A)}{n}}$	1

Beispiel 5.8 Anhand einer Stichprobe von 200 befragten Studenten soll der Anteil p_A derer geschätzt werden, die Nachhilfeunterricht geben. Von den 176, die antworten, geben 79 Nachhilfestunden. Wie lautet ein 95 %–Konfidenzintervall für p_A?

Lösung:
Nach Abzug der 24 Antwortverweigerer bleibt ein Stichprobenumfang von $n = 176$; damit ist

$$\hat{p}_A = \bar{x}_A = \frac{79}{176} = 0.449$$

ein Schätzwert für p_A .

Unter Ausnützung der Normalapproximation erhält man aus Tab. 5.12 als Konfidenzintervall

$$\left.\begin{matrix} p_o \\ p_u \end{matrix}\right\} = 0.449 \pm \underbrace{z_{0.975}}_{1.96}\sqrt{\frac{0.449 \times 0.551}{176}} = 0.449 \pm 0.073 = \left\{\begin{matrix} 0.522 \\ 0.376 \end{matrix}\right. .$$

Notwendiger Stichprobenumfang:

Bezeichnet $2d$ die gewünschte maximale Länge eines symmetrischen (1-α)–Konfidenzintervalles für p_A (also Intervall [p-d,p+d]), dann ist zumindest ein Stichprobenumfang

$$n_{\min} = \left(\frac{z_{1-\alpha/2}}{2d}\right)^2 \tag{5.14}$$

notwendig.

Beweis:
Unter Ausnützung der Normalapproximation können wir für d direkt das $1-\alpha/2$ Quantil der Normalverteilung als

$$d = z_{1-\alpha/2}\sqrt{\frac{\bar{x}_A(1-\bar{x}_A)}{n}}\ .$$

benutzen. Da für $0 < x < 1$ die Funktion $x(1-x)$, die eine konvexe Parabel darstellt, ihr Minimum bei 0.5 aufweist und dort den Wert 0.25 annimmt, muss also

$$d \geq z_{1-\alpha/2}\sqrt{\frac{0.25}{n}} = \frac{z_{1-\alpha/2}}{2\sqrt{n}}$$

gelten, woraus (5.14) folgt. △

Tabelle 5.13: *Notwendiger Stichprobenumfang*

d	$n_{\min}$ gemäß (5.14)		$n_{\min}$ bei $p_A \approx 0.1$	
	$\alpha = 0.01$	$\alpha = 0.05$	$\alpha = 0.01$	$\alpha = 0.05$
0.05	666	384	240	138
0.04	1040	600	374	216
0.03	1849	1067	666	384
0.02	4160	2401	1498	864
0.01	16641	9604	5991	3457

Die Tabelle 5.13 zeigt für $\alpha = 0.01$ und $\alpha = 0.05$ zu einigen Werten d die notwendigen Stichprobenumfänge gemäß (5.14), Normalverteilungsquantile auf 2 Nachkommastellen gerundet. Kennt man die Größenordnung von p_A, so kann man den notwendigen Stichprobenumfang zum Teil deutlich reduzieren, wenn man sich auf Konfidenzintervalle beschränkt, die *im Durchschnitt* kürzer als d sind:

$$n'_{\min} \approx z^2_{1-\alpha/2}\, p'_A(1-p'_A)/d^2$$

(p'_A steht für die Größenordnung von p_A). Diese Tatsache folgt auch unmittelbar aus dem obigen Beweis. Die reduzierten Umfänge finden sich ebenfalls in Tab. 5.13.

Ein exakter statistischer Test zur Hypothese $H_0 : p = p_0$ vergleicht die absolute Häufigkeit y_A des Ereignisses A mit kritischen Werten c_u und c_o, die quasi unteres und oberes $\alpha/2$–Quantil der Binomialverteilung darstellen:

$$c_u = b^{(u)}_{n,p_0;\alpha/2} = \max\left\{c : \sum_{i=0}^{c-1}\binom{n}{i}p_0^i(1-p_0)^{n-i} \leq \alpha/2\right\}$$

$$c_o = b^{(o)}_{n,p_0;\alpha/2} = \min\left\{c : \sum_{i=c+1}^{n}\binom{n}{i}p_0^i(1-p_0)^{n-i} \leq \alpha/2\right\}$$

Die Entscheidungsregel lautet dann:

$$\left.\begin{array}{l} c_u \leq y_A \leq c_o \\ y_A < c_u \text{ oder } y_A > c_o \end{array}\right\} \Rightarrow H_0 \left\{\begin{array}{l} \text{annehmen} \\ \text{ablehnen .} \end{array}\right.$$

Tabelle 5.14: *Tests für p_A*

H_0	exakter Test	Normalapproximation
	Teststatistik t	
	Annahmebereich (c_u, c_o)	
$p = p_0$	y_A	$\frac{\bar{x}_A - p_0}{\sqrt{p_0(1-p_0)/n}}$
	$c_o = b^{(o)}_{n,p;\alpha/2}$ $c_u = b^{(u)}_{n,p;\alpha/2}$	$\left.\begin{matrix} c_o \\ c_u \end{matrix}\right\} = \pm z_{1-\alpha/2}$
$p \leq p_0$	$c_o = b^{(o)}_{n,p;\alpha}$ $c_u = 0$	$c_o = z_{1-\alpha}$ $c_u = -\infty$
$p \geq p_0$	$c_o = n$ $c_u = b^{(u)}_{n,p;\alpha}$	$c_o = \infty$ $c_u = -z_{1-\alpha}$

$$b^{(u)}_{n,p;\gamma} = \max\{c : \sum_{i=0}^{c-1} \binom{n}{i} p^i (1-p)^{n-i} \leq \gamma\}$$
$$b^{(o)}_{n,p;\gamma} = \min\{c : \sum_{i=c+1}^{n} \binom{n}{i} p^i (1-p)^{n-i} \leq \gamma\}$$

Analog funktionieren die Tests bei einseitigen Hypothesen. In Tabelle 5.14 sind diese kritischen Werte zusammengestellt.

Kann man die Normalapproximation der Binomialverteilung (oder auch der relativen Häufigkeit) ausnützen, verwendet man die Testgröße

$$\frac{\bar{X}_A - p_0}{\sqrt{p_0(1-p_0)/n}} ,$$

die im Fall der vorhin behandelten Nullhypothese annähernd standardnormalverteilt ist. Damit ergeben sich für diese Testgröße die kritischen Werte

$$\left.\begin{matrix} c_o \\ c_u \end{matrix}\right\} = \pm z_{1-\alpha/2} .$$

Analog verläuft die Argumentation bei einseitigen Hypothesen. Auch diese kritischen Werte finden sich in Tab. 5.14.

Beispiel 5.9 In Fortsetzung von Beispiel 5.8 soll die Hypothese überprüft werden, ob mehr als die Hälfte der Studenten Nachhilfeunterricht gibt. Als Sicherheit wird $1-\alpha = 0.95$ gewählt.

Hier ist offensichtlich die Normalapproximation zu verwenden und man erhält aus Tab. 5.14 für die Testgröße

$$t = \frac{0.449 - 0.5}{\sqrt{0.5 \times 0.5/176}} = -1.353$$

als unteren kritischen Wert $c_u = -z_{0.95} = -1.645$. Die Testgröße liegt darüber, sodass die Hypothese *nicht verworfen* werden kann, es besteht also *kein signifikanter* Einwand gegen die eingangs aufgestellte Behauptung, dass mehr als die Hälfte der Studenten Nachhilfeunterricht erteilt.

Kapitel 6

Varianzanalyse

6.1 Einleitung

Die Varianzanalyse (**an**alysis **o**f **va**riance - ANOVA) stellt ein häufig verwendetes und effizientes Verfahren der angewandten Statistik zur Auswertung komplexer Versuche dar. Sie wurde von R.A. Fisher in den 1920er Jahren zur statistischen Auswertung von Feldversuchen entwickelt und seither laufend zu einer wirkungsvollen Methode zur Analyse ähnlicher und auch komplexerer Versuchsanordnungen verbessert und ausgebaut.

Das folgende Beispiel stellt eine von vielen verschiedenen Stichprobensituationen dar, die mit Modellen der Varianzanalyse behandelt werden können.

Beispiel 6.1 Vier Weizensorten werden hinsichtlich ihrer *ha*–Erträge (in *dt*) verglichen; bei verschiedenen Landwirten ergaben sich nachfolgende Werte, wobei jeder Landwirt bloß eine Sorte anbaut:

Sorte	Erträge						
1	41	47	50	43	46	48	51
2	42	39	34	40	44		
3	48	55	54	52	53	49	
4	44	49	41	45			

Liefern die Sorten durchschnittlich gleiche Erträge? Ist die Sorte 3 ertragreicher?

◇◇◇

Im obigen Beipiel steht die Frage im Vordergrund, ob die *vier* Weizensorten den gleichen durchschnittlichen Ernteertrag aufweisen. Es wird also der Einfluss des *Faktors* "Weizensorte" auf den Ernteertrag untersucht, wobei die (endlich vielen) *Stufen* des Faktors wählbar und somit *fest* vorgegeben sind. Sie weisen daher einen bestimmten, allerdings noch durch andere Unsicherheiten überlagerten Einfluss auf den Ernteertrag auf.

Allen Modellen gemeinsam ist das Prinzip, das zur Herleitung geeigneter Methoden für die Beantwortung aufgeworfener Fragen verwendet wird. In jedem Fall wird die "Gesamtvarianz"

$$\text{const} \times \sum_{y} (y - \bar{y})^2 \quad ,$$

in der y alle Beobachtungen durchläuft und $\bar{y}$ das (Gesamt-)Mittel darüber darstellt, in entsprechende Teile (Komponenten) aufgespaltet, die miteinander verglichen werden. Daraus leitet sich auch der Name dieser Verfahren ab.

6.2 Einfache Varianzanalyse

Hier ist der Einfluss *eines* Faktors A mit I Stufen auf die abhängige und beobachtbare Größe y von Interesse. Dazu werden pro Stufe J_i Versuche durchgeführt;

$$y_{ij} \qquad (i = 1, \ldots, I, \quad j = 1, \ldots, J_i)$$

bezeichne den beobachteten Wert von y im j-ten Versuch bei der Behandlung (Stufe) i.

Die Zufallsgröße y_{ij} wird dann üblicherweise als Summe eines für die Stufe i spezifischen Mittelwertes μ_i und eines zufälligen Fehlers e_{ij} interpretiert:

$$y_{ij} = \mu_i + e_{ij} \qquad (i = 1, \ldots, I, \quad j = 1, \ldots, J_i).$$

Zumeist interessieren aber die Abweichungen α_i von einem Gesamtmittel μ, die durch die Behandlung i entstehen, sodass üblicherweise die Beziehung

$$y_{ij} = \mu + \alpha_i + e_{ij} \qquad (i = 1, \ldots, I, \quad j = 1, \ldots, J_i) \tag{6.1}$$

gewählt wird. Da in diesem Fall für die $I + 1$ Parameter $\mu, \alpha_1, \ldots, \alpha_I$ nur I Beziehungen (nämlich die Stufen des Faktors A) vorhanden sind, wählt man als Nebenbedingung meist

$$\sum_{i=1}^{I} J_i \alpha_i = 0 \quad . \tag{6.2}$$

Die Fehler e_{ij} werden in der Standardanalyse unabhängig normalverteilt mit konstanter Varianz σ^2 angenommen (Homoskedastizität) . Damit lautet das Modell für die *einfache Varianzanalyse*

$$\begin{aligned} y_{ij} \quad &= \quad \mu + \alpha_i + e_{ij} \qquad (i = 1, \ldots, I, \quad j = 1, \ldots, J_i) \\ &\qquad e_{ij} \sim N(0, \sigma^2) \quad \text{unabhängig.} \end{aligned} \tag{6.3}$$

Für die Frage, ob der Faktor A einen Einfluss auf die abhängige Größe hat, testet man die Nullhypothese

$$H_A : \ \alpha_1 = \alpha_2 = \ldots = \alpha_I = 0 \tag{6.4}$$

(Gegenhypothese: mindestens ein Ungleichungszeichen). Zur Herleitung der Teststatistik versucht man, die Gesamtvariation der Beobachtungen aufzuspalten in einen Teil, der die Schwankung der Gruppen (als Gruppe werden alle Beobachtungen zu einer Stufe des Faktors A aufgefasst) um einen gemeinsamen Mittelwert beschreibt (Variation *zwischen* den Gruppen), und einen zweiten, der das Streuverhalten *innerhalb* der Gruppen erfasst. Wesentlich für die Untersuchung ist dann die Schwankung der Gruppenmittel *relativ* zum Streuverhalten innerhalb der Gruppen (die nur mehr die unkontrollierbare Zufälligkeit enthalten). Mit den Abkürzungen

$$\bar{y}_{i.} = \frac{1}{J_i} \sum_{j=1}^{J_i} y_{ij}$$

und

$$\bar{y}_{..} = \frac{1}{\sum_{i=1}^{I} J_i} \sum_{i=1}^{I} \sum_{j=1}^{J_i} y_{ij}$$

nützt man die Identität

$$(y_{ij} - \bar{y}_{..}) = \underbrace{(y_{ij} - \bar{y}_{i.})}_{\text{innerhalb}} + \underbrace{(\bar{y}_{i.} - \bar{y}_{..})}_{\text{zwischen}}$$

und erhält für die Gesamtvarianz

$$\begin{aligned} \sum_{i=1}^{I} \sum_{j=1}^{J_i} (y_{ij} - \bar{y}_{..})^2 &= \underbrace{\sum_{i=1}^{I} \sum_{j=1}^{J_i} (y_{ij} - \bar{y}_{i.})^2}_{SS_e} + \underbrace{\sum_{i=1}^{I} J_i (\bar{y}_{i.} - \bar{y}_{..})^2}_{SS_A} \\ &\quad + 2 \sum_{i=1}^{I} (\bar{y}_{i.} - \bar{y}_{..}) \underbrace{\sum_{j=1}^{J_i} (y_{ij} - \bar{y}_{i.})}_{0} \\ &= SS_e + SS_A \quad , \end{aligned} \tag{6.5}$$

also die oben erwähnte Aufspaltung in eine *Quadratsumme* (engl. *sum of squares, SS*) SS_A *zwischen* den Gruppen und eine, nämlich SS_e, *innerhalb* derselben. Bei *starken* Gruppeneinflüssen wird SS_A *größer* ausfallen als im Falle eines fehlenden Gruppeneinflusses, wogegen SS_e davon (theoretisch) unbeeinflusst bleibt. Daher wird der Einwand gegen die Nullhypothese H_A umso stärker sein, je größer SS_A ausfällt.

Für die exakte Formulierung der Teststatistik sind noch die statistischen Eigenschaften der Quadratsummen notwendig. Im Modell (6.3) gilt

$$\sum_{j=1}^{J_i} (y_{ij} - \bar{y}_{i.})^2 \sim \sigma^2 \chi^2_{J_i - 1} \quad ,$$

da die Fehler e_{ij} unabhängig normalverteilt sind. Aus dem Additionstheorem der χ^2–Verteilung folgt somit

$$SS_e = \sum_{i=1}^{I} \sum_{j=1}^{J_i} (y_{ij} - \bar{y}_{i.})^2 \sim \sigma^2 \chi^2_{\sum_{i=1}^{I} (J_i - 1)} \quad .$$

Als *mittlere* Quadratsumme (engl. *mean squares, MS*) wird der Quotient einer *SS* durch die Anzahl ihrer Freiheitsgrade (engl. *degrees of freedom, df*) bezeichnet. Damit erhält man mit

$$MS_e = SS_e / (\sum_{i=1}^{I} J_I - I)$$

einen erwartungstreuen Schätzer für σ^2, d.h. der Erwartungswert (engl. *expected mean squares, EMS*) ist $EMS_e = \sigma^2$. Aus diesem Grund wird SS_e oft auch Fehler–Quadratsumme (engl. *error sum of squares*) genannt.

Unter der Nullhypothese H_A gilt für die Verteilung von SS_A

$$SS_A = \sum_{i=1}^{I} J_i(\bar{y}_{i.} - \bar{y}_{..})^2 \sim \sigma^2 \chi^2_{I-1} \quad ,$$

also eine χ^2–Verteilung mit $I-1$ Freiheitsgraden, wobei SS_A und SS_e unabhängig sind (Satz von Cochran). Daher ist dann die Statistik (vgl. F–Verteilung)

$$F = \frac{MS_A}{MS_e} = \frac{SS_A/(I-1)}{SS_e/(\sum_{i=1}^{I} J_i - I)} \sim F_{I-1,\sum_{i=1}^{I} J_i - I}$$

F–verteilt. Wie oben angedeutet, sind große Werte für diese Statistik signifikant, sodass die Nullhypothese H_A dann zum Signifikanzniveau α zu verwerfen ist, falls

$$F = \frac{MS_A}{MS_e} > F_{I-1,\sum_{i=1}^{I} J_i - I;1-\alpha}$$

gilt. Kann H_A hingegen nicht verworfen werden, nimmt man an, dass die I Stufen des Faktors A keinen (nennenswerten) Einfluss auf das Mittel der beobachteten Variable y haben (Achtung vor einem Fehler 2. Art!)

Die im Zuge einer Varianzanalyse berechneten Zwischen- und Testgrößen werden üblicherweise in Tabellenform nach dem Schema in Tab. 6.1 angeordnet. Dabei enthält die

Tabelle 6.1: *Einfache Varianzanalyse*

Ursprung der Variabilität	SS	d.f.	MS	F	p
A	$\sum_{i=1}^{I} J_i(\bar{y}_{i.} - \bar{y}_{..})^2$	$I-1$	$\frac{SS_A}{I-1}$	$\frac{MS_A}{MS_e}$	p_A
Fehler	$\sum_{i=1}^{I}\sum_{j=1}^{J_i} (y_{ij} - \bar{y}_{i.})^2$	$\sum_{i=1}^{I} J_i - I$	$\frac{SS_e}{\sum_{i=1}^{I} J_i - I}$	–	–
Total	$\sum_{i=1}^{I}\sum_{j=1}^{J_i} (y_{ij} - \bar{y}_{..})^2$	$\sum_{i=1}^{I} J_i - 1$	–	–	–

Spalte $''F''$ den berechneten Wert der entsprechenden F–Statistik und die Spalte $''p''$ das empirische Signifikanzniveau, also die Wahrscheinlichkeit, dass unter der jeweiligen Nullhypothese die Teststatistik einen Wert größer oder gleich dem tatsächlich berechneten annimmt.

Um Schätzwerte für die in (6.3) verwendeten Parameter μ und α_i zu berechnen, wendet man die Methode der kleinsten Quadrate (engl. *least squares, LS*) an und minimiert die Summe der Residuenquadrate

$$S = \sum_{i=1}^{I}\sum_{j=1}^{J_i} (y_{ij} - \mu - \alpha_i)^2 \quad .$$

Für die partiellen Ableitungen nach den Parametern gilt

$$\begin{aligned} \frac{\partial S}{\partial \mu} &= (-2)\sum_{i=1}^{I}\sum_{j=1}^{J_i} (y_{ij} - \mu - \alpha_i) \\ \frac{\partial S}{\partial \alpha_i} &= (-2)\sum_{j=1}^{J_i} (y_{ij} - \mu - \alpha_i) \quad . \end{aligned}$$

Aus der Nebenbedingung $\sum_{i=1}^{I} J_i \alpha_i = 0$ erhält man sofort die *LS*–Schätzer

$$\hat{\mu} = \bar{y}_{..} \qquad \hat{\alpha}_i = \bar{y}_{i.} - \bar{y}_{..} \quad .$$

Beispiel 6.2 Mit dem Datenmaterial aus Bsp. 6.1 lässt sich eine einfache Varianzanalyse nach folgendem Schema durchführen:

Ursprung der Variabilität	SS	d.f.	MS	F	p
A	404.49	3	134.83	11.78	0.0002
Fehler	206.10	18	11.44		
Total	610.59	21			

Offensichtlich liegt wegen $11.78 > 3.16 = F_{3,18;0.95}$ ein *signifikanter* Einfluss der Weizensorte auf den Hektarertrag vor, was übrigens auch an dem extrem kleinen p–Wert abgelesen werden kann.

◇◇◇

6.3 Vollständige Versuchspläne

In Verallgemeinerung von Abschnitt 6.2 lässt sich auch die Abhängigkeit der beobachtbaren Zufallsgröße y von mehr als einem Einflussfaktor untersuchen. Stehen dazu Beobachtungen für *alle* Kombinationen der Stufen betrachteter Faktoren zur Verfügung, spricht man von *vollständigen Versuchsplänen*. Ist die Beobachtungsanzahl für alle Faktorkombinationen gleich, hat man es mit *balancierten*, ansonsten mit *unbalancierten* Versuchsplänen zu tun. Im folgenden werden vollständige Versuchspläne mit festen Effekten für zwei Faktoren (*zweifache Varianzanalyse*, engl. *two-way layout*) behandelt.

6.3.1 Zweifache Varianzanalyse ohne Wechselwirkungen

Zur Untersuchung des Einflusses zweier Faktoren A und B mit I und J Stufen liegen Beobachtungen y_{ijk} $(i = 1, \ldots, I;\ j = 1, \ldots, J;\ k = 1, \ldots, K)$ vor, wobei $K = 1$ sein kann, also keine wiederholten Beobachtungen pro Faktorkombination vorliegen. Allerdings wird die Beobachtungsanzahl K für alle Faktorkombinationen gleich angenommen. Analog zu (6.3) wählt man ein Modell

$$\begin{aligned} y_{ijk} \;=\; & \mu + \alpha_i + \beta_j + e_{ijk} \\ & \sum_i \alpha_i = \sum_j \beta_j = 0 \\ & e_{ijk} \sim N(0, \sigma^2) \quad \text{unabhängig} \\ & (i = 1, \ldots, I;\ j = 1, \ldots, J;\ k = 1, \ldots, K) \quad . \end{aligned} \tag{6.6}$$

Die in diesem Modell zu testenden Hypothesen lauten

$$H_A : \quad \alpha_1 = \ldots = \alpha_I = 0 \tag{6.7}$$

$$H_B : \quad \beta_1 = \ldots = \beta_J = 0 \quad . \tag{6.8}$$

Die *LS*–Schätzer für μ, α_i und β_j erhält man wie bei der einfachen Varianzanalyse durch Minimieren von

$$S = \sum_{ijk} (y_{ijk} - \mu - \alpha_i - \beta_j)^2 \quad ,$$

woraus sich durch Nullsetzen der partiellen Ableitungen nach den Parametern

$$\hat{\mu} = \bar{y}_{...} \qquad \hat{\alpha}_i = \bar{y}_{i..} - \bar{y}_{...} \qquad \hat{\beta}_j = \bar{y}_{.j.} - \bar{y}_{...}$$

ergibt. Offensichtlich gilt

$$y_{ijk} = \bar{y}_{...} + (\bar{y}_{i..} - \bar{y}_{...}) + (\bar{y}_{.j.} - \bar{y}_{...}) + (y_{ijk} - \bar{y}_{i..} - \bar{y}_{.j.} + \bar{y}_{...}) \quad ,$$

woraus nach einfacher Rechnung analog zu Abschnitt 6.2 die Quadratsummenzerlegung

$$\begin{aligned} & \sum_{ijk} (y_{ijk} - \bar{y}_{...})^2 \\ = \; & \sum_{ijk} (\bar{y}_{i..} - \bar{y}_{...})^2 + \sum_{ijk} (\bar{y}_{.j.} - \bar{y}_{...})^2 + \sum_{ijk} (y_{ijk} - \bar{y}_{i..} - \bar{y}_{.j.} + \bar{y}_{...})^2 \\ = \; & SS_A + SS_B + SS_e \end{aligned} \tag{6.9}$$

folgt. Für die Verteilung dieser Anteile gilt zunächst analog dem Abschnitt 6.2

$$SS_e \sim \sigma^2 \chi^2_{IJK-I-J+1} \; .$$

Falls die Hypothese H_A zutrifft, gilt

$$SS_A \sim \sigma^2 \chi^2_{I-1}$$

und SS_e und SS_A sind *unabhängig*. Entsprechend gilt bei zutreffender Hypothese H_B

$$SS_B \sim \sigma^2 \chi^2_{J-1}$$

und die Unabhängigkeit von SS_e und SS_B.

Zum Testen der Hypothesen H_A und H_B verwendet man wieder die F–Statistiken

$$\begin{aligned} F_A = \frac{MS_A}{MS_e} \quad & \sim \quad F_{I-1,IJK-I-J+1} \\ F_B = \frac{MS_B}{MS_e} \quad & \sim \quad F_{J-1,IJK-I-J+1} \quad , \end{aligned} \tag{6.10}$$

wobei die Verteilungen nur unter den Hypothesen H_A bzw. H_B gültig sind. Die Hypothese H_A erscheint dann nicht plausibel, wenn die (mittlere) Summe der Abweichungsquadrate SS_A (MS_A) der Stufenmittel des Faktors A vom Gesamtmittel im Vergleich zur Fehlerquadratsumme groß ausfällt, wenn also die F–Statistik F_A einen großen Wert annimmt. Daher ist die Hypothese H_A auf dem Signifikanzniveau α zu verwerfen, wenn

$$F_A = \frac{MS_A}{MS_e} > F_{I-1,IJK-I-J+1;1-\alpha} \tag{6.11}$$

ausfällt. Analoges gilt für die Hypothese H_B. Das Schema für die zweifache Varianzanalyse lautet daher wie in Tab. 6.2 angegeben. Die Bedeutung der Spalten entspricht der in Abschnitt 6.2 .

Tabelle 6.2: *Zweifache Varianzanalyse ohne Wechselwirkungen*

Ursprung der Variabilität	SS	d.f.	MS	F	p
A	$\sum_i JK(\bar{y}_{i..} - \bar{y}_{...})^2$	$I-1$	$\frac{SS_A}{I-1}$	$\frac{MS_A}{MS_e}$	p_A
B	$\sum_j IK(\bar{y}_{.j.} - \bar{y}_{...})^2$	$J-1$	$\frac{SS_B}{J-1}$	$\frac{MS_B}{MS_e}$	p_B
Fehler	$\sum_{ijk}(y_{ijk} - \bar{y}_{i..} - \bar{y}_{.j.} + \bar{y}_{...})^2$	$IJK - I - J + 1$	$\frac{SS_e}{IJK-I-J+1}$	–	–
Total	$\sum_{ijk}(y_{ijk} - \bar{y}_{...})^2$	$IJK-1$	–	–	–

6.3.2 Zweifache Varianzanalyse mit Wechselwirkungen

Neben dem rein additiven Ansatz zuvor, wo jede Stufe der Faktoren A und B den Mittelwert μ der beobachtbaren Zufallsgröße y um einen bestimmten, konstanten Wert α_i oder β_j verändert, besteht die Möglichkeit, auch den Einfluss sogenannter *Wechselwirkungen* (engl. *interactions*) zu betrachten. Dazu definiert man die Parameter

$$(\alpha\beta)_{ij} \qquad (i = 1, \ldots, I;\ j = 1, \ldots, J) \quad ,$$

die den zusätzlichen, durch die Summe der Einzeleinflüsse nicht beschreibbaren Effekt

$$\mathrm{E}(y_{ijk}) - \mu - \alpha_i - \beta j$$

der Behandlung (i, j) auf y ausdrücken sollen, und betrachtet in Ergänzung von (6.6) nun das Modell

$$\begin{aligned} y_{ijk} \quad = \quad & \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + e_{ijk} \\ & (i = 1, \ldots, I;\ j = 1, \ldots, J;\ k = 1, \ldots, K) \end{aligned} \tag{6.12}$$

mit den zusätzlichen Nebenbedingungen

$$\begin{aligned} \sum_{i=1}^{I} (\alpha\beta)_{ij} &= 0 \qquad & j = 1, \ldots, J \\ \sum_{j=1}^{J} (\alpha\beta)_{ij} &= 0 \qquad & i = 1, \ldots, I \quad . \end{aligned}$$

Als weitere Hypothese bietet sich nun

$$H_{AB}:\ (\alpha\beta)_{11} = \ldots = (\alpha\beta)_{IJ} = 0 \tag{6.13}$$

an. Zu beachten ist, dass dieses Modell nur im Fall mehrerer Beobachtungen je Zelle (d.h. $K > 1$) analysiert werden kann. Als Bezeichnung der Wechselwirkung zweier Faktoren A und B und zur Typisierung derartiger Modelle verwendet man häufig A∗B, A:B oder kürzer AB.

Die *LS*–Schätzer für μ, α_i, β_j und $(\alpha\beta)_{ij}$ erhält man wieder durch Minimieren von

$$S = \sum_{i=1}^{I} \sum_{j=1}^{J} \sum_{k=1}^{K} [(y_{ijk} - (\mu + \alpha_i + \beta_j + (\alpha\beta)_{ij})]^2 \tag{6.14}$$

als

$$\begin{aligned}\hat{\mu} &= \bar{y}_{...} \\ \hat{\alpha}_i = \bar{y}_{i..} - \bar{y}_{...} \quad \hat{\beta}_j &= \bar{y}_{.j.} - \bar{y}_{...} \\ \widehat{(\alpha\beta)}_{ij} &= \bar{y}_{ij.} - \bar{y}_{i..} - \bar{y}_{.j.} + \bar{y}_{...} \quad .\end{aligned}$$

Wie man leicht nachrechnet, beschreiben die *LS*–Schätzer für $(\alpha\beta)_{ij}$ tatsächlich genau die Differenz aus dem empirischen Zellenmittel $\bar{y}_{ij.}$ und den Schätzwerten aus dem rein additiven Modell (6.6):

$$\widehat{(\alpha\beta)}_{ij} = \bar{y}_{ij.} - (\hat{\mu} + \hat{\alpha}_i + \hat{\beta}_j) \quad .$$

Eine Quadratsummenzerlegung nach orthogonalen Komponenten erhält man wegen

$$y_{ijk} = \bar{y}_{...} + (\bar{y}_{i..} - \bar{y}_{...}) + (\bar{y}_{.j.} - \bar{y}_{...}) + (\bar{y}_{ij.} - \bar{y}_{i..} - \bar{y}_{.j.} + \bar{y}_{...}) + (y_{ijk} - \bar{y}_{ij.})$$

als

$$\begin{aligned}&\sum_{ijk}(y_{ijk} - \bar{y}_{...})^2 \\ &= \sum_{ijk}(\bar{y}_{i..} - \bar{y}_{...})^2 + \sum_{ijk}(\bar{y}_{.j.} - \bar{y}_{...})^2 + \\ &\quad \sum_{ijk}(\bar{y}_{ij.} - \bar{y}_{i..} - \bar{y}_{.j.} + \bar{y}_{...})^2 + \sum_{ijk}(y_{ijk} - \bar{y}_{ij.})^2 \\ &= SS_A + SS_B + SS_{AB} + SS_e \quad .\end{aligned} \tag{6.15}$$

Für die Verteilung dieser Anteile gilt nun analog dem Abschnitt 6.2

$$SS_e \sim \sigma^2 \chi^2_{IJ(K-1)} \ .$$

Falls die Hypothesen H_A bzw. H_B zutreffen, gelten dieselben Aussagen wie bei der zweifachen Varianzanalyse ohne Wechselwirkungen. Bei zutreffender Hypothese H_{AB} gilt

$$SS_{AB} \sim \sigma^2 \chi^2_{(I-1)(J-1)}$$

und wieder die Unabhängigkeit von SS_e und SS_{AB}.

Zum Testen der Hypothesen H_A, H_B und H_{AB} verwendet man auch hier wieder die F–Statistiken

$$\begin{aligned}F_A &= \frac{MS_A}{MS_e} \sim F_{I-1,IJ(K-1)} \\ F_B &= \frac{MS_B}{MS_e} \sim F_{J-1,IJ(K-1)} \\ F_{AB} &= \frac{MS_{AB}}{MS_e} \sim F_{(I-1)(J-1),IJ(K-1)} \quad ,\end{aligned} \tag{6.16}$$

wobei die Verteilungen nur unter den Hypothesen H_A, H_B bzw. H_{AB} gültig sind. Wiederum stellen große Werte für die Teststatistiken einen signifikanten Einwand gegen die jeweilige Hypothese an. Das Schema für die zweifache Varianzanalyse mit Wechselwirkungen zeigt Tab. 6.3 .

Tabelle 6.3: *Zweifache Varianzanalyse mit Wechselwirkungen*

Ursprung der Variabilität	SS	d.f.	MS	F	p
A	$\sum_i JK(\bar{y}_{i..} - \bar{y}_{...})^2$	$I-1$	$\frac{SS_A}{I-1}$	$\frac{MS_A}{MS_e}$	p_A
B	$\sum_j IK(\bar{y}_{.j.} - \bar{y}_{...})^2$	$J-1$	$\frac{SS_B}{J-1}$	$\frac{MS_B}{MS_e}$	p_B
AB	$\sum_{ij} K(\bar{y}_{ij.} - \bar{y}_{i..} - \bar{y}_{.j.} + \bar{y}_{...})^2$	$(I-1)(J-1)$	$\frac{SS_{AB}}{(I-1)(J-1)}$	$\frac{MS_{AB}}{MS_e}$	p_{AB}
Fehler	$\sum_{ijk}(y_{ijk} - \bar{y}_{ij.})^2$	$IJ(K-1)$	$\frac{SS_e}{IJ(K-1)}$	–	–
Total	$\sum_{ijk}(y_{ijk} - \bar{y}_{...})^2$	$IJK-1$	–	–	–

6.4 Varianztests

Offen blieb bisher die Frage, wie die immer wieder genannte Forderung *gleicher Varianzen* (Homoskedastizität) für bestimmte oder auch alle Gruppen von Beobachtungen überprüft werden kann. Üblicherweise schaltet man zu diesem Zweck einen Vortest über die Gleichheit der (oder bestimmter) Varianzen vor die eigentliche Varianzanalyse. Zu beachten ist dabei jedoch, dass dieser Vortest dem üblichen Prinzip bei Signifikanztests zuwiderläuft, da in diesem Fall versucht wird, die *Nullhypothese* (= gleiche Varianzen) zu *bestätigen*. Dennoch ist dieser Weg bei gebotener Vorsicht besser als keine Überprüfung.

Ausgangspunkt für die folgenden Tests sind I Stichproben mit den Umfängen $n_1, \ldots, n_I$, also

$$\begin{array}{llll} y_{11}\,, & \ldots\ y_{1n_1} & \text{mit} & \mathrm{Var}(Y_{1j_1}) = \sigma_1^2 \quad j_1 = 1, \ldots, n_1 \\ y_{21}\,, & \ldots\ y_{2n_2} & \text{mit} & \mathrm{Var}(Y_{2j_2}) = \sigma_2^2 \quad j_2 = 1, \ldots, n_2 \\ & \ldots & & \ldots \qquad\qquad \ldots \\ y_{I1}\,, & \ldots\ y_{In_I} & \text{mit} & \mathrm{Var}(Y_{Ij_I}) = \sigma_I^2 \quad j_I = 1, \ldots, n_I \quad , \end{array}$$

wobei über die Verteilung der y_{ij_i} vorerst noch nichts vorausgesetzt wird. (Da im Folgenden klar ist, welchen Bereich j_i durchläuft, wird zu Gunsten der Lesbarkeit auf den zusätzlichen Index i von j_i verzichtet.) Die Nullhypothese H_0 lautet für alle folgenden Tests

$$H_0 \ : \quad \sigma_1^2 = \sigma_2^2 = \ldots = \sigma_I^2 \quad .$$

Bartlett-Test

Im Folgenden wird vorausgesetzt, dass die Stichproben normalverteilt sind; diese Annahme ist wesentlich und Abweichungen davon beeinflussen die Wirksamkeit des Tests sehr stark. Weiters darf der Stichprobenumfang nicht zu klein sein; als Faustregel gilt $n_i \geq 5$ für $i = 1, \ldots, I$.

Zunächst berechnet man die Stichproben– (= Gruppen–) Varianzen

$$s_i^2 = \frac{1}{n_i - 1} \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_{i.})^2$$

und daraus die "durchschnittliche" Stichprobenvarianz

$$s^2 = \frac{1}{\sum_{i=1}^{I}(n_i - 1)} \sum_{i=1}^{I} (n_i - 1)\, s_i^2$$

(das entspricht der MS_e der Varianzanalyse), sowie die Konstante

$$c = \frac{1}{3(I-1)} \left(\sum_{i=1}^{I} \frac{1}{n_i - 1} - \frac{1}{\sum_{i=1}^{I}(n_i - 1)} \right) + 1 \quad .$$

Die Testgröße

$$b = \frac{1}{c} \sum_{i=1}^{I} (n_i - 1) \ln \frac{s^2}{s_i^2} = \tag{6.17}$$

$$\frac{1}{c} \left[\left(\sum_{i=1}^{I} (n_i - 1) \right) \ln s^2 - \sum_{i=1}^{I} (n_i - 1) \ln s_i^2 \right] \tag{6.18}$$

ist annähernd χ^2–verteilt mit $I - 1$ Freiheitsgraden. Große Werte für b sprechen gegen die Gleichheit der Varianzen σ_i, daher wird die Nullhypothese gleicher Varianzen zum Signifikanzniveau α abgelehnt, wenn

$$b > \chi^2_{I-1;1-\alpha} \tag{6.19}$$

ausfällt.

Hartley-Test

Unter der Annahme normalverteilter Stichproben *gleichen* Umfanges liegt diesem Test die einsichtige Testgröße

$$h = \frac{\max s_i^2}{\min s_i^2} \quad \text{zugrunde,} \tag{6.20}$$

wobei große Werte für h signifikant gegen die Nullhypothese sprechen. Kritische Werte für diesen Test finden sich in speziellen Tabellen.

Cochran-Test

Unter denselben Voraussetzungen wie der Hartley–Test verwendet der Test von Cochran die Testgröße

$$c = \frac{\max s_i^2}{\sum_{i=1}^{I} s_i^2} \quad , \tag{6.21}$$

wobei auch hier die großen Werte von c signifikant sind. Die kritischen Werte müssen ebenfalls aus speziellen Tabellen entnommen werden.

Für den Fall, dass die Stichprobenumfänge nicht zu stark differieren, wird in der Literatur die Verwendung des Cochran–Tests dennoch als zulässig betrachtet, wenn man als "gemeinsamen" Stichprobenumfang das harmonische Mittel

$$\tilde{n} := \left(\sum_{i=1}^{I} \frac{1}{n_i} \right)^{-1}$$

wählt und damit die Tabellen benützt.

Levene-Test

Die Annahme normalverteilter Stichproben wird hier fallen gelassen, ebenso werden auch keine gleichen Stichprobenumfänge verlangt. Ein großer Vorteil liegt in der Robustheit dieses Verfahrens. Hierzu benötigt man zunächst als Hilfsgrößen $z_{ij} = |y_{ij} - \bar{y}_{i.}|$ (bei einer Modifikation des Levene-Tests verwendet man $z^*_{ij} = |y_{ij} - \tilde{y}_i|$, nimmt also den Stichprobenmedian anstelle des Stichprobenmittelwertes). Damit führt man eine Art einfache "Varianzanalyse" für diese Abweichungen mit der naheliegenden Testgröße

$$w = \frac{\frac{1}{I-1}\sum_{i=1}^{I} n_i(\bar{z}_{i.} - \bar{z}_{..})^2}{\frac{1}{n-I}\sum_{i=1}^{I}\sum_{j=1}^{n_i}(z_{ij} - \bar{z}_{i.})^2} \tag{6.22}$$

durch, wobei $n = n_1 + \cdots + n_I$ den Gesamtumfang beschreibt. Große Werte für w sprechen gegen die Nullhypothese (= Varianzhomogenität), und man verwirft diese, wenn

$$w > F_{I-1,n-I;1-\alpha} \tag{6.23}$$

gilt.

Kapitel 7

Regressions– und Korrelationsanalyse

In diesem Kapitel werden Beschreibungs– und Analysemöglichkeiten für den Zusammenhang kontinuierlicher Größen diskutiert. Grundlage dafür ist eine *verbundene* Stichprobe $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$ zweier *ZGen* X und Y.

7.1 Korrelationsanalyse

Der (lineare) Zusammenhang zweier *ZGen* X und Y lässt sich durch die Kovarianz $\sigma_{X,Y}$ und den Korrelationskoeffizienten $\rho_{X,Y}$ (siehe Abschnitt 3.8) beschreiben. Als unverzerrter Schätzer für die Kovarianz dient die Stichprobenkovarianz

$$s_{xy} = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y}) \ , \tag{7.1}$$

deren Größe jedoch noch von der Varianz von X und Y abhängt. Dividiert man die Kovarianz durch die beiden Standardabweichungen s_x und s_y erhält man als dimensionslose Größe den (Pearson–) Stichprobenkorrelationskoeffizienten

$$\begin{aligned} r_{xy} &= \frac{s_{xy}}{s_x s_y} && (7.2) \\ &= \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}} && (7.3) \\ &= \frac{\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}}{\sqrt{\left(\sum_{i=1}^{n} x_i^2 - n\bar{x}^2\right)\left(\sum_{i=1}^{n} y_i^2 - n\bar{y}^2\right)}}. && (7.4) \end{aligned}$$

Die Werte des Korrelationskoeffizienten liegen im Wertebereich

$$-1 \leq r_{xy} \leq +1$$

Dabei bedeutet $r_{xy} = 0$ kein linearer Zusammenhang und $|r_{xy}| = 1$ vollkommener linearer Zusammenhang. Das Vorzeichen von r_{xy} gibt die Richtung des Zusammenhangs an, d.h. ob mit steigenden Werten von X auch die Werte von Y ansteigen (positive Korrelation) oder

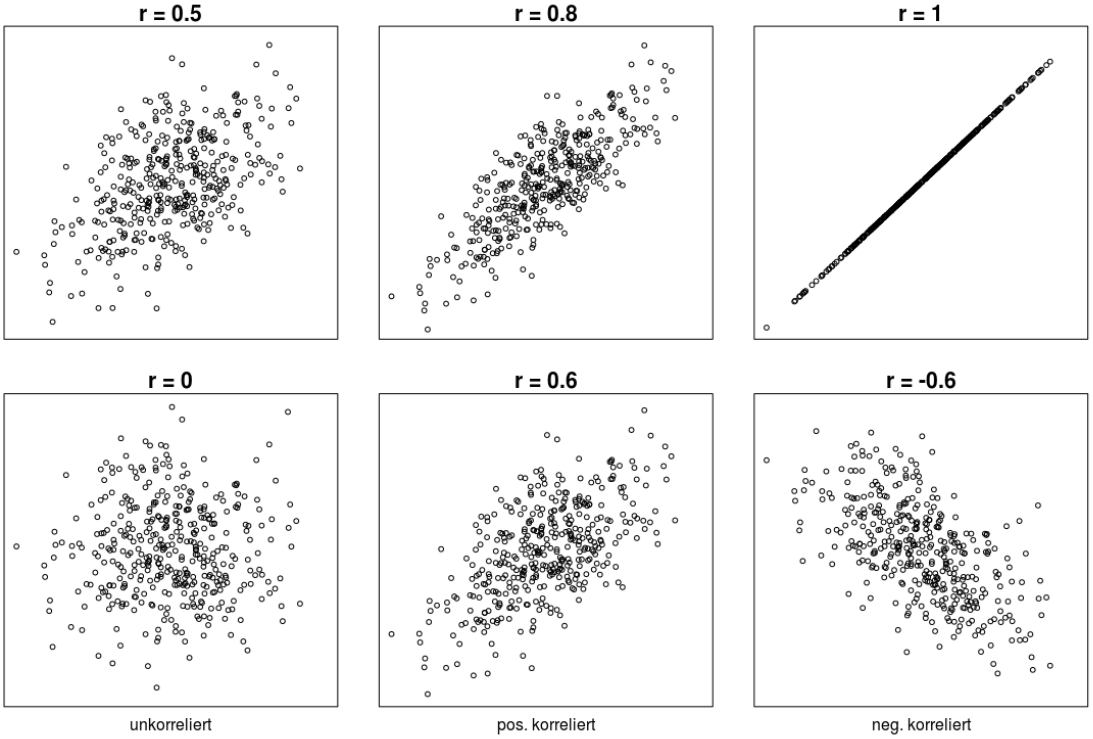


Abbildung 7.1: Interpretation der Korrelation anhand idealisierter Streudiagramme

hingegen fallen (negative Korrelation). Anhand von Streudiagrammen ist dies in Abb. 7.1 veranschaulicht.

Unter der Annahme, dass (X, Y) zweidimensional normalverteilt ist, lässt sich die Frage nach der Unabhängigkeit der beiden Größen X und Y direkt mit Hilfe der Korrelation untersuchen. In diesem Fall ist die Nullhypothese der Unabhängigkeit von X und Y gleichbedeutend mit der Unkorreliertheit ($\rho_{X,Y} = 0$), die Testgröße

$$t = \frac{r_{xy}\sqrt{n-2}}{\sqrt{1-r_{xy}^2}} \tag{7.5}$$

ist t_{n-2}–verteilt. Diese Testgröße wird absolut groß, wenn r_{xy} absolut groß ausfällt, was aber auf Grund der angenommenen Unabhängigkeit nicht zu erwarten ist. Daher stehen absolut große Werte für (7.5) im Widerspruch zur Nullhypothese und man wählt $\pm t_{n-2;1-\alpha/2}$ als kritische Werte für diesen Test:

$$|t| \left\{ \begin{array}{c} > \\ \leq \end{array} \right\} t_{n-2;1-\alpha/2} \;\Rightarrow\; H_0 \left\{ \begin{array}{l} \text{ablehnen} \\ \text{beibehalten.} \end{array} \right. \tag{7.6}$$

Das Streudiagramm kann zur Prüfung der Voraussetzungen des Korrelationskoeffizienten-Tests herangezogen werden. Dieser setzt eine Bi-Normalverteilung der Wertepaare (x_i, y_i) voraus, die einer elliptischen Form der Punktwolke entspricht (Abb. 7.2). Starke Abweichungen von der elliptischen Form deuten auf die Verletzung der Voraussetzung hin. Eine häufige

Ursache ist, dass nichtlineare Beziehungen vorliegen, wie nichlineare, monotone Beziehungen (b), oder nichtlineare, nichtmonotone Beziehungen (c). Auch Ausreißer können zu einer Verletzung der Voraussetzungen führen, und eine Erhöhung (d) oder eine Abminderung (e) des Korrelationskoeffizienten bewirken.

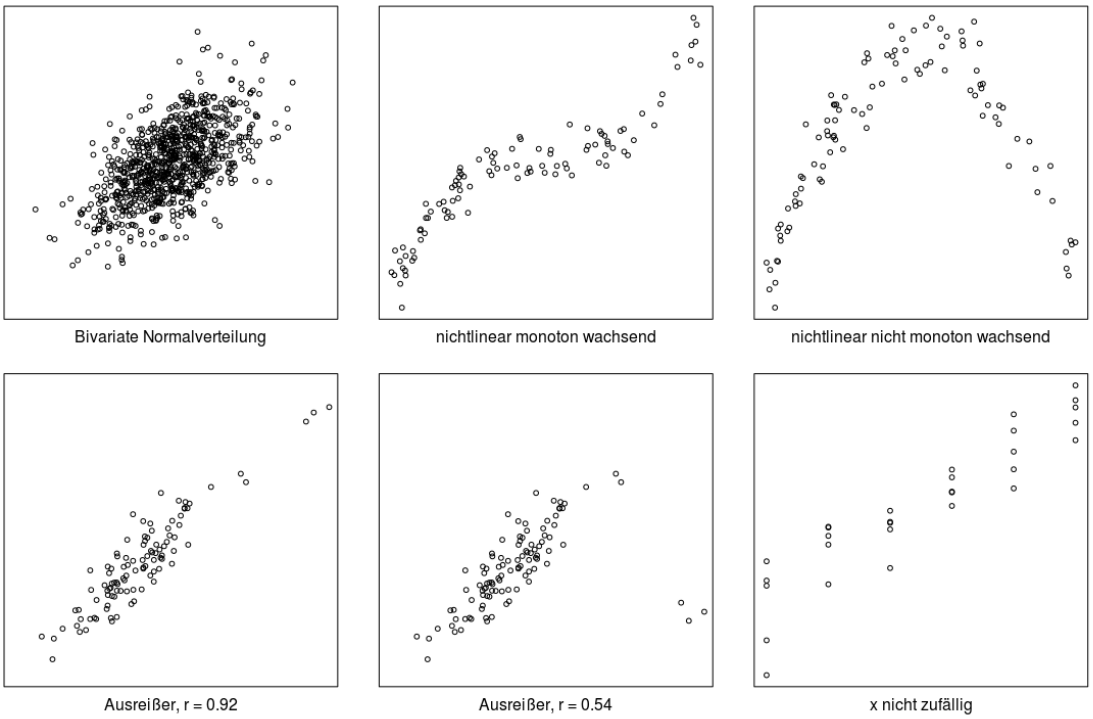


Abbildung 7.2: Prüfung der Voraussetzungen im Streudiagramm

Sind die Voraussetzungen nicht erfüllt, sollten alternative Korrelationsmaße wie z.B. der *Spearman-Korrelationskoeffizient* herangezogen werden. Beim Spearman-Korrelationskoeffizienten werden die x_i und y_i Werte jeweils der Größe nach geordnet und durch ihre Rangzahl ersetzt. Der Korrelationskoeffizient wird dann auf Basis der Rangzahlen berechnet, wodurch der Einfluss von Ausreißern vermindert wird. Der Spearman-Korrelationskoeffizient nimmt wiederum Werte im Intervall $[-1, +1]$ an und gibt die Stärke des monotonen Zusammenhangs zwischen zwei Variablen an.

Abbildung 7.2f zeigt schließlich eine Situation, bei der die Wertepaare (x_i, y_i) nicht zufällig, sondern an vorgegebenen Werten von X gemessen wurden. Dies kann zu einer deutlichen Abweichung von der Normalverteilung führen. Solche Fälle sind mittels Regressionsanalyse zu untersuchen, die im folgenden Abschnitt behandelt wird.

7.2 Einfache lineare Regressionsanalyse

Zwischen zwei kontinuierlichen Größen x und y vermutet man einen Zusammenhang, der im einfachsten Fall von der Art

$$y = a + bx$$

sein kann (Fahrzeuggeschwindigkeit/Reaktionsweg, Körpergewicht/Körpergröße). Da solche Größen in der Regel derartigen Gesetzen auf Grund von vielen nichtbeachteten bzw. unbekannten Einflüssen nicht exakt entsprechen, verwendet man häufig folgendes Modell. Unter der Voraussetzung, dass die x–Werte (fehlerfrei) gesteuert werden können, soll sich die zufallsbeeinflusste Größe Y als

$$Y|x = a + bx + E \tag{7.7}$$

mit dem *Fehlerterm* E darstellen lassen. Dabei wird vorausgesetzt, dass

R 1) der Fehler E $N(0, \sigma^2)$–normalverteilt und

R 2) σ offensichtlich von x unabhängig

ist. Der Zusammenhang kann somit durch die Mittelwertfunktion

$$\mathrm{E}(Y|x) = a + bx \ .$$

angegeben werden, die als Regressionsgerade bezeichnet wird. Die Regressionsgerade ist durch die beiden Parameter *Achsenabschnitt* a und die *Steigung* b bestimmt.[1] Wie in Abb. 7.3 ersichtlich entspricht a dem Wert der Regressionsgeraden bei $x = 0$, und b der Anstiegsrate der Regressionsgeraden, also dem Zugewinn von y wenn x um eine Einheit zunimmt.

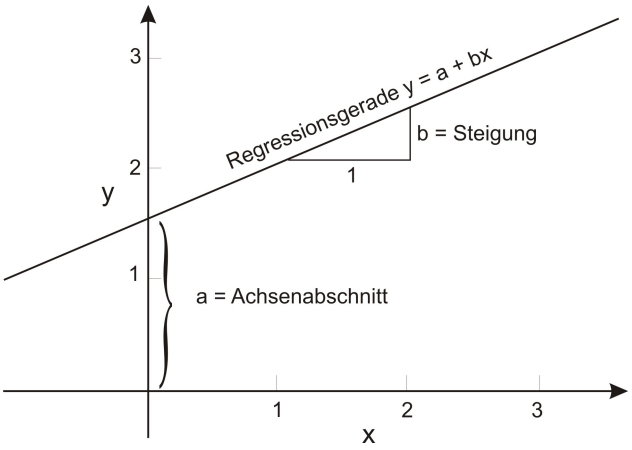


Abbildung 7.3: Definition der Regressionsparameter a und b

7.2.1 Schätzung der Regressionskoeffizienten

Die *Regressionskoeffizienten* a und b sowie die Varianz (auch *Fehlervarianz* genannt) σ^2 sind unbekannt und daher auf Grund einer Stichprobe $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$ zu schätzen, wobei üblicherweise die Unabhängigkeit der Stichprobenbeobachtungen vorausgesetzt wird (Modelle für abhängige Beobachtungen existieren, sind jedoch nicht Teil der Grundvorlesung).

Die beste Regression ist diejenige, die den geringsten Abstand zu allen Punkten hat. Hierzu betrachtet man die Abstände zwischen den *beobachteten* y–Werten und den mit Hilfe der geschätzten Regressionskoeffizienten *modellierten* $\hat{y}$–Werten

$$\hat{y}_i = \hat{a} + \hat{b}x_i$$

[1] Oft werden die Regressionsparameter a und b auch als β_0 und β_1 angeschrieben.

Diese Abstände werden als *Fehler* oder *Residuen* bezeichnet (siehe Abb. 7.4) und ergeben sich zu

$$(y_i - \hat{y}_i). \tag{7.8}$$

Da nur die y-Abstände betrachtet werden, ist die Regressionsbeziehung nicht symmetrisch. Eine Vertauschung von x- und y-Achse würde zu unterschiedlichen Regressionen führen. Aufgrund dieser Zielgerichtetheit führen die Variablen x und y bei der Regression einen besonderen Namen: x bezeichnet man als *Einflussvariable*, *unabhängige Variable* oder *Regressor*, y nennt man *abhängige Variable*.

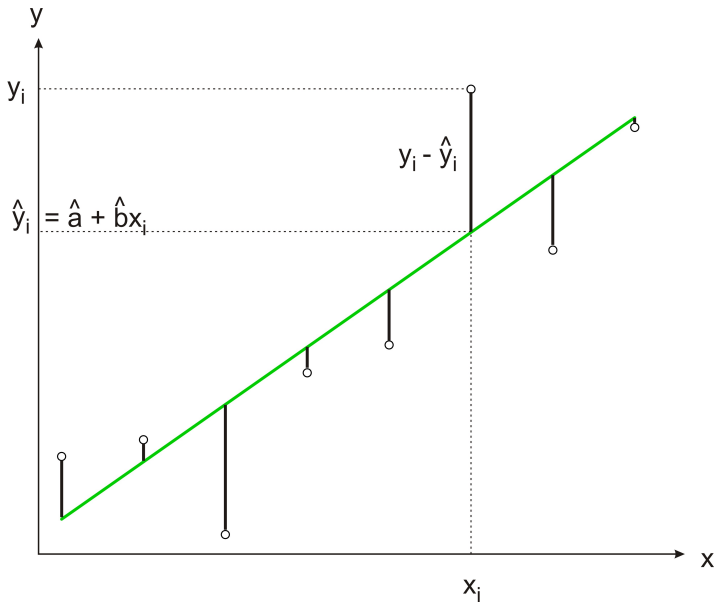


Abbildung 7.4: Residuen zwischen Beobachtung y und Modellwert $\hat{y}$

Die Schätzung der Regressionsparameter $\hat{a}$ und $\hat{b}$ erfolgt so, dass die Summe der quadrierten Residuen minimal ausfällt, also

$$S(\hat{a}, \hat{b}) = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} (y_i - \hat{a} - \hat{b}x_i)^2 \;\longrightarrow\; \min \tag{7.9}$$

gilt. Daher nennt man dieses Prinzip *Methode der kleinsten Fehlerquadrate (*engl. *least squares method)* oder kurz *LS*–Methode.

Zur Minimierung müssen die partiellen Ableitungen von S bezüglich a und b

$$\begin{aligned} \frac{\partial S(a,b)}{\partial a} &= \sum_{i=1}^{n} 2 \times (y_i - a - bx_i)(-1) \\ \frac{\partial S(a,b)}{\partial b} &= \sum_{i=1}^{n} 2 \times (y_i - a - bx_i)(-x_i) \end{aligned}$$

verschwinden, was zu dem Gleichungssystem *(Gauß'sche Normalgleichungen)*

$$\begin{array}{lcll} a\,n & + & b \sum_{i=1}^{n} x_i & = \sum_{i=1}^{n} y_i \\ a \sum_{i=1}^{n} x_i & + & b \sum_{i=1}^{n} x_i^2 & = \sum_{i=1}^{n} x_i y_i \end{array} \tag{7.10}$$

führt.

Die Auflösung des Gleichungssystems ergibt die Schätzungen der **Regressionsparameter**

$$\boxed{\hat{b} = \frac{s_{xy}}{s_x^2}} \tag{7.11}$$

und

$$\boxed{\hat{a} = \bar{y} - \hat{b}\bar{x}} \tag{7.12}$$

wobei in Gl. (7.11)

$$\begin{aligned} s_{xy} &= \tfrac{1}{n-1}\sum(x_i - \bar{x})(y_i - \bar{y}) \\ &= \tfrac{1}{n-1}\left(\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}\right) \end{aligned} \tag{7.13}$$

die Stichprobenkovarianz der x und y–Werte, und

$$\begin{aligned} s_x^2 &= \tfrac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 \\ &= \tfrac{1}{n-1}\left(\sum_{i=1}^{n} x_i^2 - n\bar{x}^2\right) \end{aligned} \tag{7.14}$$

die Stichprobenvarianz der x–Werte bezeichnet.

Der Schätzer $\hat{a}$ ist $N(a, \sigma_a^2)$-normalverteilt mit

$$\sigma_a^2 = \left(\frac{1}{n} + \frac{\bar{x}^2}{(n-1)s_x^2}\right)\sigma^2$$

Analog dazu besitzt $\hat{b}$ eine $N(b, \sigma_b^2)$-Normalverteilung mit

$$\sigma_b^2 = \frac{\sigma^2}{(n-1)s_x^2}\ .$$

Die Schätzungen $\hat{a}$ und $\hat{b}$ sind abhängig und besitzen den Korrelationskoeffizienten

$$\rho_{\hat{a},\hat{b}} = \frac{\bar{x}}{\sqrt{\sum_{i=1}^{n} x_i^2/n}}\ .$$

7.2.2 Fehlervarianz

Die Fehlervarianz σ^2 ist ein Maß für die durchschnittliche Abweichung der Messwerte von der Regressionsgeraden und errechnet sich als Durchschnitt der quadrierten Residuen $S(\hat{a}, \hat{b})$. Unter der Voraussetzung, dass die Residuen unabhängig, identisch verteilt und im Mittel Null sind, erhält man einen erwartungstreuen Schätzer für σ^2 durch Division der quadrierten Residuen $S(\hat{a}, \hat{b})$ durch die Anzahl ihrer Freiheitsgrade (FG), die sich aus der Anzahl der Beobachtungen minus der Anzahl der in der Regression geschätzten Parameter errechnet. Für die einfache lineare Regression gilt $FG = n - 2$.

Somit ist die Schätzung der Fehlervarianz σ^2

$$\boxed{\hat{\sigma}^2 = s^2 = \frac{S(\hat{a}, \hat{b})}{n-2} = \frac{1}{n-2}\sum_{i=1}^{n}(y_i - \hat{a} - \hat{b}x_i)^2} \tag{7.15}$$

wobei $(n-2)s^2/\sigma^2$ dann χ^2_{n-2}–verteilt ist. Eine algebraisch äquivalente, aber leichter auswertbare Formel für s^2 lautet

$$\boxed{s^2 = \frac{n-1}{n-2}\,(s_y^2 - \hat{b}^2 s_x^2)\ ,} \tag{7.16}$$

wobei s_y^2 die Gesamtvarianz

$$s_y^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2 \tag{7.17}$$

darstellt.

7.2.3 Konfidenzintervalle und Tests

Auf Grund obiger Verteilungsaussagen über die Schätzungen der Regressionskoeffizienten und der Fehlervarianz und unter Ausnützung der Definition der t–Verteilung gilt

$$\frac{\hat{a} - a}{s\sqrt{\frac{1}{n} + \frac{\bar{x}^2}{(n-1)s_x^2}}} \ \sim\ t_{n-2}$$

und

$$\frac{\hat{b} - b}{s/(\sqrt{n-1}\,s_x)} \ \sim\ t_{n-2}\ ,$$

was zur Herleitung von Tests und Konfidenzintervallen für a und b verwendet werden kann.

Ein beidseitiges $(1-\alpha)$ Konfidenzintervall von a ist gegeben durch

$$\hat{a} \ \pm\ t_{n-2;1-\alpha/2}\, s\sqrt{\frac{1}{n} + \frac{\bar{x}^2}{(n-1)s_x^2}}$$

Um zu Testen, ob a signifikant von 0 (oder einem beliebigen Wert a_0) verschieden ist, formuliert man die Hypothesen zu

$$H_0 : a = a_0,\ b\ \textit{beliebig}$$

$$H_A : a \neq a_0,\ b\ \textit{beliebig}$$

und berechnet die Teststatistik

$$\boxed{t = \frac{\hat{a} - a}{s\sqrt{\frac{1}{n} + \frac{\bar{x}^2}{(n-1)s_x^2}}}} \tag{7.18}$$

Absolut große Werte der Teststatistik $|t| \geq t_{n-2}$ führen zur Ablehnung der Nullhypothese.

In analoger Weise berechnet man ein beidseitiges $(1-\alpha)$ Konfidenzintervall für den Steigungsparameter b mit

$$\hat{b} \pm t_{n-2;1-\alpha/2}\frac{s}{s_x\sqrt{n-1}}$$

Um zu Testen, ob b signifikant von 0 verschieden ist, formuliert man die Hypothesen zu

$$H_0 : b = 0$$

$$H_A : b \neq 0$$

und berechnet die Teststatistik

$$\boxed{t = \frac{\hat{b} - b}{s/(\sqrt{n-1}\, s_x)}} \tag{7.19}$$

Absolut große Werte der Teststatistik $|t| \geq t_{n-2}$ führen zur Ablehnung der Nullhypothese. Andernfalls gibt es keine signifikante Abhängigkeit zwischen x und y.

7.2.4 Bestimmtheitsmaß

Eine Kenngröße zur Beurteilung der Brauchbarkeit des Modells (7.7) ist durch das *Bestimmtheitsmaß*

$$\boxed{R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} = 1 - \frac{(n-2)s^2}{(n-1)s_y^2}} \tag{7.20}$$

gegeben. Es liegt zwischen 0 und 1 und gibt den Anteil an der Gesamtvariabilität der y–Werte an, der durch das Regressionsmodell erklärt wird. Im Fall eines Regressionsmodells entspricht R^2 dem Quadrat des *formalen* Stichprobenkorrelationskoeffizienten der Größen x und y. Letzteres gilt aufgrund der Erwartungstreue $E(y) = E(\hat{y})$ des Regressionsmodells und ist daher nicht generell für beliebige, nicht erwartungstreue Modelle gültig.

Bei der Interpretation des Bestimmtheitsmaßes ist zu berücksichtigen, dass das R^2 eines linearen Regressionsmodells die Stärke des linearen Zusammenhanges der x_i und y_i–Werte angibt. Ein geringes R^2 bedeutet daher, dass kein linearer Zusammenhang zwischen den Größen x und y vorliegt. Ein nichtlinearer Zusammenhang wurde durch das Modell nicht untersucht und kann daher auch nicht ausgeschlossen werden.

Für große Stichproben kann das Bestimmtheitsmaß auch mittels

$$R^2 \approx 1 - \frac{s^2}{s_y^2} = \frac{s_y^2 - s^2}{s_y^2} \tag{7.21}$$

angenähert werden, da der Vorfaktor $(n-2)/(n-1)$ nahezu 1 wird, für große Werte von n. In dieser Form wird nochmal verdeutlicht, dass das Bestimmtheitsmaß den Anteil der durch das Modell erklärten Varianz angibt. Diese Formel ist aber nur zur Veranschaulichung gedacht; in Übungs- und Prüfungsaufgaben ist immer Formel 7.20 zu verwenden!

7.2.5 Konfidenz- und Prognoseband

Die Mittelwertfunktion des an die Daten angepassten Regressionsmodells kann zur Ermittlung des y-Wertes für einen gegebenen x-Wert verwendet werden. Hierbei werden die Fälle *Schätzung des Erwartungswertes* und *Prognose eines unbeobachteten Wertes* unterschieden.

In seltenen Fällen sind wir an der *Schätzung des Erwartungswertes* (i.e. Ausgleichswert der Regressionsgeraden) interessiert. Zur Schätzung "des" y–Wertes bei gegebenem x–Wert x_0 verwendet man in der Regel den Mittelwert $\mu_{y(x_0)} = \mathrm{E}(Y|x_0)$, den man unter Ausnützung der Schätzungen für a und b aus dem Modell (7.7) durch

$$\hat{y}_{x_0} = \hat{\mu}_{y(x_0)} = \hat{a} + \hat{b}x_0 \tag{7.22}$$

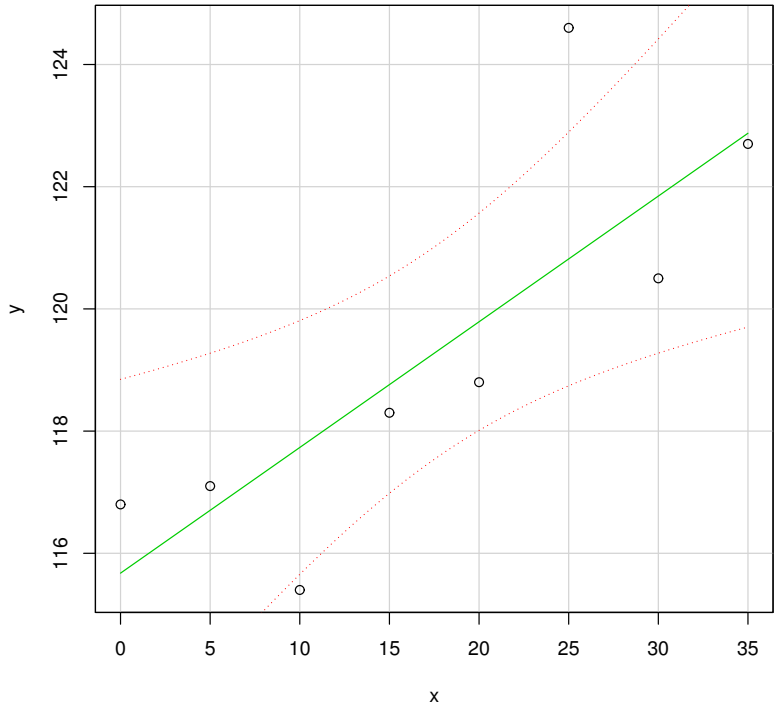


Abbildung 7.5: Konfidenzband der Regressionsgeraden

berechen kann. Diese Schätzung besitzt eine Normalverteilung mit dem Mittelwert $a + bx_0$ und der Varianz

$$\left(\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{(n-1)s_x^2}\right) \sigma^2 \ ,$$

sodass

$$\frac{\hat{y}_{x_0} - (\hat{a} + \hat{b}x_0)}{s\sqrt{\frac{1}{n} + \frac{(x_0-\bar{x})^2}{(n-1)s_x^2}}} \ \sim \ t_{n-2}$$

gilt, woraus sich wieder Tests und Konfidenzintervalle ableiten lassen. Damit kann man etwa ein $(1-\alpha)$–*Konfidenzintervall des Erwartungswertes* für einen y–Wert an der Stelle x_0 in der Form

$$\boxed{y \in \left(\hat{a} + \hat{b}x_0 \pm t_{n-2;1-\alpha/2} s \sqrt{\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{(n-1)s_x^2}}\right)} \tag{7.23}$$

angeben. Trägt man das Konfidenzintervall für jeden x-Wert auf, so erhält man das Konfidenzband der Regressionsgeraden (Abb. 7.5). Es repräsentiert die Unsicherheit der Regressionsbeziehung aufgrund der Stichprobenwerte.

In der Regel soll jedoch ein Konfidenzintervall für die *Prognose eines bisher unbeobachteten x–Wertes* x_0 angegeben werden. Dieser Wert kann, muss aber nicht in der Zukunft

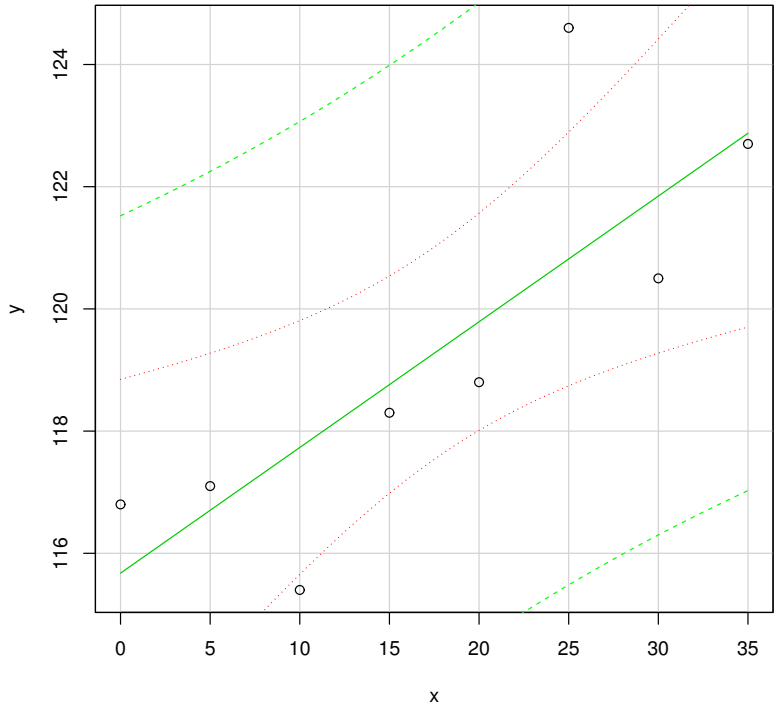


Abbildung 7.6: Prognoseband der Regressionsgeraden (äußere strichlierte Linien)

liegen. Wesentlich ist, dass der betrachtete x–Wert nicht zur Schätzung der Regressionsbeziehung verwendet wurde. Der beste Schätzwert für seinen y–Wert ist wieder der Mittelwert aus Gl. (7.22). Die Variabilität der Prädiktion hängt wiederum von der Unsicherheit der Regressionsgeraden (also von der Varianz der Schätzer $\hat{a}$ und $\hat{b}$) ab. Zusätzlich muss jedoch die Variabilität von Einzelwerten um den Mittelwert $a + bx_0$ berücksichtigt werden. Man erhält dann

$$\frac{y_{x_0} - (\hat{a} + \hat{b}x_0)}{s\sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{(n-1)s_x^2}}} \sim t_{n-2}$$

und kann damit etwa ein $(1 - \alpha)$–*Prognoseintervall* für einen y–Wert an der Stelle x_0 in der Form

$$\boxed{y \in \left(\hat{a} + \hat{b}x_0 \pm t_{n-2;1-\alpha/2}\, s \sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{(n-1)s_x^2}}\right)} \tag{7.24}$$

angeben. Trägt man das Prognoseintervall für jeden x-Wert auf, so erhält man das Prognoseband der Regressionsgeraden (Abb. 7.6). Es repräsentiert die Unsicherheit der Prognose bisher unbeobachteter Einzelwerte mittels Regressionsmodell.

7.2.6 Beispiele

Beispiel 7.1 Die Härte y von Plastikformteilen (in Brinell) hängt offensichtlich von der Aushärtezeit x (in h) ab. Zur Überprüfung des Zusammenhanges wurden für 12 Chargen mit teilweise verschiedenen Aushärtezeiten Proben untersucht (entnommen aus: Neter et al.: *Applied Linear Statistical Models*, Irwin–Verlag Homewood, 1985):

Aushärtezeit x (in h)	Härte y (in Brinell)	Aushärtezeit x (in h)	Härte y (in Brinell)
32	230	40	248
48	262	48	279
72	323	48	267
64	298	24	214
48	255	80	359
16	199	56	305

Es soll ein linearer Modellansatz untersucht werden.

a) Wie lauten Schätzwerte für a, b und σ^2?

b) Wie lauten 95%–Konfidenzintervalle für diese Parameter?

c) Mit welcher Brinellhärte ist bei einer Aushärtungszeit von 36 Stunden zu rechnen? Wie lautet ein 95%–Konfidenzintervall dafür?

d) Wie lautet ein 95%–Prognoseintervall für die Härte bei 60 Stunden Aushärtungszeit?

Lösung:

Für die Auswertung eines linearen Regressionsansatzes ergibt sich für den obigen Datensatz zunächst folgende Tabelle obigen Regressionsansatzes

i	x_i	y_i	x_i^2	y_i^2	$x_i y_i$
1	32	230	1 024	52 900	7 360
2	48	262	2 304	68 644	12 576
3	72	323	5 184	104 329	23 256
4	64	298	4 096	88 804	19 072
5	48	255	2 304	65 025	12 240
6	16	199	256	39 601	3 184
7	40	248	1 600	61 504	9 920
8	48	279	2 304	77 841	13 392
9	48	267	2 304	71 289	12 816
10	24	214	576	45 796	5 136
11	80	359	6 400	128 881	28 720
12	56	305	3 136	93 028	17 080
$\sum$	576	3 239	31 488	897 639	164 752

zu a)

Unter Verwendung der Formeln (7.11), (7.13), (7.15) und der Beziehung $n\bar{x} = \sum x_i$ erhält

man als Schätzwerte für die Regressionsparameter und die Fehlervarianz

$$\begin{aligned} s_x^2 &= \frac{1}{n-1}\left(\sum_{i=1}^{n} x_i^2 - n\bar{x}^2\right) \\ &= (31488 - 576^2/12)/11 = 349.0909 \\ s_y^2 &= \frac{1}{n-1}\left(\sum_{i=1}^{n} y_i^2 - n\bar{y}^2\right) \\ &= (897639 - 3239^2/12)/11 = 2125.356 \end{aligned}$$

$$\begin{aligned} s_{xy} &= \frac{1}{n-1}\left(\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}\right) \\ &= (164752 - 576 * 3239/12)/11 = 843.6364 \end{aligned}$$

$$\begin{aligned} \hat{b} &= s_{xy}/s_x^2 = 843.6364/349.0909 = 2.4167 \\ \hat{a} &= \bar{y} - \hat{b}\bar{x} = 269.9167 - 2.4167 \times 48 = 153.9151 \end{aligned}$$

$$\begin{aligned} s^2 &= \frac{n-1}{n-2}\,(s_y^2 - \hat{b}^2 s_x^2) = \frac{11}{10}\,(2125.356 - 2.4167^2 \times 349.09) = 95.169 \\ s &= 9.755 \end{aligned}$$

Für das Bestimmtheitsmaß erhält man

$$R^2 = 1 - \frac{(n-2)s^2}{(n-1)s_y^2} = 1 - \frac{10 \times 95.169}{11 \times 2125.356} = 0.9593,$$

was auf eine sehr gute lineare Modellierbarkeit hinweist.

zu b)

a: Mit $t_{10;0.975} = 2.228$ ergibt sich ein 95%–Konfidenzintervall für a als

$$\begin{aligned} \hat{a} &\pm t_{n-2;1-\alpha/2}\, s \sqrt{\frac{1}{n} + \frac{\bar{x}^2}{(n-1)s_x^2}} \\ &= 153.92 \pm 2.228 \times 9.755 \times \sqrt{\frac{1}{12} + \frac{(576/12)^2}{31488 - 576^2/12}} \\ &= 153.92 \pm 2.228 \times 8.04 \;=\; 153.92 \pm 17.91\,, \end{aligned}$$

also zu

$$(136.00, 171.84) \quad .$$

b: Ein 95%–Konfidenzintervall für b ergibt sich als

$$\begin{aligned} \hat{b} \pm t_{n-2;1-\alpha/2}\frac{s}{s_x\sqrt{n-1}} &= 2.417 \pm 2.228 \times 9.755/\sqrt{11 \times 349.09} \\ &= 2.417 \pm 0.350 \end{aligned}$$

und somit als

$$(2.067, 2.767) \quad .$$

σ^2: Als 95%–Konfidenzintervall für σ^2 erhält man mit $\chi^2_{10;0.025} = 3.247$ und $\chi^2_{10;0.975} = 20.483$

$$\frac{(n-2)s^2}{\chi^2_{n-2;1-\alpha/2}} \leq \sigma^2 \leq \frac{(n-2)s^2}{\chi^2_{n-2;\alpha/2}} \quad ,$$

also

$$(46.1927, 291.3967) \quad .$$

Daraus ergibt sich durch Wurzelziehen der Grenzen als 95%–Konfidenzintervall für σ

$$(6.797, 17.070) \quad .$$

zu c)
Durch Einsetzen in die Modellbeziehung ergibt sich

$$\hat{\mu}_{y(x=36)} = \hat{a} + \hat{b} \times 36 = 240.92$$

und als Konfidenzintervall für $\mu_{y(x=36)}$ erhält man

$$\begin{aligned}
&\hat{\mu}_{y(x=36)} \pm t_{n-2;1-\alpha/2}\, s \sqrt{\frac{1}{n} + \frac{(x-\bar{x})^2}{(n-1)s_x^2}} = \\
&240.92 \pm 2.228 \times 9.755 \sqrt{\frac{1}{12} + \frac{(36-48)^2}{11 \times 349.09}} = \\
&240.92 \pm 7.536 \quad ,
\end{aligned}$$

also

$$(233.39, 248.47) \quad .$$

zu d)
Man erhält ein Prognoseintervall für eine *Beobachtung* y bei 60 h (zur Sicherheit 1-α), indem man zunächst wie unter c) den Erwartungswert $\mu_{y(x=60)}$ schätzt und damit

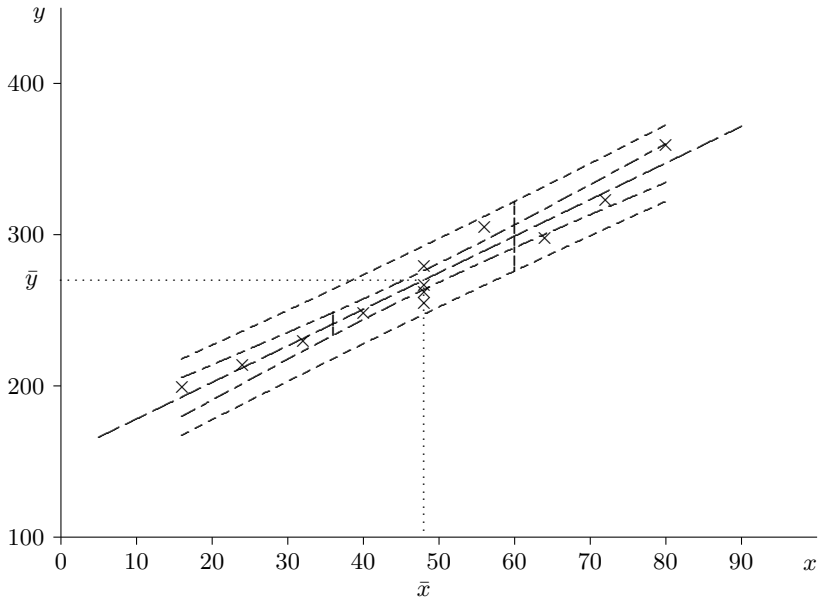
$$\begin{aligned}
&\hat{\mu}_{y(x=60)} \pm t_{n-2;1-\alpha/2}\, s \sqrt{1 + \frac{1}{n} + \frac{(x-\bar{x})^2}{(n-1)s_x^2}} = \\
&298.92 \pm 22.95 \quad ,
\end{aligned}$$

also

$$(275.99, 321.89)$$

bildet.

Die grafische Darstellung der Daten und ihrer Auswertung ergibt folgendes Bild:



7.2.7 Kalibration (Inverse Regression)

Wir haben eine lineare Beziehung zwischen der unabhängigen Größe ("Standard") x und dem Messwert y; aus Daten (x_i, y_i) mit $i = 1, \ldots, n$ werden die Regressionskoeffizienten a (Abschnittsparameter = konstanter Term) und b (Anstiegsparameter) geschätzt und wir erhalten damit die geschätzte Regressionsbeziehung

$$\hat{y} = \hat{a} + \hat{b}\, x \ .$$

Für den umgekehrten Vorgang (zu einem beobachteten y–Wert y_0 ist der zugehörige "wahre" x–Wert zu "bestimmen") gilt:

1) Die beste Schätzung ergibt sich einfach zu

$$\hat{x}_0 = \frac{y_0 - \hat{a}}{\hat{b}} \tag{7.25}$$

2) Ein Konfidenzintervall zur Irrtumswahrscheinlichkeit α (z.B. 5 %) erhält man formal einfach aus der entsprechenden Formel zum Konfidenzintervall für einen *beobachteten* Wert durch Umkehrung; das führt auf (sieht komplizierter aus, als es ist):

$$\left.\begin{matrix} x_0^{\text{oben}} \\ x_0^{\text{unten}} \end{matrix}\right\} = \hat{x}_0 + \frac{(\hat{x}_0 - \bar{x})\, g \pm (ts/\hat{b})\, \sqrt{(\hat{x}_0 - \bar{x})/((n-1)s_x^2) + (1-g)(1+1/n)}}{1-g} \tag{7.26}$$

wobei

$$g = \frac{(t\,\hat{b})^2}{(n-1)s_x^2/s^2} \qquad \text{und} \qquad t = t_{n-2;1-\alpha/2}$$

bedeuten. Der Nenner in g sollte in der Regel groß werden, weil man meist von einem gut passenden Modell mit kleinem σ und daher auch kleinem s ausgehen kann.

Beweis:

Ein $(1-\alpha)$–Konfidenzintervall für einen einzelnen Messwert ergibt sich mit der obigen Abkürzung zu

$$\hat{a}+\hat{b}x_0 - t\,s\sqrt{1+\frac{1}{n}+\frac{(x_0-\bar{x})^2}{(n-1)s_x^2}} \le y_0 \le \hat{a}+\hat{b}x_0 + t\,s\sqrt{1+\frac{1}{n}+\frac{(x_0-\bar{x})^2}{(n-1)s_x^2}}$$

bzw.

$$-t\,s\sqrt{1+\frac{1}{n}+\frac{(x_0-\bar{x})^2}{(n-1)s_x^2}} \le y_0 - \hat{a} - \hat{b}x_0 \le +t\,s\sqrt{1+\frac{1}{n}+\frac{(x_0-\bar{x})^2}{(n-1)s_x^2}}\ .$$

Das ist aber gleichbedeutend mit

$$(y_0-\hat{a}-\hat{b}x_0)^2 \le t^2\,s^2\left(1+\frac{1}{n}+\frac{(x_0-\bar{x})^2}{(n-1)s_x^2}\right)$$

oder wegen $\hat{x}_0 = (y_0-\hat{a})/\hat{b}$ eben

$$(\hat{x}_0-x_0)^2 \le \frac{t^2\,s^2}{\hat{b}^2}\left(1+\frac{1}{n}+\frac{(x_0-\bar{x})^2}{(n-1)s_x^2}\right)\ .$$

Das ist aber äquivalent zu (kleiner Trick!)

$$(\hat{x}_0-x_0)^2 \le \frac{t^2\,s^2}{\hat{b}^2}\left(1+\frac{1}{n}+\frac{[(\hat{x}_0-\bar{x})-(\hat{x}_0-x_0)]^2}{(n-1)s_x^2}\right)$$

und nach Quadrieren des Ausdrucks in eckigen Klammern, sowie Subtraktion des Terms mit $(\hat{x}_0-x_0)^2$

$$(\hat{x}_0-x_0)^2\left(1-\frac{t^2\,s^2}{(n-1)s_x^2\,\hat{b}^2}\right) \le \frac{t^2\,s^2}{\hat{b}^2}\left(1+\frac{1}{n}+\frac{(\hat{x}_0-\bar{x})^2}{(n-1)s_x^2}-\frac{2\,(\hat{x}_0-\bar{x})(\hat{x}_0-x_0)}{(n-1)s_x^2}\right)\ .$$

Mit der Abkürzung g von oben und nach Subtraktion der rechten Seite in der letzten Formel vereinfacht sich diese zu

$$(\hat{x}_0-x_0)^2\,(1-g)+2g\,(\hat{x}_0-\bar{x})(\hat{x}_0-x_0)-\frac{t^2\,s^2}{\hat{b}^2}\left(1+\frac{1}{n}\right)-g\,(\hat{x}_0-\bar{x})^2 \le 0\ .$$

Zwischen den Nullstellen der hier beschriebenen quadratischen Funktion in $\hat{x}_0-x_0$ ist die letzte Ungleichung erfüllt. Man erhält sie für

$$A\,u^2+B\,u+C$$

bekanntlich als

$$u_{1,2}=\frac{-B\pm\sqrt{B^2-4AC}}{2A}$$

und damit ergeben sie sich für $\hat{x}_0-x_0$ mit

$$\begin{aligned} A:&\quad 1-g\\ B:&\quad 2g\,(\hat{x}_0-\bar{x})\\ C:&\quad -\frac{t^2\,s^2}{\hat{b}^2}\left(1+\frac{1}{n}\right)-g\,(\hat{x}_0-\bar{x})^2 \end{aligned}$$

als

$$\hat{x}_0-x_0=\frac{-g\,(\hat{x}_0-\bar{x})\pm\sqrt{g^2\,(\hat{x}_0-\bar{x})^2+(1-g)\,(t^2\,s^2/\hat{b}^2)\,(1+1/n)+(1-g)g\,(\hat{x}_0-\bar{x})^2}}{1-g}$$

woraus unmittelbar die Grenzen des ursprünglich gesuchten Konfidenzintervalls für x_0 in der oben angegeben Form folgen. △

.

Kapitel 8

Nichtparametrische Verfahren

Manchmal ist der Verteilungstyp, der durch einen oder mehrere Parameter beschrieben werden kann, nicht angebbar oder gehört zu einer Klasse, die nicht einfach zu handhaben ist. In solchen Fällen lassen sich *nichtparametrische* Verfahren sehr vorteilhaft einsetzen. Sie sind zwar in der Regel im Vergleich zu parametrischen Methoden nicht so effizient, aber häufig einfacher und vor allem in mehr Situationen einsetzbar.

8.1 Rang–Tests

Bei den Verfahren in diesem Abschnitt handelt es sich um Beispiele sogenannter *Rang–Tests*. Bei diesen geht nicht der Beobachtungswert selbst in die Testgröße ein, sondern es wird nur Information über seine Rangposition in der/den Stichprobe(n) verwendet. Es wird dabei keine Verteilungsannahme getroffen, es müssen bloß kontinuierliche Merkmale vorausgesetzt werden. Klarerweise gibt es bei der Anwendung dieser Verfahren einen Informationsverlust und damit geringere Aussageschärfe. Dieser Nachteil wird in der Regel durch größere Robustheit aufgewogen.

8.1.1 Wilcoxon–Vorzeichenrangtest

Dieser Test eignet sich zur Überprüfung von Hypothesen über die Lage (symmetrischer) Merkmale und stellt damit eine Verallgemeinerung von Mittelwerttests bei normalverteilten Merkmalen (siehe Abschnitt 5.1) dar. Als einzige Voraussetzung ist ein *symmetrisches* kontinuierliches Merkmal verlangt. Es muss also ein Symmetriezentrum ζ geben, das dann natürlich auch gleichzeitig (theoretischer) Mittelwert und Median ist und für das $F_X(\zeta - y) = 1 - F_X(\zeta + y)$ (F_X steht für die theoretische *VF* des Merkmals X) gilt. Man kann diesen Test daher auch als einen Median–Test bezeichnen.

Zur Überprüfung der Nullhypothese H_0: $x_{0.5} = \zeta_0$ (für die anderen beiden Hypothesen in Tab. 8.1 gilt dies ebenfalls) bildet man zunächst die Hilfsgrößen

$$x_i' = x_i - \zeta_0$$

(falls dabei $x_i' = 0$ auftritt, wird diese Probe einfach aus der Stichprobe eliminiert!). Danach ordnet man die *Absolutbeträge* $|x_i'|$ dieser Werte der Größe nach und weist ihnen ihre Rangposition r_i zu, also dem kleinsten Wert unter $|x_i'|$ den Rang 1, dem größten dagegen den Rang n. Im Falle gleich großer Absolutwerte (Bindungen) wird einfach der Durchschnitt der für diese

Tabelle 8.1: *Wilcoxon–Vorzeichenrangtest*

H_0	Teststatistik t^+	
	Annahmebereich (c_u, c_o)	
	c_u	c_o
	$t^+ = \sum_{i=1}^{n} c_i r_i$	
$x_{0.5} = \zeta_0$	$w_{n;\alpha/2}$	$w_{n;1-\alpha/2}$
$x_{0.5} \geq \zeta_0$	$w_{n;\alpha}$	$n(n+1)/2$
$x_{0.5} \leq \zeta_0$	0	$w_{n;1-\alpha}$

Werte vorgesehenen Rangpositionen (*mittlerer Rang*, engl: *midrank*) zugewiesen. Schließlich benötigt man für die Testgröße noch zu jedem x'_i–Wert ein Gewicht

$$c_i = \begin{cases} 1 & \text{falls } x'_i > 0 \\ 0 & \text{falls } x'_i < 0 \end{cases}$$

und bildet nun die Summe der Rangzahlen zu positiven x'–Werten:

$$t^+ = \sum(\text{Rangzahlen der positive } x'_i) = \sum_{i=1}^{n} c_i r_i \ . \tag{8.1}$$

Im Falle $x_{0.5} = \zeta_0$ erwartet man für t^+ einen Wert um die halbe Rangsumme, also $n(n+1)/4$. Große Werte deuten auf zuviele oder zu große positive Abweichungen, also auf einen tatsächlichen Median über ζ_0 hin, kleine auf einen Median unter ζ_0 . Damit erhält man die in Tab. 8.1 angegebenen kritischen Werte, wobei die Werte selbst in Tab. 8.2 (Hartung, 2002) zusammengestellt sind.

Beispiel 8.1 In einer Stichprobe von Äpfeln aus einer Lieferung der Sorte Golddelicious wird das Apfelgewicht (in g) gemessen:

147 152 185 138 164 163 153 160

Anhand der verfügbaren Daten soll die Frage untersucht werden, ob der Median (ist dann gleich dem theoretischen Mittelwert) mit 150 g angenommen werden kann, wobei von einem symmetrisch verteilten Gewicht ausgegangen werden kann (Signifikanzniveau $\alpha = 5\,\%$).

Lösung:

Die Tabelle 8.3 zeigt die benötigten Hilfsgrößen (x'_i–, $|x'_i|$–, r_i– und c_i–Werte). Damit erhält man als Testgröße

$$t^+ = 0 \times 2.5 + 1 \times 1 + 1 \times 8 + 0 \times 5 + 1 \times 7 + 1 \times 6 + 1 \times 2.5 + 1 \times 4 = 28.5$$

Für die Hypothese $x_{0.5} = 150\,g$ erhält man aus den Tabellen 8.1 und 8.2 die kritischen Werte $w_{8;0.025} = 4$ und $w_{8;0.975} = 31$ zum vorgeschriebenen Signifikanzniveau 5 %. Die Testgröße liegt dazwischen, also kann man die Hypothese ($x_{0.5} = 150\,g$) nicht verwerfen.

Tabelle 8.2: *Wilcoxon–Vorzeichenrangtest; kritische Werte* $w_{n;\gamma}$

n	$w_{n;0.01}$	$w_{n;0.025}$	$w_{n;0.05}$	$w_{n;0.1}$	$w_{n;0.9}$	$w_{n;0.95}$	$w_{n;0.975}$	$w_{n;0.99}$
4	0	0	0	1	8	9	10	10
5	0	0	1	3	11	13	14	14
6	0	1	3	4	16	17	19	20
7	1	3	4	6	21	23	24	26
8	2	4	6	9	26	29	31	33
9	4	6	9	11	33	35	38	40
10	6	9	11	15	39	43	45	48
11	8	11	14	18	47	51	54	57
12	10	14	18	22	55	59	62	66
13	13	18	22	27	63	68	72	77
14	16	22	26	32	72	78	82	88
15	20	26	31	37	82	88	93	99
16	24	30	36	43	92	99	105	111
17	28	35	42	49	103	110	117	124
18	33	41	48	56	114	122	129	137
19	38	47	54	63	126	135	142	151
20	44	53	61	70	139	148	156	165

Tabelle 8.3: *Wilcoxon–Vorzeichenrangtest*

i	x_i	x'_i	$\lvert x'_i\rvert$	r_i	c_i	i	x_i	x'_i	$\lvert x'_i\rvert$	r_i	c_i
1	147	-3	3	2.5	0	5	164	14	14	7	1
2	152	2	2	1	1	6	163	13	13	6	1
3	185	35	35	8	1	7	153	3	3	2.5	1
4	138	-12	12	5	0	8	160	10	10	4	1

8.1.2 Wilcoxon–Rangsummentest

Zwei unabhängige Stichproben $x_{1,1}, x_{1,2}, \ldots, x_{1,n_1}$ und $x_{2,1}, x_{2,2}, \ldots, x_{2,n_2}$ für die *ZGen* X_1 und X_2 bilden den Ausgangspunkt der Analyse. Die zugrundeliegenden Verteilungen von X_1 und X_2 werden durch die *VFen* F_1 und F_2 beschrieben. Die Nullhypothese lautet nun:

$$H_0 \; : \; F_1(z) = F_2(z) \; . \tag{8.2}$$

Die Alternative ist daß sich die beiden Verteilungen in der Lokation (Median, Mittelwert) unterscheiden, die äußere Form der beiden Verteilungen jedoch gleich ist:

$$H_A \; : \; F_1(z) = F_2(z - c) \; . \tag{8.3}$$

Dabei ist die konstante reelle Zahl c die Verschiebung in der Lokation. Welche Form die Verteilungen F_1 und F_2 haben (symmetrisch, schief) ist egal, solange es dieselbe (oder zumindest ähnliche) Form ist. Insbesonders müssen natürlich auch die Varianzen in beiden Gruppen

gleich sein. Zur Überprüfung der Annahme gleicher Verteilungsform bieten sich z.B. Boxplots für beide Stichproben an. Ist die Form der Boxen und Whiskers zumindest ähnlich und nur parallel verschoben, kann der Wilcoxon–Rangsummentest meistens gut verwendet werden.

Für den Test werden die beiden Stichproben zu einer gemeinsamen Stichprobe zusammengefasst, die anschließend der Größe nach geordnet wird:

$$y_{(1)} \leq y_{(2)} \leq \cdots \leq y_{(n_1+n_2)} \ .$$

Die geklammerten Indizes stellen die Position in dieser Rangordnung dar und werden demnach als *Rangzahlen* bezeichnet.

Kommt ein Wert in der gemeinsamen Probe öfter vor *(Bindung,* engl. *tie)*, wird der *Durchschnittswert* aller Rangzahlen, die für diesen Wert vorgesehen wären, jedem Auftreten dieses Wertes als Rangzahl zugeordnet (siehe auch das folgende Beispiel).

Man betrachtet nun die Rangzahlen r_i der (geordneten) Stichprobenwerte $x_{1,(i)}$ $(i = 1, \ldots, n_1)$ der ersten Probe, für die somit $x_{1,(i)} = y_{(r_i)}$ gilt. Falls die Nullhypothese zutrifft, müssten diese Rangzahlen für die erste Probe mehr oder weniger gleichmäßig verteilt im Bereich $1, \ldots, n_1 + n_2$ liegen. Eine brauchbare Testgröße zur Beurteilung dieser Forderung ist die Summe

$$w_{n_1,n_2} = \sum_{i=1}^{n_1} r_i \tag{8.4}$$

der Rangzahlen der ersten Stichprobe. Für diese *Rangsumme* gilt:

$$\mu_W = \mathrm{E}(W_{n_1,n_2}) = \frac{n_1(n_1+n_2+1)}{2} \tag{8.5}$$

und

$$\sigma_W^2 = \mathrm{Var}(W_{n_1,n_2}) = \frac{n_1 n_2(n_1+n_2+1)}{12} \ . \tag{8.6}$$

Die Rangsumme darf nicht zu klein und nicht zu groß ausfallen, weil im ersten Fall die Werte der ersten Probe eher im unteren Bereich der zweiten Probe liegen würden, im zweiten Fall eher im oberen Bereich. Beides stünde im Widerspruch zur Annahme übereinstimmender Verteilungen. Daher lautet die Entscheidungsregel dieses nach der Testgröße benannten *Wilcoxon–Rangsummen-Tests (WRS-Test)* zum Signifikanzniveau α:

$$\left.\begin{array}{c} w_{n_1,n_2} < w_{n_1,n_2}^{2,u} \\ \text{oder} \\ w_{n_1,n_2} > w_{n_1,n_2}^{2,o} \end{array}\right\} \quad \Rightarrow \quad \text{Ablehnung von } H_0 \, . \tag{8.7}$$

Dabei beschreibt

$$w_{n_1,n_2}^{2,u} = \max\left\{c : \ P(W_{n_1,n_2} < c) \leq \frac{\alpha}{2}\right\}$$

gleichsam das untere und

$$w_{n_1,n_2}^{2,o} = \min\left\{d : \ P(W_{n_1,n_2} > d) \leq \frac{\alpha}{2}\right\}$$

das obere $\frac{\alpha}{2}$–Quantil der Verteilung der Rangsumme W_{n_1,n_2} der ersten Stichprobe unter der Annahme indentischer *VFen* F_1 und F_2 (Nullhypothese).

Es gilt:

$$w^{2,u}_{n_1,n_2} = \frac{(n_1+n_2)(n_1+n_2+1)}{2} - w^{2,o}_{n_2,n_1}$$

und aus Symmetriegründen auch

$$w^{2,o}_{n_1,n_2} = \frac{(n_1+n_2)(n_1+n_2+1)}{2} - w^{2,u}_{n_2,n_1} \ .$$

Beweis:
Offensichtlich gilt für die Rangsumme w_{n_1,n_2} der ersten und w_{n_2,n_1} der zweiten Probe

$$w_{n_1,n_2} + w_{n_2,n_1} = \frac{(n_1+n_2)(n_1+n_2+1)}{2} \ .$$

Daher folgt

$$\begin{aligned} P(W_{n_1,n_2} < c) \leq \frac{\alpha}{2} \quad &\Leftrightarrow \quad P\left(\frac{(n_1+n_2)(n_1+n_2+1)}{2} - W_{n_2,n_1} < c\right) \leq \frac{\alpha}{2} \\ &\Leftrightarrow \quad P\left(W_{n_2,n_1} > \frac{(n_1+n_2)(n_1+n_2+1)}{2} - c\right) \leq \frac{\alpha}{2} \ , \end{aligned}$$

woraus sich unmittelbar die obige Beziehung ergibt. △

Die kritischen Werte für den *WRS–Test* sind in Tabelle A.5 zusammengestellt (Hartung, 1990). Falls die Stichprobenumfänge den angegebenen Bereich übersteigen, nützt man die Tatsache aus, dass

$$\frac{W_{n_1,n_2} - \mu_W}{\sigma_W} \ \rightsquigarrow \ N(0,1) \tag{8.8}$$

gilt, wobei "$\rightsquigarrow$" bedeutet, dass die Verteilung der linken Seite mit wachsenden Stichprobenumfängen gegen die rechte Verteilung strebt, in diesem Fall die Standardnormalverteilung. Also kann dann W_{n_1,n_2} als ungefähr $N(\mu_W, \sigma^2_W)$–normalverteilt angenommen werden.

Selbstverständlich kann aus Symmetriegründen die Rangsumme der zweiten Stichprobe als Testgröße herangezogen werden, indem man die Bedeutung der ersten und zweiten Probe bei der Beschreibung des *WRS–Tests* vertauscht.

Beispiel 8.2 Bei zwei Obstplantagen werden Stichproben von der Sorte Golddelicious gezogen. Dabei konnte man folgende Werte für das Apfelgewicht (in *dag*) beobachten:

Plantage A:	12.3	11.6	11.5	15.3	13.1	12.5	13.0	11.4
	14.5	12.2	12.3	14.9	14.1	13.5	13.2	
Plantage B:	13.5	14.2	12.6	14.5	12.9	13.6	12.7	14.8
	15.9	13.7	16.2	13.2				

Besteht zwischen den Verteilungen der Apfelgewichte in den beiden Plantagen ein signifikanter Unterschied? (Sicherheit $1-\alpha = 0.95$)

Lösung:

Die Tabelle 8.4 zeigt die aus den beiden Datensätzen geformte gemeinsame Stichprobe der Größe nach geordnet. Daraus ergeben sich die Rangzahlen für die erste Probe. Als Summe

dieser Rangzahlen erhält man 169.5 . Die kritischen Werte erhält man aus Tab A.5 (wegen $\alpha/2 = 0.025$ jeweils der oberste Wert des Dreierpakets) als

$$w^{2,u}_{15,12} = 170 \quad \text{und} \quad w^{2,u}_{12,15} = 128 \ ,$$

woraus sich

$$w^{2,o}_{15,12} = (15+12)(15+12+1)/2 - w^{2,u}_{12,15} = 378 - 128 = 250$$

ergibt. Da offensichtlich $169.5 < 170$ gilt, ist die Nullhypothese knapp zu verwerfen, es besteht ein leicht signifikanter Einwand *gegen* die Gleichheit der Verteilung des Apfelgewichtes bei den untersuchten Plantagen.

Tabelle 8.4: *Wilcoxon–Rangsummentest*

i	y		r_i	i	y		r_i
	1. Probe	2. Probe			1. Probe	2. Probe	
1	11.4		1	11	13.5		15.5
2	11.5		2			13.5	
3	11.6		3			13.6	
4	12.2		4			13.7	
5	12.3		5.5	12	14.1		19
6	12.3		5.5			14.2	
7	12.5		7	13	14.5		21.5
		12.6				14.5	
		12.7				14.8	
		12.9		14	14.9		24
8	13.0		11	15	15.3		25
9	13.1		12			15.9	
10	13.2		13.5			16.2	
		13.2					

8.1.3 Kruskal–Wallis–Test

Dieser Test (Kruskal und Wallis, 1952) stellt eine Alternative zur einfachen Varianzanalyse dar und ist insbesonders dann sinnvoll einzusetzen, wenn die bei der Varianzanalyse notwendige Normalverteilungsannahme nicht erfüllt ist. Er ist die direkte Erweiterung des Wilcoxon-Tests für mehr als 2 Gruppen. Als einzige Voraussetzung reicht die Annahme *stetiger* Verteilungen *mit derselben Form*. Die Nullhypothese lautet (analog zur einfachen Varianzanalyse)

$$H_0 : \text{"alle } k \text{ Verteilungen stimmen überein"} \ .$$

gegen die Alternative

$$H_A : \text{"mindestens eine Verteilung hat eine andere Lokation"} \ .$$

Zur Überprüfung der Annahme derselben Verteilungsform können wieder Boxplots verwendet werden, die Boxen und Whiskers sollten ähnlich geformt und nur parallel verschoben sein.

Bei kleineren Gruppengrößen (unter 10) sollten statt Boxplots besser direkt die Rohdaten als Punkte in einem Streudiagramm betrachtet werden.

Testvorschrift:

Tabelle 8.5: *Kruskal–Wallis–Test; kritische Werte* $h_{3;(n_1,n_2,n_3);1-\alpha}$

n	n_1	n_2	n_3	$h_{3;(n_1,n_2,n_3);1-\alpha}$	
				$\alpha = 0.10$	$\alpha = 0.05$
7	1	2	4	4.50	4.82
	1	3	3	4.57	5.14
	2	2	3	4.50	4.71
8	1	2	5	4.20	5.00
	1	3	4	4.06	5.21
	2	2	4	4.46	5.13
	2	3	3	4.56	5.14
9	1	3	5	4.02	4.87
	1	4	4	4.07	4.87
	2	2	5	4.37	5.04
	2	3	4	4.51	5.40
	3	3	3	4.62	5.60
10	1	4	5	3.96	4.86
	2	3	5	4.49	5.11
	2	4	4	4.55	5.24
	3	3	4	4.70	5.72
11	1	5	5	4.04	4.91
	2	4	5	4.52	5.27
	3	3	5	4.41	5.52
	3	4	4	4.48	5.58
12	2	5	5	4.51	5.25
	3	4	5	4.52	5.63
	4	4	4	4.50	5.65
13	3	5	5	4.55	5.63
	4	4	5	4.62	5.62
14	4	5	5	4.52	5.64
15	5	5	5	4.56	5.66

1) Alle $n = n_1 + \cdots + n_k$ Messwerte werden *gemeinsam* der Größe nach geordnet; jedem Messwert x_{ij} $(i = 1, \ldots, k;\ j = 1, \ldots, n_i)$ kommt damit ein *Rang* r_{ij}, sein Platz in dieser geordneten Reihe, zu. Damit bedeutet $r_{ij} = 1$, dass x_{ij} der kleinste aller n Messwerte ist, und $r_{ij} = n$, dass x_{ij} der größte unter ihnen ist.

   Liegen dabei insgesamt g *verschiedene* Messwerte $z_1 < z_2 < \cdots < z_g$ vor und

gilt $g < n$, so muss es offensichtlich unter den n Messwerten gleiche geben *(Bindung,* engl. *tie)*; t_l $(l = 1, \ldots, g)$ bezeichne die Anzahl der Messwerte unter den x_{ij} $(i = 1, \ldots, k;\ j = 1, \ldots, n_i)$, die gleich z_l sind; $t_l \geq 2$ bedeutet somit, dass z_l *mehrfach* auftritt. Für derartige Werte werden die vorgesehenen t_l Ränge gemittelt, und dieser (durchschnittliche) Rang den (zu z_l gleichen) Messwerten zugeordnet. Es gilt nun

$$\sum_{i=1}^{k} \sum_{j=1}^{n_i} r_{ij} = \frac{n(n+1)}{2} \ .$$

2) Für jede Gruppe $i = 1, \ldots, k$ bildet man die Summe der Rangzahlen ihrer Messwerte

$$r_{i\cdot} = \sum_{j=1}^{n_i} r_{ij} \ .$$

3) Als Prüfgröße betrachtet man

$$t = \frac{1}{b} \left[\frac{12}{n(n+1)} \sum_{i=1}^{k} \frac{r_{i\cdot}^2}{n_i} - 3(n+1) \right] \tag{8.9}$$

mit der Korrekturgröße

$$b = 1 - \frac{1}{n^3 - n} \sum_{l=1}^{g} (t_l^3 - t_l)$$

beim Vorliegen von Bindungen (sonst gilt $b = 1$). Die Prüfgröße stellt eine Art Varianz der durchschnittlichen Rangzahlen der einzelnen Gruppen dar. Je mehr diese Durchschnittswerte voneinander abweichen, umso größer wird also t; große Werte für t sprechen damit gegen die Hypothese der Gleichheit.

4) Unter Verwendung des (1-α)–Quantils $h_{k;(n_1,\ldots,n_k);1-\alpha}$ der Verteilung der Prüfgröße (8.9) bei Gültigkeit der Nullhypothese (siehe Tab. 8.5) lautet die Entscheidungsregel

$$t \left\{ \begin{array}{c} \leq \\ > \end{array} \right\} h_{k;(n_1,\ldots,n_k);1-\alpha} \quad \Rightarrow \quad H_0 \left\{ \begin{array}{l} \text{beibehalten} \\ \text{ablehnen} \ . \end{array} \right.$$

Neben dem Tabellenausschnitt in Tab. 8.5 (Hartung, 2002) finden sich ausführliche Tabellen in Hollander und Wolfe (1972).

Für den Fall großer Stichprobenumfänge approximiert man die Verteilung der Prüfgröße (8.9) durch die χ^2_{k-1}–Verteilung. Dann kann man als kritischen Wert $\chi^2_{k-1;1-\alpha}$ wählen.

Beispiel 8.3 Drei Weizensorten werden hinsichtlich ihrer *ha*–Erträge (in *dt*) verglichen; bei verschiedenen Landwirten ergaben sich nachfolgende Werte, wobei jeder Landwirt bloß eine Sorte anbaut:

Sorte	*ha*–Ertrag				
1	41	47	50	43	48
2	42	39	40	44	
3	48	55	54	52	

Liefern die Sorten "gleiche" Erträge? (Vergleiche Beispiel 6.1)

Lösung:
Eine Analyse mit Hilfe des Tests von Kruskal und Wallis liefert folgende Tabelle:

Sorte 1		Sorte 2		Sorte 3	
		39	*1*		
		40	*2*		
41	*3*				
		42	*4*		
43	*5*				
		44	*6*		
47	*7*				
48	*8.5*				
				48	*8.5*
50	*10*				
				52	*11*
				54	*12*
				55	*13*
$\sum$	*33.5*	$\sum$	*13*	$\sum$	*44.5*

Die schräggestellten Zahlen geben dabei den Rang der Messwerte in dem gemeinsamen Datensatz an. Mit dem Wert 48 gibt es die einzige Bindung in diesem Datensatz. Damit ergibt sich die Korrekturgröße zu

$$b = 1 - \frac{6}{13^3 - 13} = \frac{363}{364} = 0.997$$

und als Testgröße erhält man schließlich

$$t = \frac{364}{363}\left[\frac{12}{13 \times 14}\left(\frac{33.5^2}{5} + \frac{13^2}{4} + \frac{44.5^2}{4}\right) - 3 \times 14\right] = 8.249$$

Laut Tab.8.5 beträgt der kritische Wert bei einem Signifikanzniveau $\alpha = 0.05$ und den Stichprobenumfängen $(5, 4, 4)$ 5.62 und ist damit kleiner als der Wert der Testgröße t. Die Hypothese gleicher Verteilungen ist daher zu *verwerfen*; es gibt einen *signifikanten* Unterschied der Sorten hinsichtlich des *ha*–Ertrages.

8.2 Vorzeichentest

Der Vorzeichentest als Spezialfall des Binomialtests (Test über den Anteilsparameter) lässt sich zur Überprüfung von Hypothesen einsetzen, dass positive und negative Werte eines betrachteten Merkmals X gleich wahrscheinlich sind. Sein großer Vorteil liegt darin, dass er frei von jeglichen Verteilungsannahmen ist.

Als Testgröße verwendet man die Anzahl v von positiven Beobachtungswerten. Wenn man jedem Wert x_i einer Stichprobe $x_1, x_2, \ldots, x_n$ einen Wert v_i ("Vorzeichen") zuordnet von der Form

$$x_i \;\rightarrow\; v_i = \begin{cases} 1 & \text{für } x_i < 0 \\ 0.5 & \text{für } x_i = 0 \\ 0 & \text{für } x_i > 0\ , \end{cases}$$

so ist v gegeben durch

$$v = \sum_{i=1}^{n} v_i$$

und diese Größe ist unter der Nullhypothese ("positive und negative Werte sind gleich wahrscheinlich"), abgesehen von den möglichen Summanden 0.5, $Bi(n, 0.5)$–binomialverteilt. Die Zuordnung von 0.5 im Falle, dass ein beobachteter Wert exakt null ist, trägt der Tatsache Rechnung, dass eine derartige Situation theoretisch nicht auftreten dürfte (Wahrscheinlichkeit null!), aber in der Praxis auf Grund von Genauigkeitsgrenzen dennoch vorkommt. Man teilt diesen Wert daher "gerecht" auf die unmittelbar an null angrenzenden Bereiche auf.

Falls die Nullhypothese zutrifft, *erwartet* man $n/2$ '+'–Zeichen in der Stichprobe, es dürften also nicht zu wenige und auch nicht zu viele auftreten. Daher sprechen sowohl eine große Anzahl von '+'–Zeichen als auch eine geringe Anzahl eher gegen diese Hypothese, also wählt man als kritische Wert das obere $\alpha/2$–Quantil

$$b^o_{n,0.5;\alpha/2} = \min \left\{ d \,:\, \sum_{l=d}^{n} \binom{n}{l} 0.5^n \leq \alpha/2 \right\}$$

bzw. das untere $\alpha/2$–Quantil

$$b^u_{n,0.5;\alpha/2} = \max \left\{ c \,:\, \sum_{l=0}^{c} \binom{n}{l} 0.5^n \leq \alpha/2 \right\}$$

der $Bi(n, 0.5)$–Verteilung.

Häufig wird dieser Vorzeichentest zur Überprüfung des Medians verwendet. Eine Anwendung davon ergibt sich aus der Situation des folgenden Abschnittes. Ein Spezialfall dafür ist der Vergleich verbundener Stichproben (z.B. Körpergröße einer Person am Morgen und am Abend, Reaktionszeit einer Person mit und ohne Alkoholeinfluss), wenn die Normalapproximation (vgl. Abschnitt 5.4) wegen eines zu geringen Stichprobenumfanges nicht greift.

Für zwei verbundene Proben $(x_1, y_1), \ldots, (x_n, y_n)$ berechnet man die Differenzen $d_i = x_i - y_i$. Die Übereinstimmung der Verteilung von X und Y hat zur Folge, dass die Differenz D symmetrisch um null verteilt ist. Die Wahrscheinlichkeit einer *positiven* Differenz wäre gleich der einer *negativen* Differenz, sodass sich der Vorzeichentest einsetzen lässt.

Beispiel 8.4 *Beispiel 8.0:*

Für die Agrarstatistik wird im Rahmen einer Befragung das bäuerliche Jahreseinkommen (in 1.000 €) für die Jahre 2000 und 2001 erhoben (wie Beispiel 5.5):

Landwirt Jahr	1	2	3	4	5	6
2000	18.47	19.48	11.94	16.39	8.45	32.50
2001	20.54	20.09	13.01	17.33	9.10	32.97
$v(d_{01-00})$	+	+	+	+	+	+

Ist das Jahreseinkommen 2000 und 2001 identisch verteilt ("gleich")? (Signifikanzniveau $\alpha = 0.05$)

Lösung:
Betrachtet man die Differenz Einkommen–01 minus Einkommen–00, so bedeutet die Nullhypothese, dass Zuwächse und Verluste gleichwahrscheinlich wären. Es gibt 6 '+'–Vorzeichen; "verträgt" sich das mit der Nullhypothese? Hier lässt sich ausnahmsweise der p–Wert, das empirische Signifikanzniveau, leicht berechnen:

$$\text{P("6 oder mehr '+'–Zeichen"}|n=6,p=0.5) = \binom{6}{0}0.5^6 = 0.0156 \ .$$

Da es sich um eine zweiseitige Fragestellung handelt (zuwenige und zuviele '+'–Zeichen sprechen gegen die Nullhypothese), muss man den p–Wert mit $\alpha/2$ vergleichen: 0.0156 ist offensichtlich kleiner als 0.025, also kann die Hypothese verworfen werden, es gibt also einen signifikanten Unterschied in der bäuerlichen Einkommensverteilung 2000 und 2001.

Selbstverständlich hätte man die Nullhypothese auch einseitig formulieren können: "das Einkommen 2001 ist niedriger als 2000" (negativ formuliert, damit die Chance auf Ablehnung besteht!). Ausschließlich zuviele '+'–Zeichen sprechen dann gegen diese Nullhypothese. Der oben berechnete p–Wert wäre also mit dem *gesamt*zulässigen Risiko α zu vergleichen. Da 0.0156 natürlich auch kleiner als 0.05 ist, kann auch diese Nullhypothese abgelehnt werden, das Einkommen ist somit *signifikant* gestiegen.

8.3 Tests auf Verteilung

8.3.1 Der χ^2–Test

Der im folgenden beschriebene *Anpassungstest* zählt zu den naheliegendsten und einfachsten Verfahren zur Beurteilung von Modellen, zu denen letztlich auch Wahrscheinlichkeitsverteilungen gehören.

Der einfache χ^2–Test

Im einfachsten Fall wird diese Methode zur Überprüfung einer konkreten Verteilungsannahme eingesetzt. Folgende Ausgangssituation liegt zugrunde:

1) die *ZG* X besitzt die *W-Vt* P_X;
2) für den Merkmalraum M_X von X existiert eine Zerlegung $M_X = K_1 \cup K_2 \cup \cdots \cup K_k$ in disjunkte Bereiche (Klassen).

Zu untersuchen ist die Behauptung (Nullhypothese)

$$H_0 \ : \ P_X = P_0 \quad , \tag{8.10}$$

dass also P_0 die Verteilung von X ist. Da hier die Nullhypothese bloß *eine* Verteilung umfasst, spricht man auch von einer *einfachen* Nullhypothese und demnach auch vom *einfachen* χ^2–Test.

Zur Beurteilung dieser Behauptung wird eine Stichprobe $x_1, x_2, \ldots, x_n$ herangezogen. Für diese werden die *beobachteten* (absoluten) Häufigkeiten $y_{n;l}$ für die k Klassen mit den auf Grund der Annahme H_0 zu *erwartenden* Häufigkeiten e_l $(l = 1, \ldots, k)$ verglichen. Starke

Abweichungen sprechen dabei gegen das Zutreffen der Nullhypothese und führen daher zu ihrer Ablehnung.

Da die absolute Häufigkeit $Y_{n;l}$ der Klasse K_l $(l = 1, \ldots, k)$ mit den Parametern n und

$$p_l = P(X \in K_l | H_0) = P_0(K_l)$$

binomialverteilt ist, gilt für den Erwartungswert

$$e_l = \mathrm{E}(Y_{n;l} | H_0) = np_l \ .$$

Aus technischen Gründen misst man den Unterschied durch die *normierten quadratischen Abstände ('Residuen')*

$$\frac{(y_{n;l} - e_l)^2}{e_l} \tag{8.11}$$

und verwendet deren Summe

$$t = \sum_{l=1}^{k} \frac{(y_{n;l} - e_l)^2}{e_l} \tag{8.12}$$

als Testgröße. Unter der Nullhypothese gilt

$$\lim_{n \to \infty} P(T \leq \chi^2_{k-1;\gamma}) = \gamma \ , \tag{8.13}$$

dass also die Testgröße asymptotisch χ^2–verteilt ist mit $k - 1$ Freiheitsgraden. Daher rührt auch die Bezeichnung für diesen Test.

Abweichungen von der ursprünglichen Verteilungsannahme werden sich in größeren Unterschieden zwischen beobachteter und auf Grund eben dieser Annahme erwarteter Häufigkeit zumindest für eine, wenn nicht mehrere Klassen auswirken. Daher sprechen *große* Werte der Teststatistik (8.12) *gegen* die Nullhypothese und führen zur *Ablehnung* der ursprünglichen Annahme. Als kritischen Wert wählt man $c_o = \chi^2_{k-1;1-\alpha}$. Die Auswertung erfolgt meist in einem Schema, wie es in Tab. 8.6 verwendet wird.

Bemerkung 1: Da der χ^2–Test auf asymptotischen Argumenten beruht, muss sichergestellt sein, dass die Besetzungszahl $y_{n;l}$ jeder Klasse groß genug ist. Eine Faustregel verlangt mindestens *fünf* zu erwartende Beobachtungen $(e_l \geq 5)$ für jede Klasse. Falls diese Bedingung nicht erfüllt ist, muss man durch geeignete Zusammenfassung von Klassen trachten, ihr zu entsprechen. Bei ordinal skalierten Merkmalen kann man die Bedingung für *Randklassen* abschwächen und verlangt dann, dass in ihnen jeweils wenigstens *eine* Beobachtung *erwartet* wird.

Bemerkung 2: Aus technischen Gründen wird in Statistikpaketen keine Entscheidung getroffen, die auf einem *Vergleich* mit *kritischen Werten* (hier: $\chi^2_{k-1;1-\alpha}$) beruht. Vielmehr ist einfach die Wahrscheinlichkeit p angegeben, dass *unter der Nullhypothese* die Teststatistik für die Nullhypothese schlechter ausfällt als der tatsächlich beobachtete (d.h. ermittelte) Wert. Im Falle des χ^2–Tests ergibt sich p somit als

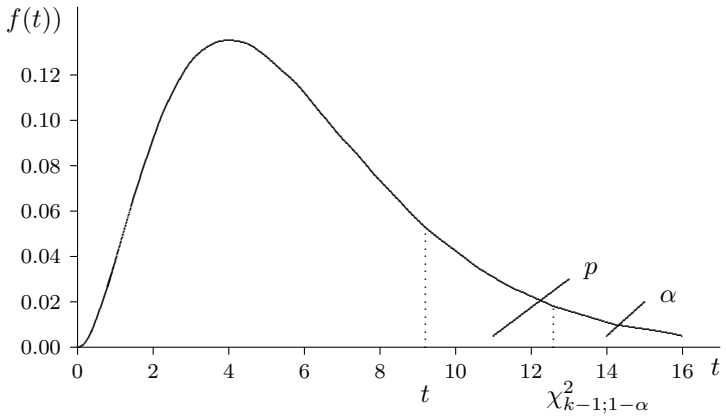
$$p = \mathrm{P}(T > t | H_0) \tag{8.14}$$

Dieser Wert wird als *p–Wert (p–value)* oder *empirisches Signifikanzniveau* bezeichnet. Nun gilt offensichtlich (siehe Abb.8.1)

$$p \left\{ \begin{array}{c} \geq \\ < \end{array} \right\} \alpha \;\Rightarrow\; t \left\{ \begin{array}{c} \leq \\ > \end{array} \right\} \chi^2_{k-1;1-\alpha} \;\Rightarrow\; \left\{ \begin{array}{l} \text{Beibehaltung} \\ \text{Ablehnung} \end{array} \right\} \text{ von } H_0 \quad , \tag{8.15}$$

also lässt sich die Entscheidung vom *Programmbenutzer* auf Grund des p–Wertes in völlig äquivalenter Weise treffen.

Abbildung 8.1: *Verteilung der Teststatistik unter* H_0



Beispiel 8.5 Ein Spielwürfel wird auf seine Ausgewogenheit hin überprüft. Die Grundlage dafür ist die beobachtete Augenzahl X und zu beurteilen ist die Hypothese

$$H_0 \;:\; X \sim D(6) \;,$$

dass also X für die sechs möglichen Augenzahlen (diskret) gleichverteilt ist. Die für die Anwendung des χ^2–Tests notwendige Klasseneinteilung des Merkmalbereiches M_X ergibt sich hier auf natürliche Art, indem jede mögliche Augenzahl eine eigene Klasse beschreibt.

Tabelle 8.6: *Auswertung für einen einfachen* χ^2*–Test*

K_l	$y_{n;l}$	p_l	$e_l = np_l$	$y_{n;l} - e_l$	$(y_{n;l} - e_l)^2/e_l$
1	23	1/6	20	3	0.45
2	19	1/6	20	-1	0.05
3	27	1/6	20	7	2.45
4	14	1/6	20	-6	1.80
5	16	1/6	20	-4	0.80
6	21	1/6	20	1	0.05
$\sum$	120	1.0	120	0	5.60

Es werden $n = 120$ Würfelversuche durchgeführt und die beobachteten Häufigkeiten $y_{n;l}$ jeder Klasse festgehalten (siehe Tab. 8.6). Unter der Hypothese eines fairen Würfels (H_0) gilt

für die Wahrscheinlichkeit aller Klassen

$$p_1 = p_2 = \cdots = p_6 = \frac{1}{6}$$

und damit werden für die einzelnen Klassen

$$e_l = np_l = \frac{n}{6}$$

Beobachtungen erwartet. Der Vergleich zwischen $y_{n;l}$ und e_l erfolgt gemäß (8.11) und ist in Tab. 8.6 ersichtlich. Als Wert der Teststatistik (8.12) erhält man

$$t = 5.60 < 11.07 = \chi^2_{5;0.95}$$

und dieser ist offensichtlich kleiner als der kritische Wert $\chi_{k-1;1-\alpha}$, sodass beim Signifikanzniveau $\alpha = 0.05$ die Nullhypothese beibehalten wird, also *kein* signifikanter Einwand gegen einen fairen Würfel vorliegt.

Der zusammengesetzte χ^2–Test

Die Ausgangssituation gleicht der beim einfachen χ^2–Test, bloß lautet die Nullhypothese hier

$$H_0 \ : \ P_X \in \{P_\theta : \ \theta \in \Theta\} \ . \tag{8.16}$$

Es wird also die Behauptung untersucht, dass die Verteilung von X zu der durch den Parameter(vektor) θ beschriebenen ("parametrisierten") Familie $\{P_\theta : \ \theta \in \Theta\}$ von *W-Vten* gehört.

Die einzelnen Schritte zur Testauswertung erfolgen wie beim einfachen χ^2–Test. Ein Problem stellt die Ermittlung der Wahrscheinlichkeiten der einzelnen Klassen dar, da hier eine eindeutig bestimmte Verteilung fehlt. Durch *ML*-Schätzung der Parameter erhält man Schätzwerte $\hat{\theta}$ für θ, sodass mit $P_{\hat{\theta}}$ quasi wieder *eine* Verteilung durch die Nullhypothese beschrieben wird. Mit dieser ermittelt man die gesuchten Klassenwahrscheinlichkeiten p_l, die aber wegen der geschätzten $\hat{\theta}$ selbst nur Schätzungen

$$\hat{p}_l = P_{\hat{\theta}}(X \in K_l)$$

darstellen. Damit erhält man $\hat{e}_l = n\hat{p}_l$ als Schätzungen für die erwarteten Häufigkeiten und kann dann analog zum einfachen χ^2–Test die Teststatistik

$$t = \sum_{l=1}^{k} \frac{(y_{n;l} - \hat{e}_l)^2}{\hat{e}_l} \tag{8.17}$$

berechnen. Für diese gilt im Falle des Zutreffens der Nullhypothese

$$\lim_{n\to\infty} P(T \le \chi^2_{k-s-1;\gamma}) = \gamma \ , \tag{8.18}$$

also ist auch hier die Teststatistik asymptotisch χ^2–verteilt mit $k - s - 1$ Freiheitsgraden, wobei s die Anzahl der zu schätzenden Parameter in der Nullhypothese angibt. Mit jedem unbekannten und daher zu schätzenden Parameter reduziert sich somit im Vergleich zum einfachen χ^2–Test die Zahl der Freiheitsgrade um eins.

Tabelle 8.7: *Auswertung für einen zusammengesetzten χ^2–Test*

l	K_l	$y_{n;l}$	$\hat{p}_l$	$\hat{e}_l = n\hat{p}_l$	$y_{n;l} - \hat{e}_l$	$(y_{n;l} - \hat{e}_l)^2/\hat{e}_l$
1	$(-\infty, 139]$	10	0.1079	8.632	1.368	0.2168
2	$(139, 147]$	12	0.1409	11.272	0.728	0.0470
3	$(147, 155]$	18	0.2039	16.312	1.688	0.1747
4	$(155, 163]$	22	0.2175	17.400	4.600	1.2161
5	$(163, 171]$	5	0.1711	13.688	-8.688	5.5144
6	$(171, 179]$	6	0.0993	7.944	-1.944	0.4757
7	$(179, 187]$	4	0.0424	3.392	0.608	0.1090
8	$(187, \infty)$	3	0.0170	1.360	1.640	1.9776
$\sum$		80	1.0000	80.000	0.000	9.7293

Beispiel 8.6 Der Datensatz aus Bsp. 1.1 soll auf Normalverteilung hin untersucht werden. Es bietet sich ein zusammengesetzter χ^2–Test mit den zwei zu schätzenden Parametern μ und σ^2 einer Normalverteilung an. Diese ergeben sich zu $\hat{\mu} = 156.7$ $[g]$ und $\hat{\sigma}^2 = 204.78$ bzw. $\hat{\sigma} = 14.3$ $[g]$.

Damit ist man nun in der Lage, die Klassenwahrscheinlichkeiten p_l zu schätzen. So ergibt sich etwa $\hat{p}_1$ als

$$\hat{p}_1 = P(X \leq 139|\hat{\mu}, \hat{\sigma}) = \Phi\left(\frac{139 - 156.7}{14.3}\right) = \Phi(-1.238) = 0.1079 \ ,$$

wobei zu beachten ist, dass die erste (ebenso wie die letzte) Klasse zur Außenseite hin nicht beschränkt ist. Für $\hat{p}_2$ erhält man

$$\begin{aligned} \hat{p}_2 &= P(139 < X \leq 147|\hat{\mu}, \hat{\sigma}) = \Phi\left(\frac{147 - 156.7}{14.3}\right) - \Phi\left(\frac{139 - 156.7}{14.3}\right) \\ &= \Phi(-0.678) - \Phi(-1.238) = 0.2488 - 0.1079 = 0.1409 \ , \end{aligned}$$

für $\hat{p}_3$ ergibt sich

$$\begin{aligned} \hat{p}_3 &= P(147 < X \leq 155|\hat{\mu}, \hat{\sigma}) = \Phi\left(\frac{155 - 156.7}{14.3}\right) - \Phi\left(\frac{147 - 156.7}{14.3}\right) \\ &= \Phi(-0.119) - \Phi(-0.678) = 0.4527 - 0.2488 = 0.2039 \end{aligned}$$

und für $\hat{p}_4$

$$\begin{aligned} \hat{p}_4 &= P(155 < X \leq 163|\hat{\mu}, \hat{\sigma}) = \Phi\left(\frac{163 - 156.7}{14.3}\right) - \Phi\left(\frac{155 - 156.7}{14.3}\right) \\ &= \Phi(0.441) - \Phi(-0.119) = 0.6702 - 0.4527 = 0.2175 \end{aligned}$$

Mit den restlichen Wahrscheinlichkeiten verfährt man analog. Für die rechte Randklasse erhält man

$$\begin{aligned} \hat{p}_8 &= P(187 < X|\hat{\mu}, \hat{\sigma}) = 1 - \Phi\left(\frac{187 - 156.7}{14.3}\right) \\ &= 1 - \Phi(2.119) = 1 - 0.9830 = 0.0170 \ . \end{aligned}$$

Damit lässt sich die Auswertung der Teststatistik (8.17), wie in Tab. 8.7 dargestellt, durchführen. Es gibt $k = 8$ Klassen und $s = 2$ Parameter müssen geschätzt werden. Man erhält schließlich mit

$$t = 9.7293 < 11.07 = \chi^2_{5;0.95}$$

einen Wert für die Teststatistik, der bei $\alpha = 0.05$ unter dem kritischen Wert $\chi^2_{k-s-1;1-\alpha}$ liegt, sodass die Nullhypothese beibehalten wird, also kein signifikanter Einwand gegen die Normalverteilungsannahme besteht.

8.3.2 Kolmogorov–Smirnov–Test

Der Kolmogorov–Smirnov–Test (*KS*–Test) zur Verteilungsüberprüfung zieht für die statistische Beurteilung einer Verteilungsannahme die empirische *VF* $\hat{F}_n$ (siehe 2.3.6) heran.

Einstichproben–*KS*–Test (*KS-1*–Test)

Die Ausgangssituation ist dieselbe wie beim einfachen χ^2–Test, bloß wird die *ZG* X als kontinuierlich vorausgesetzt. Die Nullhypothese H_0 für die *VF* lautet:

$$H_0 \; : \; F_X = F_0 \; . \tag{8.19}$$

Für eine Stichprobe $x_1, \ldots, x_n$ verwendet man als Teststatistik

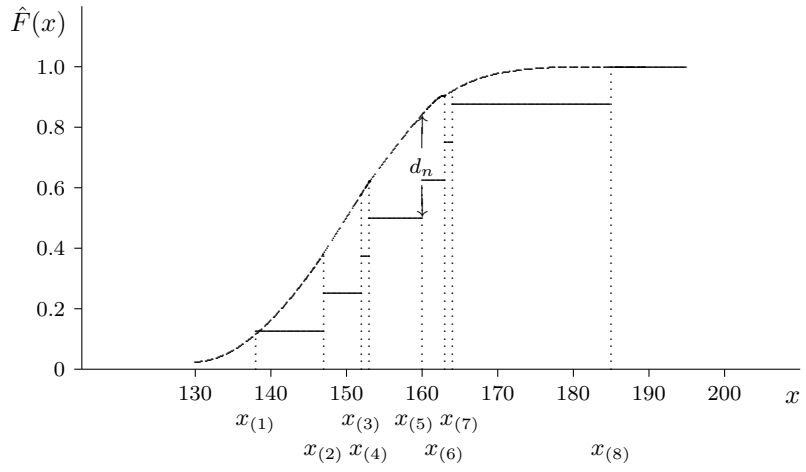
$$t = \sqrt{n}\, d_n \tag{8.20}$$

mit

$$d_n = \sup_{x \in \boldsymbol{R}} |F_0(x) - \hat{F}_n(x)| = \max_{i=1,\ldots,n} |F_0(x_i) - \hat{F}_n(x_i)| \; , \tag{8.21}$$

also dem maximalen Unterschied zwischen der behaupteten *VF* und der empirischen *VF*. Die Multiplikation mit $\sqrt{n}$ dient der Normierung. Die Idee wird anhand der Abb. 8.2 deutlich.

Abbildung 8.2: *Einstichproben–Kolmogorov–Smirnov–Test*



Die Verteilung dieser Teststatistik hängt bei zutreffender Nullhypothese *nicht* von der konkreten *VF* F_0 ab und ist zudem exakt beschreibbar. Solche Verfahren werden naheliegenderweise als *verteilungsfreie* Methoden bezeichnet.

Im Falle der Gültigkeit der Nullhypothese dürfte d_n nicht zu groß ausfallen, daher wird der kritische Bereich bei diesem Test durch *große* Werte für d_n bzw. t gebildet. Die kritischen Werte für einige typische Stichprobenumfänge finden sich in Tabelle 8.8 (Hartung, 2002).

Tabelle 8.8: *Kritische Werte* $d^{(1)}_{n;1-\alpha}$ *für KS-1–Test*

n	$d^{(1)}_{n;0.80}$	$d^{(1)}_{n;0.90}$	$d^{(1)}_{n;0.95}$	$d^{(1)}_{n;0.98}$	$d^{(1)}_{n;0.99}$
5	1.00	1.14	1.26	1.40	1.50
8	1.01	1.16	1.28	1.43	1.53
10	1.02	1.17	1.29	1.45	1.55
20	1.04	1.19	1.31	1.47	1.57
40	1.05	1.20	1.33	1.49	1.59
> 40	1.08	1.23	1.36	1.52	1.63

Da die Verteilung der Teststatistik zu diesem Test bekannt ist, sind die Aussagen exakt. Daher ist der *KS-1*–Test für kleine Stichprobenumfänge dem χ^2–Test in der Regel vorzuziehen.

Der *KS-1*–Test lässt sich analog dem zusammengesetzten χ^2–Test für (zusammengesetzte) Hypothesen erweitern, die eine Verteilungsfamilie beschreiben (z.B.: H_0 – "Normalverteilung"). Allerdings geht dabei die Unabhängigkeit der Teststatistik von der zugrundeliegenden Verteilung verloren. Für jede Verteilungsfamilie ergeben sich eigene kritische Werte, die für die Normalverteilung u.a. von Lilliefors (1967) analysiert wurden. Details findet man bei Hartung (2002).

Beispiel 8.7 Anhand der 8 Beobachtungen in Beispiel 8.1 soll mit Hilfe des *KS-1*–Tests überprüft werden, ob eine $N(150, 10^2)$–Verteilung zugrundeliegt (Sicherheit $1 - \alpha = 0.95$).

Tabelle 8.9: *Einstichproben–Kolmogorov–Smirnov–Test*

i	$x_{(i)}$	$F_0(x_{(i)})$	$\hat{F}_n(x_{(i)})$	d_u	d_o
1	138	0.115	0.125	0.115	0.010
2	147	0.382	0.250	0.257	0.132
3	152	0.579	0.375	0.329	0.204
4	153	0.618	0.500	0.143	0.118
5	160	0.841	0.625	*0.341*	0.216
6	163	0.903	0.750	0.278	0.153
7	164	0.919	0.875	0.169	0.044
8	185	1.000	1.000	0.125	0

Die Tab. 8.9 enthält die 8 Stichprobenwerte in geordneter Form, die Werte der postulierten

VF

$$F_0(x_{(i)}) = \Phi\left(\frac{x_{(i)} - 150}{10}\right)$$

und der empirischen *VF*, sowie die Differenzen d_u bzw. d_o zwischen beiden, die dadurch entstehen, dass an den Sprungstellen der empirischen *VF* sowohl der kleinere als auch der größere Funktionswert betrachtet werden kann:

$$\begin{aligned} d_u(x_{(i)}) &= |F_0(x_{(i)}) - \hat{F}_n(x_{(i)}-)| = |F_0(x_{(i)}) - \hat{F}_n(x_{(i-1)})| \\ d_o(x_{(i)}) &= |F_0(x_{(i)}) - \hat{F}_n(x_{(i)})| \end{aligned}$$

wobei $g(x-) = \lim_{y\uparrow x} g(y)$ für den *linksseitigen Grenzwert* der Funktion g steht.

Wie man aus dieser Tabelle und in Abb. 8.2 sieht, ergibt sich der größte Unterschied zwischen der *VF* der $N(150, 10^2)$–Verteilung und der empirischen *VF* bei $x_{(5)} = 160$ und beträgt $d_n = 0.341$, sodass sich der Wert der Teststatistik als $t = \sqrt{8} \times 0.341 = 0.964$ ergibt. Da

$$t = 0.964 < 1.28 = d^{(1)}_{8;0.95}$$

ausfällt, wird die Nullhypothese beibehalten, es besteht also *kein signifikanter Einwand* gegen das Vorliegen einer $N(150, 10^2)$–Verteilung.

Zweistichproben–*KS*–Test (*KS-2*–Test)

Die Idee des *KS*–Tests, die empirische *VF* als Beurteilungsgrundlage einer *W-Vt* heranzuziehen, lässt sich auch auf den Vergleich zweier Verteilungen übertragen. Die *VFen* zu den (kontinuierlichen) *ZGen* X und Y seien F_X und F_Y. Es ergeben sich weitergehende Vergleichsmöglichkeiten, wenn man folgende Definition betrachtet.

Definition: Die *ZG* X heißt *stochastisch größer* als die *ZG* Y, wenn

$$F_X(t) \leq F_Y(t) \quad \text{für alle } t \in \boldsymbol{R}\,, \tag{8.22}$$

wenn es also für jeden Wert t wahrscheinlicher ist, dass Y kleiner als t ausfällt, als dass dies für X gilt.

Damit lassen sich folgende Nullhypothesen untersuchen:

a) H_0 : $F_X = F_Y$
b) $H_0^{>}$: X ist stochastisch größer als Y, d.h. $F_X(t) \leq F_Y(t) \quad \forall t \in \boldsymbol{R}$
c) $H_0^{<}$: X ist stochastisch kleiner als Y, d.h. $F_X(t) \geq F_Y(t) \quad \forall t \in \boldsymbol{R}$

Bei der Überprüfung der Nullhypothese im Fall a) geht man von Stichproben $x_1, \ldots, x_{n_X}$ und $y_1, \ldots, y_{n_Y}$ für X und Y aus und betrachtet den Maximalabstand zwischen den beiden zugeordneten empirischen *VFen*

$$d_{n_X,n_Y} = \max_{t\in\boldsymbol{R}} |\hat{F}_{X;n}(t) - \hat{F}_{Y;n}(t)|\,. \tag{8.23}$$

Große Werte für d_{n_X,n_Y} deuten auf eine Abweichung zwischen den beiden zugrundeliegenden theoretischen *VFen* hin und werden demnach zur Ablehnung der Nullhypothese führen.

Fall $n_X = n_Y = n$:

Die Tabelle 8.10 (siehe Hartung, 2002) enthält in der linken Randspalte kritische Werte $d^{(2)}_{n;1-\alpha}$, die ab den Stichprobenumfängen n gültig sind, die im Tabellenrumpf in der zum gegebenen Signifikanzniveau α gehörenden Spalte aufscheinen. Für Stichprobenumfänge $n > 40$ enthält die letzte Zeile Formeln zur Bestimmung der kritischen Werte $d^{(2)}_{n;1-\alpha}$. Als Testgröße wählt man

$$t = n\, d_{n_X,n_Y} \tag{8.24}$$

und verwirft die Nullhypothese der Gleichheit von F_X und F_Y, falls

$$t > d^{(2)}_{n;1-\alpha} \tag{8.25}$$

gilt.

Tabelle 8.10: *Kritische Werte $d^{(2)}_{n;1-\alpha}$ für KS-2–Test*

n	α				
	0.20	0.10	0.05	0.02	0.01
2	3	3			
3	4	4	4		
4	7	6	5	5	5
5	11	9	7	6	6
6	16	13	10	9	8
7	22	17	14	11	10
8	28	22	18	15	13
9	36	28	23	18	16
10		34	28	22	20
11			33	27	24
12			40	32	28
13				37	33
14					38
$n > 40$					
	$1.527\sqrt{n}$	$1.739\sqrt{n}$	$1.923\sqrt{n}$	$2.150\sqrt{n}$	$2.305\sqrt{n}$

Fall $n_X \neq n_Y$:

Die Tabelle 8.11 (siehe Hartung, 2002) enthält für einige Signifikanzwerte α approximative kritische Werte $d^{(2)}_{n_X,n_Y;1-\alpha}$, wobei mit $n_{(1)}$ der kleinere der beiden Stichprobenumfänge bezeichnet wird. Als Testgröße wählt man

$$t = \sqrt{\frac{1}{\frac{1}{n_X}+\frac{1}{n_Y}}}\, d_{n_X,n_Y} = \sqrt{\frac{n_X n_Y}{n_X+n_Y}}\, d_{n_X,n_Y} \tag{8.26}$$

und verwirft die Nullhypothese, wenn

$$t > d^{(2)}_{n_X,n_Y;1-\alpha} \tag{8.27}$$

ausfällt.

Tabelle 8.11: *Kritische Werte* $d^{(2)}_{n_X,n_Y;1-\alpha}$ *für KS-2–Test*

α	$n_{(1)}$	$n_{(2)}$	$d^{(2)}_{n_{(1)},n_{(2)};1-\alpha}$
0.20	2–4	5–40	1.02
	5–15	5–40	1.03
	sonst		1.08
0.10	2–3	3–12	1.10
	4–8	5–9	1.12
	4–16	10–20	1.16
	sonst		1.23
0.05	2–4	3–15	1.22
	5–16	6–20	1.30
	sonst		1.36

Festsetzung: $n_{(1)} < n_{(2)}$

Beispiel 8.8 Bei zwei Obstplantagen werden Stichproben von der Sorte Golddelicious gezogen. Dabei konnte man folgende Werte für das Apfelgewicht (in *dag*) beobachten:

Plantage A:	12.3	11.6	11.5	15.3	13.1	12.5	13.0	11.4
	14.5	12.2	12.3	14.9	14.1	13.5	13.2	
Plantage B:	13.5	14.2	12.6	14.5	12.9	13.6	12.7	14.8
	15.9	13.7	16.2	13.2				

Besteht zwischen den Verteilungen der Apfelgewichte in den beiden Plantagen ein signifikanter Unterschied? (Sicherheit $1 - \alpha = 0.95$)

Lösung:

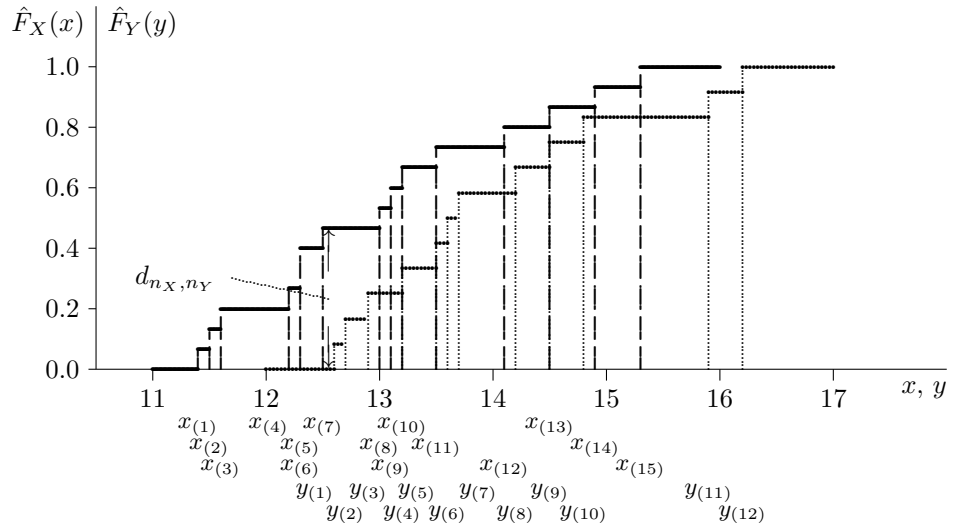
Die nachstehende Tabelle zeigt für die der Größe nach geordneten Beobachtungen die Werte der beiden empirischen *VF*en $\hat{F}_X(t)$ (Plantage A) und $\hat{F}_Y(t)$ (Plantage B), sowie deren Differenz $\Delta(t)$. Die Knollen zeigen an, in welcher Plantage (A oder B) der jeweilige Wert beobachtet wurde; mehrere Knollen deuten auf ebenso viele Beobachtungen hin (der Wert 12.3 wurde also zweimal, und zwar in Plantage A beobachtet).
Die Abb. 8.3 zeigt diesen Sachverhalt auch grafisch. Die maximale absolute Differenz zwischen den beiden empirischen *VF*en beträgt 0.467 und somit ergibt sich die Testgröße t zu

$$t = \sqrt{\frac{15 \cdot 12}{15 + 12}} \times 0.467 = 1.206$$

Da der (obere) kritische Wert $d^{(2)}_{12,15;0.95} = 1.30$ (Tab, 8.11) gößer als die Testgröße ist, wird H_0 beibehalten, es gibt also keinen signifikanten Einwand gegen gleiche Verteilungen des Apfelgewichtes auf Plantage A und B. Allerdings hat der KS-Test bei derart kleinen Stichproben auch keine besonders große Macht, verwirft also die Nullhypothese nur bei sehr starken Unterschieden in den empirischen Verteilungen.

t	A	B	$\hat{F}_X(t)$	$\hat{F}_Y(t)$	$\Delta(t)$	t	A	B	$\hat{F}_X(t)$	$\hat{F}_Y(t)$	$\Delta(t)$
11.4	•		0.067	0	0.067	13.5	•	•	0.733	0.417	0.316
11.5	•		0.133	0	0.133	13.6		•	0.733	0.500	0.233
11.6	•		0.200	0	0.200	13.7		•	0.733	0.583	0.150
12.2	•		0.267	0	0.267	14.1	•		0.800	0.583	0.217
12.3	••		0.400	0	0.400	14.2		•	0.800	0.667	0.133
12.5	•		0.467	0	*0.467*	14.5	•	•	0.867	0.750	0.117
12.6		•	0.467	0.083	0.384	14.8		•	0.867	0.833	0.034
12.7		•	0.467	0.167	0.300	14.9	•		0.933	0.833	0.100
12.9		•	0.467	0.250	0.217	15.3	•		1	0.833	0.167
13.0	•		0.533	0.250	0.283	15.9		•	1	0.917	0.083
13.1	•		0.600	0.250	0.350	16.2		•	1	1	0
13.2	•	•	0.667	0.333	0.334						

Abbildung 8.3: *Zweistichproben–Kolmogorov–Smirnov–Test*



Kapitel 9

Kontingenztafeln

Zur Analyse von Abhängigkeiten zweier oder mehrerer Variablen mit einem endlichen diskreten Wertebereich (String-, Integer-Variablen) dienen Kontingenztafeln. Im einfachsten Fall liegen bei zwei betrachteten Variablen sogenannte *Kreuztabellen* vor, im Fall von drei oder mehreren Variablen hat man es mit höherdimensionalen Tabellen zu tun, die in geeigneter Weise für eine zweidimensionale Darstellung aufgebaut werden müssen.

9.1 Kreuztabellen

Zur einfacheren Schreibweise seien die Merkmalsbereiche der beiden betrachteten Variablen X_1 und X_2 ein Anfangsabschnitt der natürlichen Zahlen, also $M_1 = \{1, \ldots, r\}$ und $M_2 = \{1, \ldots, c\}$. Die $r \times c$ möglichen Kombinationen von Wertepaaren des zweidimensionalen Merkmals (X_1, X_2) werden in Matrixform angeordnet und ergeben eine Tabelle, wie sie in Tab.9.1 dargestellt ist. Eine Variable – (meist die erstgenannte) – bezeichnet die Zeilen und wird daher auch häufig *Zeilenvariable* genannt, die andere Variable beschreibt die Spalten und heißt *Spaltenvariable*.

Mit p_{ij} bezeichnet man die Wahrscheinlichkeit, dass gleichzeitig für X_1 der Wert i und für X_2 der Wert j beobachtet wird. Unter der Hypothese der *Unabhängigkeit* von X_1 und X_2 muss für diese p_{ij} gelten

$$\begin{aligned} p_{ij} &= P(X_1 = i \wedge X_2 = j) = P(X_1 = i) \times P(X_2 = j) \\ &= p_{i.}\, p_{.j} \end{aligned} \tag{9.1}$$

(siehe (3.6)), wobei $p_{i.}$ und $p_{.j}$ für die entsprechenden Randwahrscheinlichkeiten stehen (eindimensionale Randverteilungen für X_1 bzw. X_2). Von den n in der zugrundeliegenden Stichprobe betrachteten Fällen werden n_{ij} Objekte mit der Ausprägung i für die Variable X_1 und j für X_2 beobachtet ($i = 1, \ldots, r; j = 1, \ldots, c$). Die Randhäufigkeiten für X_1 bzw. X_2 sind dann

$$n_{i.} = \sum_{j=1}^{c} n_{ij} \quad \text{bzw.} \quad n_{.j} = \sum_{i=1}^{r} n_{ij} \quad ,$$

sodass man als Schätzung für die Wahrscheinlichkeitsverteilungen von X_1 und X_2

$$\hat{p}_{i.} = n_{i.}/n \quad \text{bzw.} \quad \hat{p}_{.j} = n_{.j}/n \tag{9.2}$$

erhält.

Tabelle 9.1: *Schema einer Kreuztabelle*

X_1 \ X_2	1	2		c	
1	p_{11} $\boxed{n_{11}}$ e_{11}	p_{12} $\boxed{n_{12}}$ e_{12}	$\cdots$	p_{1c} $\boxed{n_{1c}}$ e_{1c}	$p_{1.}$ $n_{1.}$ $np_{1.}$
2	p_{21} $\boxed{n_{21}}$ e_{21}	p_{22} $\boxed{n_{22}}$ e_{22}	$\cdots$	p_{2c} $\boxed{n_{2c}}$ e_{2c}	$p_{2.}$ $n_{2.}$ $np_{2.}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
r	p_{r1} $\boxed{n_{r1}}$ e_{r1}	p_{r2} $\boxed{n_{r2}}$ e_{r2}	$\cdots$	p_{rc} $\boxed{n_{rc}}$ e_{rc}	$p_{r.}$ $n_{r.}$ $np_{r.}$
	$p_{.1}$ $n_{.1}$ $np_{.1}$	$p_{.2}$ $n_{.2}$ $np_{.2}$	$\cdots$ $\cdots$	$p_{.c}$ $n_{.c}$ $np_{.c}$	n

9.1.1 Mosaik-Plots

Der Mosaik-Plot ist ein exploratives, graphisches Verfahren zur Visualisierung von Datensätzen mit zwei oder mehreren kategoriellen Variablen (Merkmalen). Abbildung 9.1 zeigt einen Mosaik-Plot mit zwei kategoriellen Merkmalen.

Die einzelnen kategoriellen Variablen werden zunächst in eine Reihenfolge gebracht und danach wird jede Variable abwechselnd der horizontalen bzw. vertikalen Achse zugeordnet. Die Reihenfolge der Variablen spielt eine entscheidende Rolle für das Aussehen des Mosaik-Plots, d.h., eine andere Reihenfolge oder Zuordnung ergibt auch einen anderen Mosaikplot.

Der Mosaik-Plot stellt im Wesentlichen eine flächen-proportionale Visualisierung der beobachteten Häufigkeiten dar. Er besteht aus einzelnen Flächen (Kacheln), die durch wiederholte horizontale und vertikale Teilungen von Rechtecken entstehen. Die Flächen der rechteckigen Kacheln, die für je eine Merkmalkombination stehen, sind somit proportional zur Anzahl der Beobachtungen, die diese Merkmalkombination aufweisen.

9.1.2 χ^2-Test

Unter Zugrundelegung der zuvor erwähnten Unabhängigkeitshypothese gilt für die *erwartete Häufigkeit* von Beobachtungen des Wertepaares (i, j) offensichtlich

$$e_{ij} = \mathrm{E}(N_{ij}) = n\, p_{ij} = n\, p_{i.} p_{.j} \quad .$$

Da die Randwahrscheinlichkeiten nicht bekannt sind, müssen sie gemäß (9.2) geschätzt werden, sodass man zur *geschätzten erwarteten Häufigkeit*

$$\hat{e}_{ij} = n\, \hat{p}_{i.}\, \hat{p}_{.j} = \frac{n_{i.} n_{.j}}{n}$$

kommt. Wegen

$$p_{r.} = 1 - \sum_{i=1}^{r-1} p_{i.} \quad \text{und} \quad p_{.c} = 1 - \sum_{j=1}^{c-1} p_{.j}$$

müssen durch (9.2) nur $r - 1 + c - 1 = r + c - 2$ Parameter geschätzt werden.

Der Vergleich mit den beobachteten Häufigkeiten n_{ij} ist Grundlage für die Entscheidung über die Unabhängigkeit. Es liegt daher das Prinzip des χ^2–Tests (vgl. Abschnitt 6.2) vor. Betrachtet man also die Testgröße

$$t = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(n_{ij} - \hat{e}_{ij})^2}{\hat{e}_{ij}} \quad , \tag{9.3}$$

so ist diese im Falle der Nullhypothese (Unabhängigkeit) asymptotisch χ^2-verteilt mit $rc - (r + c - 2) - 1 = (r - 1)(c - 1)$ Freiheitsgraden.

Große Werte für die Teststatistik (9.3) beruhen auf großen Abweichungen zwischen den beobachteten und erwarteten Häufigkeiten und weisen auf einen Widerspruch zur ursprünglichen Annahme der Unabhängigkeit von X_1 und X_2 hin. Somit erhält man einen Test für die Unabhängigkeit von X_1 und X_2 mit dem (asymptotischen) Signifikanzniveau α durch die Regel

$$t \left\{ \begin{array}{c} > \\ \leq \end{array} \right\} \chi^2_{(r-1)(c-1);1-\alpha} \quad \Rightarrow \quad \left\{ \begin{array}{l} \text{Ablehnung (“abhängig”)} \\ \text{Annahme (“unabhängig”)} \end{array} \right\} \text{ von } H_0 \quad .$$

Da die Verteilung der Teststatistik auf asymptotischen Argumenten beruht, verlangt eine weitgehend anerkannte Faustregel, dass die einzelnen Zellen der Kreuztabelle mindestens *fünf* Beobachtungen *erwarten* lassen sollen.

Beispiel 9.1 Im folgenden soll die Frage untersucht werden, ob zwischen der Haarfarbe und der Augenfarbe ein Zusammenhang besteht (Signifikanzniveau $\alpha = 0.05$).

Der Datensatz `HairEyeColor` ist im R-Paket `vcd` enthalten. Abbildung 9.1 zeigt einen Mosaik-Plot der Daten. Die beobachteten Häufigkeiten können aus Tabelle 9.2 abgelesen werden.

Lösung: Die Aufbereitung des Datensatzes analog der Tabelle 9.1 ergibt die Tabelle 9.2. Als Teststatistik ergibt sich bei $f = (4 - 1) \times (4 - 1) = 9$ Freiheitsgraden

$$t = 138.290 \ ,$$

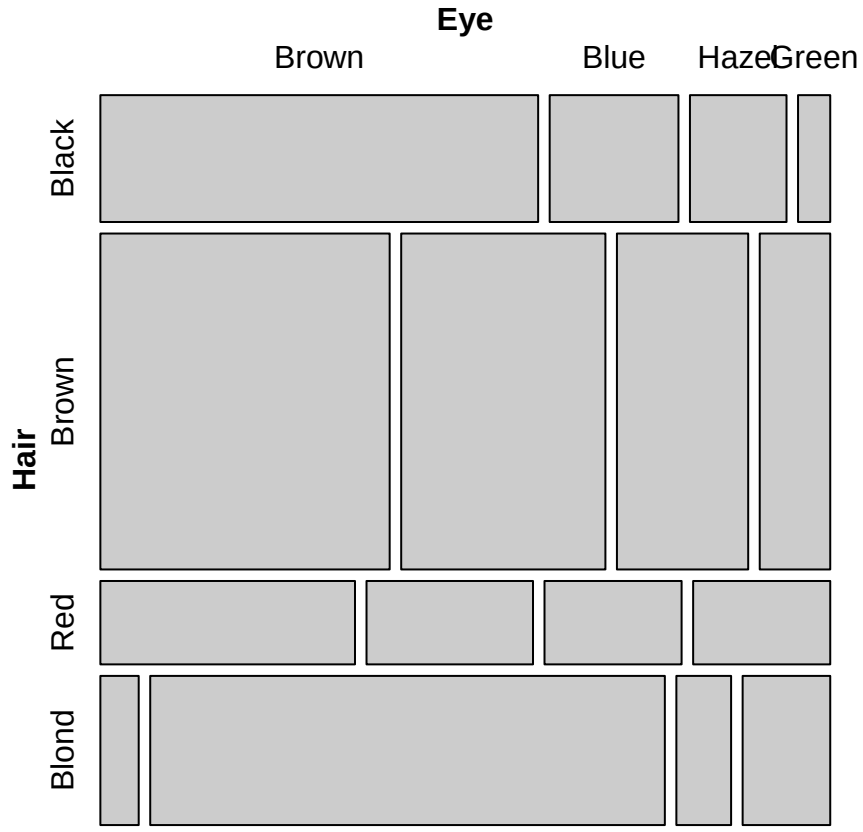
was mit dem kritischen Wert $\chi^2_{9;0.95} = 16.92$ zu vergleichen ist. Da letzterer überschritten wird, ist die Nullhypothese der Unabhängigkeit von “Haarfarbe” und “Augenfarbe” zu verwerfen. Es besteht somit ein *signifikanter Zusammenhang* zwischen diesen beiden Merkmalen, was auch unmittelbar anhand des p–Wertes

$$p = 2.33 \times 10^{-25}$$

abgelesen werden kann.

◇◇◇

Abbildung 9.1: *Mosaik-Plot*



9.1.3 Exakte Tests

Neben dem am häufigsten eingesetzten χ^2–Test gibt es noch exakte Tests, die im Wesentlichen eine Verallgemeinerung des im Folgeabschnitt diskutierten Tests nach Fisher darstellen. Ihr Hauptnachteil ist die aufwendige Bestimmung von kritischen Werten bzw. die Berechnung des empirischen Signifikanzniveaus.

Beispiel 9.2 Im Anschluss an Bsp. 1.2 soll die Frage untersucht werden, ob zwischen der Absicht, nach dem Studium einen selbständigen Beruf auszuüben, und dem Beruf des Vaters ein Zusammenhang besteht (Signifikanzniveau $\alpha = 0.05$).

Lösung: Die Aufbereitung des Datensatzes analog der Tabelle 9.1 ergibt nach Eleminierung der Spezialfälle 17, 26, 42 und 49 die Tabelle 9.3. Als Teststatistik ergibt sich bei $f = (4-1) \times (4-1) = 9$ Freiheitsgraden

$$t = 24.869 \ ,$$

was mit dem kritischen Wert $\chi^2_{9;0.95} = 16.92$ zu vergleichen ist. Da letzterer überschritten wird, ist die Nullhypothese der Unabhängigkeit von "Absicht" und "Beruf/Vater" zu verwerfen. Es besteht somit ein *signifikanter Zusammenhang* zwischen diesen beiden Merkmalen, was auch

Tabelle 9.2: *Kreuztabelle "HairEyeColor"*

Eye color / Hair color	Brown	Blue	Hazel	Green	$n_{i.}$
Black	68 19.346	20 9.421	15 0.228	5 3.817	108
Brown	119 1.521	84 3.800	54 1.831	29 0.119	286
Red	26 0.006	17 2.993	14 0.726	14 5.211	71
Blond	7 34.234	94 49.697	10 4.963	16 0.375	127
$n_{.j}$	220	215	93	64	592

Tabelle 9.3: *Kreuztabelle "Absicht/Beruf-V" zu Bsp. 1.2*

Beruf-V / Absicht	1	2	3	4	$n_{i.}$
1	0 0.125	5 1.500	1 3.750	1 1.625	7
2	0 0.589	4 7.071	22 17.679	7 7.661	33
3	0 0.196	2 2.357	6 5.893	3 2.554	11
4	1 0.089	1 1.071	1 2.679	2 1.161	5
$n_{.j}$	1	12	30	13	56

unmittelbar anhand des p–Wertes

$$p = 0.0031$$

abgelesen werden kann.

◇◇◇

9.2 Vergleich diskreter Merkmale

Formal genauso wie der Test auf Unabhängigkeit in einer Kontingenztafel verläuft der Test nach Gleichheit der Verteilung eines Merkmals X mit c Ausprägungen, das in r Gruppen beobachtet wird. Beispiele für derartige Situationen sind etwa Haltungen zu bestimmten Fragen in unterschiedlichen Teilkollektiven oder Fehleranzahlen bei verschiedenen Produktionen, wobei vorher durch eine geeignete Klassenbildung sicherzustellen ist, dass nur endlich viele Ausprägungen für X betrachtet werden.

Die beobachteten Häufigkeiten der einzelnen Ausprägungen von X in den betrachteten Gruppen werden wie bei den Kreuztabellen in Matrixform zusammengestellt:

Tabelle 9.4: *Vergleich diskreter Merkmale*

Gruppe i	Merkmal X 1	2		c	
1	p_{11} $\boxed{n_{11}}$ e_{11}	p_{12} $\boxed{n_{12}}$ e_{12}	$\cdots$	p_{1c} $\boxed{n_{1c}}$ e_{1c}	n_1
2	p_{21} $\boxed{n_{21}}$ e_{21}	p_{22} $\boxed{n_{22}}$ e_{22}	$\cdots$	p_{2c} $\boxed{n_{2c}}$ e_{2c}	n_2
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
r	p_{r1} $\boxed{n_{r1}}$ e_{r1}	p_{r2} $\boxed{n_{r2}}$ e_{r2}	$\cdots$	p_{rc} $\boxed{n_{rc}}$ e_{rc}	n_r
	p_1 $n_{.1}$ np_1	p_2 $n_{.2}$ np_2	$\cdots$ $\cdots$ $\cdots$	p_c $n_{.c}$ np_c	n

Die Nullhypothese behauptet die *Gleichheit* der Verteilung von X in den r Gruppen, also

$$H_0 : p_{1j} = p_{2j} = \cdots = p_{rj} = p_j \tag{9.4}$$

Als Testgröße verwendet man auch hier die aufsummierten quadrierten und normierten Differenzen zwischen beobachteter und unter der Nullhypothese erwarteter Häufigkeit

$$t = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(n_{ij} - \hat{e}_{ij})^2}{\hat{e}_{ij}} \quad . \tag{9.5}$$

Unter der Nullhypothese gleicher Verteilungen für alle Gruppen gilt für die erwartete Häufigkeit

$$e_{ij} = n_i p_j \quad .$$

Da die p_j aber nicht bekannt sind, werden sie durch

$$\hat{p}_j = \frac{n_{1j} + n_{2j} + \cdots + n_{rj}}{n} = \frac{n_{\cdot j}}{n}$$

geschätzt, sodass bei dieser Fragestellung die *geschätzten* erwarteten Häufigkeiten von der Form

$$\hat{e}_{ij} = \frac{n_i n_{\cdot j}}{n}$$

sind, sich also wieder als Spaltensumme mal Zeilensumme (= Gruppenumfang) dividiert durch Gesamtumfang ergeben.

Im Fall bekannter Wahrscheinlichkeiten p_j sind die einzelnen Zeilensummen in der Teststatistik (9.5) unabhängig asymptotisch χ^2_{c-1}–verteilt, und damit ist die Gesamtsumme t asysmptotisch $\chi^2_{r(c-1)}$–verteilt. Wegen der $c-1$ zu schätzenden Parameter p_j ($j = 1, \ldots, c;\ p_1 + p_2 + \cdots + p_c = 1$!) reduziert sich die Zahl der Freiheitsgrade um diese Anzahl, womit schließlich t asymptotisch $\chi^2_{(r-1)(c-1)}$–verteilt ist.

Nun sind große Werte für t signifikant, sodass man als kritischen Wert wieder $\chi^2_{(r-1)(c-1);1-\alpha}$ erhält und wie in Abschnitt 9.1 verfährt.

Beispiel 9.3
In der nebenstehenden Tabelle findet man die Ergebnisse einer Meinungsbefragung hinsichtlich der Partei– bzw. Gruppenpräferenz. Unterscheiden sich die vier Städte bezüglich ihrer politischen Struktur signifikant? (Signifikanzniveau $\alpha = 0.05$)

Stadt	Partei/Gruppe					
	SP	VP	FP	GA	KP	o.A.
Graz	41	31	26	8	5	39
Linz	54	31	22	11	3	29
Salzburg	19	16	8	12	1	44
Wien	143	57	62	29	5	204

Lösung:
Zunächst muss geklärt werden, wie man die Fälle behandelt, bei denen die Antwort verweigert wurde. Üblicherweise nimmt man sie aus der Analyse heraus, andererseits ergeben unterschiedliche Verweigerungsraten möglicherweise doch Strukturunterschiede im Wahlverhalten.

Lässt man die Antwortverweigerer unberücksichtigt, so ergibt sich die Teststatistik (9.5) bei $f = (4-1) \times (5-1) = 12$ Freiheitsgraden als

$$t = 20.538$$

mit einem p–Wert 0.0576 , sodass die Hypothese identer Verteilungen *nicht verworfen* werden kann.

◇◇◇

9.3 Vierfeldertafel

Im einfachsten Fall einer Kontingenztafel, nämlich der 2×2–Tafel, lässt sich einerseits die Teststatistik (9.3) stark vereinfachen, andererseits aber auch ein exakter Test für die Unabhängigkeit zweier diskreter Merkmale anschaulich beschreiben. Dabei ergibt sich u.a. der bekannte Test nach *Fisher* als Spezialfall.

χ^2–Test:

Für die (geschätzten) Residuen

$$d_{ij} = n_{ij} - \hat{\mathrm{E}}(N_{ij}) = n_{ij} - \frac{n_{i.}n_{.j}}{n}$$

gilt

$$d_{11} = -d_{12} = -d_{21} = d_{22} \quad ,$$

wie sich aus

$$d_{12} = n_{12} - \frac{n_{1.}n_{.2}}{n} = (n_{1.} - n_{11}) - \frac{n_{1.}(n - n_{.1})}{n} = -n_{11} + \frac{n_{1.}n_{.1}}{n} = -d_{11}$$

und der Symmetrie von erster und zweiter Komponente einfach ableiten lässt. Damit ergibt sich die Teststatistik (9.3) als

$$\begin{aligned}
t &= d_{11}^2 \left(\frac{1}{n_{1.}n_{.1}/n} + \frac{1}{n_{1.}n_{.2}/n} + \frac{1}{n_{2.}n_{.1}/n} + \frac{1}{n_{2.}n_{.2}/n} \right) \\
&= \frac{n}{n_{1.}n_{.1}n_{2.}n_{.2}} \frac{\underbrace{(n_{2.}n_{.2} + n_{2.}n_{.1} + n_{1.}n_{.2} + n_{1.}n_{.1})}(n_{11} - \frac{n_{1.}n_{.1}}{n})^2}{n_{2.}\underbrace{(n_{.2} + n_{.1})}_{n} + n_{1.}\underbrace{(n_{.2} + n_{.1})}_{n}} \\
&= \frac{n}{n_{1.}n_{.1}n_{2.}n_{.2}} n^2 (n_{11} - \frac{n_{1.}n_{.1}}{n})^2 \\
&= \frac{n}{n_{1.}n_{.1}n_{2.}n_{.2}} (n\, n_{11} - n_{1.}n_{.1})^2 \\
&= \frac{n}{n_{1.}n_{.1}n_{2.}n_{.2}} (n_{11}n_{11} + n_{12}n_{11} + n_{21}n_{11} + n_{22}n_{11} \\
&\qquad - n_{11}n_{11} - n_{11}n_{21} - n_{12}n_{11} - n_{12}n_{21})^2 \\
&= \frac{n\,(n_{11}n_{22} - n_{12}n_{21})^2}{n_{1.}n_{.1}n_{2.}n_{.2}} \quad . \qquad (9.6)
\end{aligned}$$

An Stelle dieser Größe findet man auch manchmal die nach *Yates* korrigierte Form

$$t_Y = \frac{n(|n_{11}n_{22} - n_{12}n_{21}| - \frac{n}{2})^2}{n_{1.}n_{.1}n_{2.}n_{.2}} \quad . \qquad (9.7)$$

Exakter Test nach Fisher:

Bei gegebenen Randsummen $n_{i.}$ und $n_{.j}$ ist die Vierfeldertafel offensichtlich durch n_{11} *eindeutig* festgelegt. Unter der Unabhängigkeitshypothese ist aber N_{11} nach $H(n, n_{1.}, n_{.1})$, also hypergeometrisch verteilt (Wahrscheinlichkeit, dass bei n Beobachtungen unter den $n_{.1}$ unabhängig, d.h. zufällig der ersten Spalte zugeordneten Fällen genau n_{11} in der ersten Zeile liegen, wenn sich insgesamt $n_{1.}$ in der ersten Zeile befinden). Damit lässt sich ein exakter Test

für die Unabhängigkeit konstruieren, indem zu einem vorgegebenen Signifikanzniveau α die kritischen Werte als $\alpha/2$– und $(1-\alpha/2)$–Quantil der hypergeometrischen Verteilung mit den Parametern n, $n_{1.}$ und $n_{.1}$ gewählt werden, sodass

$$\left.\begin{array}{l} n_{11} \le h_{n,n_{1.},n_{.1};\alpha/2} \quad \text{oder} \quad n_{11} \ge h_{n,n_{1.},n_{.1};1-\alpha/2} \\ h_{n,n_{1.},n_{.1};\alpha/2} < n_{11} < h_{n,n_{1.},n_{.1};1-\alpha/2} \end{array}\right\} \quad \Rightarrow$$

$$\Rightarrow \quad \mathrm{H}_0 \text{ (Unabhängigkeit)} \left\{\begin{array}{l} \text{ablehnen} \\ \text{annehmen .} \end{array}\right. \tag{9.8}$$

Dieser Test ist vor allem dann vorzuziehen, wenn auf Grund eines geringen Datenumfangs die Asymptotik des χ^2–Tests noch nicht greift.

Beispiel 9.4 Im Anschluss an Bsp. 1.2 soll die Frage untersucht werden, ob ein Zusammenhang der beruflichen Tätigkeit bei Vater und Mutter besteht, wenn man nur zwischen "selbständig" und "nicht selbständig" unterscheidet (Sicherheit $1-\alpha = 0.95$).

Lösung: Die Anwendung des χ^2–Tests liefert bei einem Freiheitsgrad als Teststatistik

$$t = 1.364$$

und den kritischen Wert $\chi^2_{1;0.95} = 3.84$. Damit ist die Unabhängigkeitshypothese anzunehmen, es besteht also *kein signifikanter* Einwand *gegen* die Unabhängigkeit der Berufstätigkeit von Mutter und Vater. Der p–Wert ergibt sich als

$$p = 0.2429$$

und weist auch auf die Annahme von H_0.

Tabelle 9.5: *Kreuztabelle "Beruf-M/Beruf-V" zu Bsp. 1.2*

Beruf-M \ Beruf-V	1	2	$n_{i.}$
1	2 1	3 4	5
2	10 11	45 44	55
$n_{.j}$	12	48	60

Bei Anwendung des exakten Tests nach Fisher benötigt man das obere und untere 0.025–Quantil der $H(60,5,12)$–Verteilung. Hier findet man nur das obere als

$$h_{60,5,12;0.975} = 3$$

und da

$$n_{11} < 3 \ ,$$

ist auch hier die Nullhypothese anzunehmen.

◇ ◇ ◇

Literatur

Afifi, A.A., und S.P. Azen (1979): *Statistical Analysis. A Computer Oriented Approach*, Acad. Press, New York.

Betha, R.M. und R.R. Rhinehart (1991): *Applied Engineering Statistics.* (Statistics: Textbooks and Monographs Series, 121) Marcel Dekker, New York/Basel/Hong Kong.

Bläsing, J.P. (1989): *Statistische Qualitätskontrolle.* Verlag: gfmt – Gesellschaft für Management und Technologie AG, St. Gallen.

Bleymüller, Gehlert und Gütlicher: *Statistik für Wirtschaftswissenschaftler*, Verlag Vahlen.

Bosch, K. (1976): *Elementare Einführung in die Wahrscheinlichkeitsrechnung*, rororo, Verlag Vieweg, Wiesbaden.

Bosch, K. (1976): *Angewandte Mathematische Statistik*, rororo, Verlag Vieweg, Wiesbaden.

Eßl, A., (1987): *Statistische Methoden in der Tierproduktion*, Österreichischer Agrarverlag, Wien.

Hartung, J., B. Elpelt und H.-K. Klösener (2002): *Statistik/ Lehrbuch und Handbuch der angewandten Statistik*, 13. Aufl., Oldenbourg Verlag, München.

ISO 2859: *Sampling Procedures for Inspection by Attributes*, Teil 0 bis Teil 3, International Organization for Standardization, Genf, 1988.

ISO 3951: *Sampling Procedures and Charts for Inspection by Variables for Percent Nonconforming.* International Organization for Standardization, Genf, 1989.

Kramer, C,Y. (1956): Extension of multiple range tests to group means with unequal number of replications, *Biometrics*, **12**, 307–310.

Kreyszig, E. (1972): *Statistische Methoden und ihre Anwendungen*, Vandenhœk, Göttingen.

Läuter, H. und R. Pincus (1989): *Mathematisch–statistische Datenanalyse*, Oldenbourg–Verlag, München.

Pearson, E.S. und H.O. Hartley (1958): *Biometrika Tables for Statisticians*, Vol. 1, Cambridge Univ. Press, Cambridge.

Pearson, E.S. und H.O. Hartley (1972): *Biometrika Tables for Statisticians*, Vol. 2, Cambridge Univ. Press, Cambridge.

Pfanzagl, J. (1974): *Allgemeine Methodenlehre der Statistik II*, De Gruyter, Berlin.

Tukey, J.W. (1953): Multiple comparisons, *J. Amer. Statist. Assoc.*, **48**, 624–625.

Tukey, J.W. (1977): *Exploratory Data Analysis.* Addison-Wesley, Reading, Mass.

Viertl, R. (1990): *Einführung in die Stochastik mit Elementen der Bayes–Statistik und Ansätzen für die Analyse unscharfer Daten*, Springer Verlag, Wien.

Anhang A

Tabellen

Tabelle A.1: $N(0,1)$–*Verteilung;*

$$\gamma = \Phi(z_\gamma) = \Pr(Z \le z_\gamma)$$

$\mathbf{z_\gamma}$	**0.00**	**0.01**	**0.02**	**0.03**	**0.04**	**0.05**	**0.06**	**0.07**	**0.08**	**0.09**
.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7852
.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9706
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9761	0.9767
2.0	0.9772	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812	0.9817
2.1	0.9821	0.9826	0.9830	0.9834	0.9838	0.9842	0.9846	0.9850	0.9854	0.9857
2.2	0.9861	0.9864	0.9868	0.9871	0.9875	0.9878	0.9881	0.9884	0.9887	0.9890
2.3	0.9893	0.9896	0.9898	0.9901	0.9904	0.9906	0.9909	0.9911	0.9913	0.9916
2.4	0.9918	0.9920	0.9922	0.9925	0.9927	0.9929	0.9931	0.9932	0.9934	0.9936
2.5	0.9938	0.9940	0.9941	0.9943	0.9945	0.9946	0.9948	0.9949	0.9951	0.9952
2.6	0.9953	0.9955	0.9956	0.9957	0.9959	0.9960	0.9961	0.9962	0.9963	0.9964
2.7	0.9965	0.9966	0.9967	0.9968	0.9969	0.9970	0.9971	0.9972	0.9973	0.9974
2.8	0.9974	0.9975	0.9976	0.9977	0.9977	0.9978	0.9979	0.9979	0.9980	0.9981
2.9	0.9981	0.9982	0.9982	0.9983	0.9984	0.9984	0.9985	0.9985	0.9986	0.9986
3.0	0.9987	0.9987	0.9987	0.9988	0.9988	0.9989	0.9989	0.9989	0.9990	0.9990
3.1	0.9990	0.9991	0.9991	0.9991	0.9992	0.9992	0.9992	0.9992	0.9993	0.9993
3.2	0.9993	0.9993	0.9994	0.9994	0.9994	0.9994	0.9994	0.9995	0.9995	0.9995
3.3	0.9995	0.9995	0.9995	0.9996	0.9996	0.9996	0.9996	0.9996	0.9996	0.9997
3.4	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9998

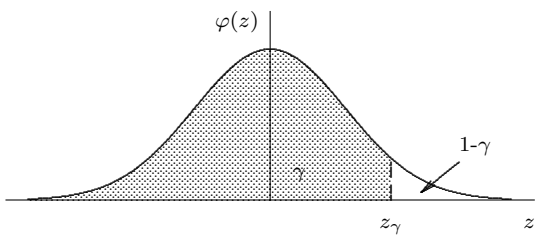


Tabelle A.2: *t–Verteilung;* γ–Quantile $t_{f;\gamma}$

$$\gamma = \Pr(T \leq t_{f;\gamma})$$

FG f	γ					
	.750	**.900**	**.950**	**.975**	**.990**	**.995**
1	1.000	3.078	6.314	12.706	31.824	63.659
2	0.816	1.886	2.920	4.303	6.965	9.925
3	0.765	1.638	2.353	3.182	4.541	5.841
4	0.741	1.533	2.132	2.776	3.747	4.604
5	0.727	1.476	2.015	2.571	3.365	4.032
6	0.718	1.440	1.943	2.447	3.143	3.707
7	0.711	1.415	1.895	2.365	2.998	3.499
8	0.706	1.397	1.860	2.306	2.896	3.355
9	0.703	1.383	1.833	2.262	2.821	3.250
10	0.700	1.372	1.812	2.228	2.764	3.169
11	0.697	1.363	1.796	2.201	2.718	3.106
12	0.695	1.356	1.782	2.179	2.681	3.055
13	0.694	1.350	1.771	2.160	2.650	3.012
14	0.692	1.345	1.761	2.145	2.624	2.977
15	0.691	1.341	1.753	2.131	2.602	2.947
16	0.690	1.337	1.746	2.120	2.583	2.921
17	0.689	1.333	1.740	2.110	2.567	2.898
18	0.688	1.330	1.734	2.101	2.552	2.878
19	0.688	1.328	1.729	2.093	2.539	2.861
20	0.687	1.325	1.725	2.086	2.528	2.845
21	0.686	1.323	1.721	2.080	2.518	2.831
22	0.686	1.321	1.717	2.074	2.508	2.819
23	0.685	1.319	1.714	2.069	2.500	2.807
24	0.685	1.318	1.711	2.064	2.492	2.797
25	0.684	1.316	1.708	2.060	2.485	2.787
26	0.684	1.315	1.706	2.056	2.479	2.779
27	0.684	1.314	1.703	2.052	2.473	2.771
28	0.683	1.313	1.701	2.048	2.467	2.763
29	0.683	1.311	1.699	2.045	2.462	2.756
30	0.683	1.310	1.697	2.042	2.457	2.750
35	0.682	1.306	1.690	2.030	2.438	2.724
40	0.681	1.303	1.684	2.021	2.423	2.704
45	0.680	1.301	1.679	2.014	2.412	2.690
50	0.679	1.299	1.676	2.009	2.403	2.678
55	0.679	1.297	1.673	2.004	2.396	2.668
60	0.679	1.296	1.671	2.000	2.390	2.660
65	0.678	1.295	1.669	1.997	2.385	2.654
70	0.678	1.294	1.667	1.994	2.381	2.648
75	0.678	1.293	1.665	1.992	2.377	2.643
80	0.678	1.292	1.664	1.990	2.374	2.639
85	0.677	1.292	1.663	1.988	2.371	2.635
90	0.677	1.291	1.662	1.987	2.368	2.632
95	0.677	1.291	1.661	1.985	2.366	2.629
100	0.677	1.290	1.660	1.984	2.364	2.626

Tabelle A.3: χ^2–*Verteilung;* γ–Quantile $\chi_{f;\gamma}$

$$\gamma = \mathrm{P}(X \leq \chi_{f;\gamma})$$

FG	γ									
f	**.005**	**.01**	**.025**	**.05**	**.1**	**.9**	**.95**	**.975**	**.99**	**.995**
1	.000	.000	.001	.004	.016	2.706	3.841	5.024	6.635	7.879
2	.010	.020	.051	.103	.211	4.605	5.991	7.378	9.210	10.597
3	.072	.115	.216	.352	.584	6.251	7.815	9.348	11.345	12.838
4	.207	.297	.484	.711	1.064	7.779	9.488	11.143	13.277	14.860
5	.412	.554	.831	1.145	1.610	9.236	11.070	12.832	15.086	16.750
6	.676	.872	1.237	1.635	2.204	10.645	12.592	14.449	16.812	18.548
7	.989	1.239	1.690	2.167	2.833	12.017	14.067	16.013	18.475	20.278
8	1.344	1.647	2.180	2.733	3.490	13.362	15.507	17.535	20.090	21.955
9	1.735	2.088	2.700	3.325	4.168	14.684	16.919	19.023	21.666	23.589
10	2.156	2.558	3.247	3.940	4.865	15.987	18.307	20.483	23.209	25.188
11	2.603	3.053	3.816	4.575	5.578	17.275	19.675	21.920	24.725	26.757
12	3.074	3.571	4.404	5.226	6.304	18.549	21.026	23.337	26.217	28.299
13	3.565	4.107	5.009	5.892	7.042	19.812	22.362	24.736	27.688	29.819
14	4.075	4.660	5.629	6.571	7.790	21.064	23.685	26.119	29.141	31.319
15	4.601	5.229	6.262	7.261	8.547	22.307	24.996	27.488	30.578	32.801
16	5.142	5.812	6.908	7.962	9.312	23.542	26.296	28.845	32.000	34.267
17	5.697	6.408	7.564	8.672	10.085	24.769	27.587	30.191	33.409	35.718
18	6.265	7.015	8.231	9.390	10.865	25.989	28.869	31.526	34.805	37.156
19	6.844	7.633	8.907	10.117	11.651	27.204	30.144	32.852	36.191	38.582
20	7.434	8.260	9.591	10.851	12.443	28.412	31.410	34.170	37.566	39.997
21	8.034	8.897	10.283	11.591	13.240	29.615	32.671	35.479	38.932	41.401
22	8.643	9.543	10.982	12.338	14.041	30.813	33.924	36.781	40.289	42.796
23	9.261	10.196	11.689	13.091	14.848	32.007	35.172	38.076	41.638	44.181
24	9.886	10.856	12.401	13.848	15.659	33.196	36.415	39.364	42.980	45.559
25	10.520	11.524	13.120	14.611	16.473	34.382	37.652	40.646	44.314	46.928
30	13.787	14.953	16.791	18.493	20.599	40.256	43.773	46.979	50.892	53.672
35	17.192	18.509	20.569	22.465	24.797	46.059	49.802	53.203	57.342	60.275
40	20.707	22.164	24.433	26.509	29.051	51.805	55.758	59.342	63.691	66.766
45	24.311	25.901	28.366	30.612	33.350	57.505	61.656	65.410	69.957	73.166
50	27.991	29.707	32.357	34.764	37.689	63.167	67.505	71.420	76.154	79.490
55	31.735	33.570	36.398	38.958	42.060	68.796	73.311	77.380	82.292	85.749
60	35.534	37.485	40.482	43.188	46.459	74.397	79.082	83.298	88.379	91.952
65	39.383	41.444	44.603	47.450	50.883	79.973	84.821	89.177	94.422	98.105
70	43.275	45.442	48.758	51.739	55.329	85.527	90.531	95.023	100.425	104.215
75	47.206	49.475	52.942	56.054	59.795	91.061	96.217	100.839	106.393	110.286
80	51.172	53.540	57.153	60.391	64.278	96.578	101.879	106.629	112.329	116.321
85	55.170	57.634	61.389	64.749	68.777	102.079	107.522	112.393	118.236	122.325
90	59.196	61.754	65.647	69.126	73.291	107.565	113.145	118.136	124.116	128.299
95	63.250	65.898	69.925	73.520	77.818	113.038	118.752	123.858	129.973	134.247
100	67.328	70.065	74.222	77.929	82.358	118.498	124.342	129.561	135.807	140.169

Tabelle A.4: F_{f_1,f_2}–*Verteilung;* 0.9–Quantile $F_{f_1,f_2;0.9}$

$$P(F \leq F_{f_1,f_2;0,9}) = 0.9; \quad F_{f_z,f_n;0,1} = 1/F_{f_n,f_z;0,9}$$

FG f_1 \ f_2	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
1	39.863	8.526	5.538	4.545	4.060	3.776	3.589	3.458	3.360
2	49.500	9.000	5.462	4.325	3.780	3.463	3.257	3.113	3.006
3	53.593	9.162	5.391	4.191	3.619	3.289	3.074	2.924	2.813
4	55.833	9.243	5.343	4.107	3.520	3.181	2.961	2.806	2.693
5	57.240	9.293	5.309	4.051	3.453	3.108	2.883	2.726	2.611
6	58.204	9.326	5.285	4.010	3.405	3.055	2.827	2.668	2.551
7	58.906	9.349	5.266	3.979	3.368	3.014	2.785	2.624	2.505
8	59.439	9.367	5.252	3.955	3.339	2.983	2.752	2.589	2.469
9	59.858	9.380	5.240	3.936	3.316	2.958	2.725	2.561	2.440
10	60.195	9.392	5.230	3.920	3.297	2.937	2.703	2.538	2.416
12	60.706	9.408	5.216	3.896	3.268	2.905	2.668	2.502	2.379
15	61.220	9.425	5.200	3.870	3.238	2.871	2.632	2.464	2.340
20	61.740	9.441	5.184	3.844	3.207	2.836	2.595	2.425	2.298
30	62.265	9.458	5.168	3.817	3.174	2.800	2.555	2.383	2.255
60	62.794	9.475	5.151	3.790	3.140	2.762	2.514	2.339	2.208
120	63.061	9.483	5.143	3.775	3.123	2.742	2.493	2.316	2.184
200	63.168	9.486	5.139	3.769	3.116	i 2.734	2.484	2.307	2.174
500	63.265	9.489	5.136	3.764	3.109	2.727	2.476	2.298	2.165

FG f_1 \ f_2	**10**	**12**	**15**	**20**	**30**	**60**	**120**	**200**	**500**
1	3.285	3.177	3.073	2.975	2.881	2.791	2.748	2.731	2.716
2	2.924	2.807	2.695	2.589	2.489	2.393	2.347	2.329	2.313
3	2.728	2.606	2.490	2.380	2.276	2.177	2.130	2.111	2.095
4	2.605	2.480	2.361	2.249	2.142	2.041	1.992	1.973	1.956
5	2.522	2.394	2.273	2.158	2.049	1.946	1.896	1.876	1.859
6	2.461	2.331	2.208	2.091	1.980	1.875	1.824	1.804	1.786
7	2.414	2.283	2.158	2.040	1.927	1.819	1.767	1.747	1.729
8	2.377	2.245	2.119	1.999	1.884	1.775	1.722	1.701	1.683
9	2.347	2.214	2.086	1.965	1.849	1.738	1.684	1.663	1.644
10	2.323	2.188	2.059	1.937	1.819	1.707	1.652	1.631	1.612
12	2.284	2.147	2.017	1.892	1.773	1.657	1.601	1.579	1.559
15	2.244	2.105	1.972	1.845	1.722	1.603	1.545	1.522	1.501
20	2.201	2.060	1.924	1.794	1.667	1.543	1.482	1.458	1.435
30	2.155	2.011	1.873	1.738	1.606	1.476	1.409	1.383	1.358
60	2.107	1.960	1.817	1.677	1.538	1.395	1.320	1.289	1.260
120	2.082	1.932	1.787	1.643	1.499	1.348	1.265	1.228	1.194
200	2.071	1.921	1.774	1.629	1.482	1.326	1.239	1.199	1.160
500	2.062	1.911	1.763	1.616	1.467	1.306	1.212	1.168	1.122

Tabelle A.4: F_{f_1,f_2}–*Verteilung;* 0.95–Quantile $F_{f_1,f_2;0.95}$

$$P(F \leq F_{f_1,f_2;0,95}) = 0.95; \quad F_{f_z,f_n;0,05} = 1/F_{f_n,f_z;0,95}$$

FG f_1 \ f_2	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
1	161.449	18.513	10.128	7.709	6.608	5.987	5.591	5.318	5.117
2	199.501	19.000	9.552	6.944	5.786	5.143	4.737	4.459	4.256
3	215.708	19.164	9.277	6.591	5.409	4.757	4.347	4.066	3.863
4	224.583	19.247	9.117	6.388	5.192	4.534	4.120	3.838	3.633
5	230.162	19.296	9.013	6.256	5.050	4.387	3.972	3.687	3.482
6	233.987	19.330	8.941	6.163	4.950	4.284	3.866	3.581	3.374
7	236.769	19.353	8.887	6.094	4.876	4.207	3.787	3.500	3.293
8	238.883	19.371	8.845	6.041	4.818	4.147	3.726	3.438	3.230
9	240.544	19.385	8.812	5.999	4.772	4.099	3.677	3.388	3.179
10	241.882	19.396	8.786	5.964	4.735	4.060	3.637	3.347	3.137
12	243.906	19.413	8.745	5.912	4.678	4.000	3.575	3.284	3.073
15	245.950	19.429	8.703	5.858	4.619	3.938	3.511	3.218	3.006
20	248.014	19.446	8.660	5.803	4.558	3.874	3.445	3.150	2.936
30	250.096	19.462	8.617	5.746	4.496	3.808	3.376	3.079	2.864
60	252.196	19.479	8.572	5.688	4.431	3.740	3.304	3.005	2.787
120	253.253	19.487	8.549	5.658	4.398	3.705	3.267	2.967	2.748
200	253.678	19.491	8.540	5.646	4.385	3.690	3.252	2.951	2.731
500	254.060	19.494	8.532	5.635	4.373	3.678	3.239	2.937	2.717

FG f_1 \ f_2	**10**	**12**	**15**	**20**	**30**	**60**	**120**	**200**	**500**
1	4.965	4.747	4.543	4.351	4.171	4.001	3.920	3.888	3.860
2	4.103	3.885	3.682	3.493	3.316	3.150	3.072	3.041	3.014
3	3.708	3.490	3.287	3.098	2.922	2.758	2.680	2.650	2.623
4	3.478	3.259	3.056	2.866	2.690	2.525	2.447	2.417	2.390
5	3.326	3.106	2.901	2.711	2.534	2.368	2.290	2.259	2.232
6	3.217	2.996	2.790	2.599	2.421	2.254	2.175	2.144	2.117
7	3.135	2.913	2.707	2.514	2.334	2.167	2.087	2.056	2.028
8	3.072	2.849	2.641	2.447	2.266	2.097	2.016	1.985	1.957
9	3.020	2.796	2.588	2.393	2.211	2.040	1.959	1.927	1.899
10	2.978	2.753	2.544	2.348	2.165	1.993	1.910	1.878	1.850
12	2.913	2.687	2.475	2.278	2.092	1.917	1.834	1.801	1.772
15	2.845	2.617	2.403	2.203	2.015	1.836	1.750	1.717	1.686
20	2.774	2.544	2.328	2.124	1.932	1.748	1.659	1.623	1.592
30	2.700	2.466	2.247	2.039	1.841	1.649	1.554	1.516	1.482
60	2.621	2.384	2.160	1.946	1.740	1.534	1.429	1.386	1.345
120	2.580	2.341	2.114	1.896	1.683	1.467	1.352	1.302	1.255
200	2.563	2.323	2.095	1.875	1.660	1.438	1.316	1.263	1.210
500	2.548	2.307	2.078	1.856	1.637	1.409	1.280	1.221	1.159

Tabelle A.4: F_{f_1,f_2}–*Verteilung;* 0.975–Quantile $F_{f_1,f_2;0.975}$

$$P(F \leq F_{f_1,f_2;0,975}) = 0.975; \quad F_{f_z,f_n;0,025} = 1/F_{f_n,f_z;0,975}$$

FG f_1 \ f_2	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
1	647.789	38.506	17.443	12.218	10.007	8.813	8.073	7.571	7.209
2	799.500	39.000	16.044	10.649	8.433	7.260	6.542	6.059	5.715
3	864.163	39.166	15.439	9.979	7.764	6.599	5.890	5.416	5.078
4	899.584	39.248	15.101	9.604	7.388	6.227	5.523	5.053	4.718
5	921.811	39.298	14.885	9.364	7.146	5.988	5.285	4.817	4.484
6	937.111	39.331	14.735	9.197	6.978	5.820	5.119	4.652	4.320
7	948.217	39.355	14.624	9.074	6.853	5.695	4.995	4.529	4.197
8	956.656	39.373	14.540	8.980	6.757	5.600	4.899	4.433	4.102
9	963.217	39.387	14.473	8.905	6.681	5.523	4.823	4.357	4.026
10	968.628	39.398	14.419	8.844	6.619	5.461	4.761	4.295	3.964
12	976.708	39.415	14.337	8.751	6.525	5.366	4.666	4.200	3.868
15	984.867	39.431	14.253	8.657	6.428	5.269	4.568	4.101	3.769
20	993.103	39.448	14.167	8.560	6.329	5.168	4.467	3.999	3.667
30	1001.415	39.466	14.080	8.461	6.227	5.065	4.362	3.894	3.560
60	1009.800	39.481	13.992	8.360	6.123	4.959	4.254	3.784	3.449
120	1014.020	39.490	13.947	8.309	6.069	4.904	4.199	3.728	3.392
200	1015.713	39.493	13.929	8.289	6.048	4.882	4.176	3.705	3.368
500	1017.254	39.496	13.913	8.270	6.028	4.862	4.156	3.684	3.347

FG f_1 \ f_2	**10**	**12**	**15**	**20**	**30**	**60**	**120**	**200**	**500**
1	6.937	6.554	6.200	5.871	5.568	5.286	5.152	5.100	5.054
2	5.456	5.096	4.765	4.461	4.182	3.925	3.805	3.758	3.716
3	4.826	4.474	4.153	3.859	3.589	3.343	3.227	3.182	3.142
4	4.468	4.121	3.804	3.515	3.250	3.008	2.894	2.850	2.811
5	4.236	3.891	3.576	3.289	3.026	2.786	2.674	2.630	2.592
6	4.072	3.728	3.415	3.128	2.867	2.627	2.515	2.472	2.434
7	3.950	3.607	3.293	3.007	2.746	2.507	2.395	2.351	2.313
8	3.855	3.512	3.199	2.913	2.651	2.412	2.299	2.256	2.217
9	3.779	3.436	3.123	2.837	2.575	2.334	2.222	2.178	2.139
10	3.717	3.374	3.060	2.774	2.511	2.270	2.157	2.113	2.074
12	3.621	3.277	2.963	2.676	2.412	2.169	2.055	2.010	1.971
15	3.522	3.177	2.862	2.573	2.307	2.061	1.945	1.900	1.859
20	3.419	3.073	2.756	2.464	2.195	1.944	1.825	1.778	1.736
30	3.311	2.963	2.644	2.349	2.074	1.815	1.690	1.640	1.596
60	3.198	2.848	2.524	2.223	1.940	1.667	1.530	1.474	1.423
120	3.140	2.787	2.461	2.156	1.866	1.581	1.433	1.370	1.311
200	3.116	2.763	2.435	2.128	1.835	1.543	1.388	1.320	1.254
500	3.094	2.740	2.411	2.103	1.806	1.507	1.343	1.269	1.192

Tabelle A.4: F_{f_1,f_2}–*Verteilung;* 0.99–Quantile $F_{f_1,f_2;0.99}$

$$P(F \leq F_{f_1,f_2;0,99}) = 0.99; \quad F_{f_z,f_n;0,01} = 1/F_{f_n,f_z;0,99}$$

FG f_1 \ f_2	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
1	4052	98.505	34.116	21.198	16.258	13.745	12.246	11.259	10.561
2	4998	99.002	30.817	18.000	13.274	10.925	9.546	8.649	8.022
3	5402	99.169	29.457	16.694	12.060	9.779	8.451	7.591	6.992
4	5623	99.252	28.710	15.977	11.392	9.148	7.847	7.006	6.422
5	5763	99.300	28.237	15.522	10.967	8.746	7.460	6.632	6.057
6	5858	99.335	27.911	15.207	10.672	8.466	7.191	6.371	5.802
7	5927	99.359	27.672	14.976	10.455	8.260	6.993	6.178	5.613
8	5980	99.376	27.489	14.799	10.289	8.101	6.840	6.029	5.467
9	6021	99.389	27.347	14.659	10.157	7.976	6.719	5.911	5.351
10	6055	99.400	27.229	14.546	10.051	7.874	6.620	5.814	5.257
12	6105	99.416	27.052	14.374	9.888	7.718	6.469	5.667	5.111
15	6156	99.434	26.872	14.198	9.722	7.559	6.314	5.515	4.962
20	6208	99.452	26.690	14.020	9.552	7.396	6.155	5.359	4.808
30	6259	99.468	26.506	13.838	9.379	7.228	5.992	5.198	4.649
60	6312	99.484	26.316	13.652	9.202	7.056	5.824	5.032	4.483
120	6338	99.491	26.221	13.558	9.111	6.969	5.737	4.946	4.398
200	6349	99.495	26.183	13.520	9.075	6.934	5.702	4.911	4.363
500	6358	99.499	26.148	13.486	9.042	6.902	5.671	4.880	4.332

FG f_1 \ f_2	**10**	**12**	**15**	**20**	**30**	**60**	**120**	**200**	**500**
1	10.044	9.330	8.683	8.096	7.562	7.077	6.851	6.763	6.686
2	7.559	6.927	6.359	5.849	5.390	4.978	4.787	4.713	4.648
3	6.552	5.953	5.417	4.938	4.510	4.126	3.949	3.881	3.821
4	5.994	5.412	4.893	4.431	4.018	3.649	3.480	3.414	3.357
5	5.636	5.064	4.556	4.103	3.699	3.339	3.174	3.110	3.054
6	5.386	4.821	4.318	3.871	3.473	3.119	2.956	2.893	2.838
7	5.200	4.640	4.142	3.699	3.304	2.953	2.792	2.730	2.675
8	5.057	4.499	4.004	3.564	3.173	2.823	2.663	2.601	2.547
9	4.942	4.388	3.895	3.457	3.067	2.718	2.559	2.497	2.443
10	4.849	4.296	3.805	3.368	2.979	2.632	2.472	2.411	2.356
12	4.706	4.155	3.666	3.231	2.843	2.496	2.336	2.275	2.220
15	4.558	4.010	3.522	3.088	2.700	2.352	2.192	2.129	2.075
20	4.405	3.858	3.372	2.938	2.549	2.198	2.035	1.971	1.915
30	4.247	3.701	3.214	2.778	2.386	2.028	1.860	1.794	1.735
60	4.082	3.535	3.047	2.608	2.208	1.836	1.656	1.583	1.517
120	3.996	3.449	2.959	2.517	2.111	1.726	1.533	1.453	1.377
200	3.962	3.414	2.923	2.479	2.070	1.678	1.477	1.391	1.308
500	3.930	3.382	2.891	2.445	2.032	1.633	1.421	1.328	1.232

Tabelle A.4: F_{f_1,f_2}–*Verteilung;* 0.995–Quantile $F_{f_1,f_2;0.995}$

$$P(F \leq F_{f_1,f_2;0,995}) = 0.995; \quad F_{f_z,f_n;0,005} = 1/F_{f_n,f_z;0,995}$$

FG f_1 \ f_2	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
1	16205	198.5	55.553	31.333	22.785	18.635	16.235	14.688	13.614
2	19991	199.0	49.803	26.284	18.314	14.544	12.404	11.042	10.107
3	21606	199.1	47.473	24.259	16.530	12.916	10.882	9.596	8.717
4	22491	199.2	46.196	23.157	15.556	12.027	10.050	8.805	7.956
5	23046	199.3	45.394	22.456	14.940	11.463	9.522	8.301	7.471
6	23428	199.3	44.838	21.975	14.513	11.073	9.155	7.952	7.134
7	23705	199.3	44.436	21.622	14.200	10.786	8.885	7.694	6.885
8	23915	199.3	44.131	21.352	13.961	10.565	8.678	7.496	6.693
9	24081	199.3	43.882	21.139	13.772	10.391	8.514	7.339	6.541
10	24215	199.4	43.692	20.967	13.618	10.250	8.380	7.211	6.417
12	24417	199.4	43.388	20.705	13.384	10.034	8.176	7.015	6.227
15	24620	199.4	43.085	20.438	13.146	9.814	7.967	6.814	6.032
20	24826	199.4	42.777	20.167	12.903	9.588	7.754	6.608	5.832
30	25034	199.4	42.467	19.891	12.656	9.358	7.534	6.396	5.625
60	25243	199.4	42.149	19.611	12.402	9.122	7.309	6.177	5.410
120	25348	199.4	41.989	19.468	12.274	9.001	7.193	6.065	5.300
200	25391	199.4	41.925	19.411	12.222	8.952	7.147	6.019	5.255
500	25429	199.4	41.867	19.359	12.175	8.908	7.104	5.978	5.215

FG f_1 \ f_2	**10**	**12**	**15**	**20**	**30**	**60**	**120**	**200**	**500**
1	12.826	11.754	10.798	9.944	9.180	8.495	8.179	8.057	7.950
2	9.426	8.510	7.701	6.986	6.355	5.795	5.539	5.441	5.355
3	8.081	7.226	6.476	5.818	5.239	4.729	4.497	4.408	4.330
4	7.343	6.521	5.803	5.174	4.623	4.140	3.921	3.837	3.763
5	6.872	6.071	5.372	4.762	4.228	3.760	3.548	3.467	3.396
6	6.545	5.757	5.071	4.472	3.949	3.492	3.285	3.206	3.137
7	6.303	5.525	4.847	4.257	3.742	3.291	3.087	3.010	2.941
8	6.116	5.345	4.675	4.090	3.580	3.134	2.933	2.856	2.789
9	5.968	5.202	4.536	3.956	3.450	3.008	2.808	2.732	2.665
10	5.847	5.086	4.424	3.847	3.344	2.904	2.705	2.629	2.562
12	5.661	4.906	4.250	3.678	3.179	2.742	2.544	2.468	2.402
15	5.471	4.721	4.070	3.502	3.006	2.570	2.373	2.297	2.230
20	5.274	4.530	3.883	3.318	2.823	2.387	2.188	2.112	2.044
30	5.071	4.331	3.687	3.123	2.628	2.187	1.984	1.905	1.835
60	4.859	4.123	3.480	2.916	2.415	1.962	1.747	1.661	1.584
120	4.750	4.015	3.372	2.806	2.300	1.834	1.605	1.512	1.425
200	4.706	3.971	3.328	2.760	2.251	1.779	1.541	1.442	1.346
500	4.666	3.931	3.287	2.719	2.207	1.726	1.478	1.369	1.260

Tabelle A.5: *Wilcoxon–Test;* γ–Quantile $w_{n_1,n_2;\gamma}$
(nach Pearson & Hartley, 1976, und Hartung, 1991)

1. Zeile: $\gamma = 0.025$; 2. Zeile: $\gamma = 0.05$; 3. Zeile: $\gamma = 0.1$

n_1 \ n_2	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**
	3	3	3	3	3	3	4	4	4	4	5	5	5	5
2	3	3	3	4	4	4	5	5	5	5	6	6	7	7
	3	4	4	5	5	5	6	6	7	7	8	8	8	9
	6	6	6	7	8	8	9	9	10	10	11	11	12	12
3	6	7	7	8	9	9	10	11	11	12	12	13	14	14
	7	8	8	9	10	11	12	12	13	14	15	16	17	17
	10	10	11	12	13	14	15	15	16	17	18	19	20	21
4	10	11	12	13	14	15	16	17	18	19	20	21	22	23
	11	12	14	15	16	17	18	20	21	22	23	24	26	27
	15	16	17	18	19	21	22	23	24	25	27	28	29	30
5	16	17	18	20	21	22	24	25	27	28	29	31	32	34
	17	18	20	21	23	24	26	28	29	31	33	34	36	38
	21	23	24	25	27	28	30	32	33	35	36	38	39	41
6	22	24	25	27	29	30	32	34	36	38	39	41	43	45
	23	25	27	29	31	33	35	37	39	41	43	45	47	49
	28	30	32	34	35	37	39	41	43	45	47	49	51	53
7	29	31	33	35	37	40	42	44	46	48	50	53	55	57
	30	33	35	37	40	42	45	47	50	52	55	57	60	62
	37	39	41	43	45	47	50	52	54	56	59	61	63	66
8	38	40	42	45	47	50	52	55	57	60	63	65	68	70
	39	42	44	47	50	53	56	59	61	64	67	70	73	76
	46	48	50	53	56	58	61	63	66	69	72	74	77	80
9	47	50	52	55	58	61	64	67	70	73	76	79	82	85
	48	51	55	58	61	64	68	71	74	77	81	84	87	91
	56	59	61	64	67	70	73	76	79	82	85	89	92	95
10	57	60	63	67	70	73	76	80	83	87	90	93	97	100
	59	62	66	69	73	77	80	84	88	92	95	99	103	107
	67	70	73	76	80	83	86	90	93	97	100	104	107	111
11	68	72	75	79	83	86	90	94	98	101	105	109	113	117
	70	74	78	82	86	90	94	98	103	107	111	115	119	124
	80	83	86	90	93	97	101	105	108	112	116	120	124	128
12	81	84	88	92	96	100	105	109	113	117	121	126	130	134
	83	87	91	96	100	105	109	114	118	123	128	132	137	142
	93	96	100	104	108	112	116	120	125	129	133	137	142	146
13	94	98	102	107	111	116	120	125	129	134	139	143	148	153
	96	101	105	110	115	120	125	130	135	140	145	150	155	160
	107	111	115	119	123	128	132	137	142	146	151	156	161	165
14	109	113	117	122	127	132	137	142	147	152	157	162	167	172
	110	116	121	126	131	137	142	147	153	158	164	169	175	180
	122	126	131	135	140	145	150	155	160	165	170	175	180	185
15	124	128	133	139	144	149	154	160	165	171	176	182	187	193
	126	131	137	143	148	154	160	166	172	178	184	189	195	201

Tabelle A.6: *KS-1–Test; kritische Werte* $d^{(1)}_{n;1-\alpha}$
(alle fünf nach Hartung, 2002)

n	$d^{(1)}_{n;0.80}$	$d^{(1)}_{n;0.90}$	$d^{(1)}_{n;0.95}$	$d^{(1)}_{n;0.98}$	$d^{(1)}_{n;0.99}$
5	1.00	1.14	1.26	1.40	1.50
8	1.01	1.16	1.28	1.43	1.53
10	1.02	1.17	1.29	1.45	1.55
20	1.04	1.19	1.31	1.47	1.57
40	1.05	1.20	1.33	1.49	1.59
> 40	1.08	1.23	1.36	1.52	1.63

Tabelle A.7: *KS-2–Test* ($n_X = n_Y$)*; kritische Werte* $d^{(2)}_{n;1-\alpha}$

n	α				
	0.20	0.10	0.05	0.02	0.01
2	3	3			
3	4	4	4		
4	7	6	5	5	5
5	11	9	7	6	6
6	16	13	10	9	8
7	22	17	14	11	10
8	28	22	18	15	13
9	36	28	23	18	16
10		34	28	22	20
11			33	27	24
12			40	32	28
13				37	33
14					38
$n > 40$					
	$1.527\sqrt{n}$	$1.739\sqrt{n}$	$1.923\sqrt{n}$	$2.150\sqrt{n}$	$2.305\sqrt{n}$

Tabelle A.8: *KS-2–Test; kritische Werte* $d^{(2)}_{n_X,n_Y;1-\alpha}$

α	$n_{(1)}$	$n_{(2)}$	$d^{(2)}_{n_{(1)},n_{(2)};1-\alpha}$
0.20	2–4	5–40	1.02
	5–15	5–40	1.03
	sonst		1.08
0.10	2–3	3–12	1.10
	4–8	5–9	1.12
	4–16	10–20	1.16
	sonst		1.23
0.05	2–4	3–15	1.22
	5–16	6–20	1.30
	sonst		1.36

Festsetzung: $n_{(1)} < n_{(2)}$

Tabelle A.9: *Wilcoxon–Vorzeichenrangtest; kritische Werte* $w_{n;\gamma}$

n	$w_{n;0.01}$	$w_{n;0.025}$	$w_{n;0.05}$	$w_{n;0.1}$	$w_{n;0.9}$	$w_{n;0.95}$	$w_{n;0.975}$	$w_{n;0.99}$
4	0	0	0	1	8	9	10	10
5	0	0	1	3	11	13	14	14
6	0	1	3	4	16	17	19	20
7	1	3	4	6	21	23	24	26
8	2	4	6	9	26	29	31	33
9	4	6	9	11	33	35	38	40
10	6	9	11	15	39	43	45	48
11	8	11	14	18	47	51	54	57
12	10	14	18	22	55	59	62	66
13	13	18	22	27	63	68	72	77
14	16	22	26	32	72	78	82	88
15	20	26	31	37	82	88	93	99
16	24	30	36	43	92	99	105	111
17	28	35	42	49	103	110	117	124
18	33	41	48	56	114	122	129	137
19	38	47	54	63	126	135	142	151
20	44	53	61	70	139	148	156	165

Tabelle A.10: *Kruskal–Wallis–Test; kritische Werte* $h_{3;(n_1,n_2,n_3);1-\alpha}$

n	n_1	n_2	n_3	$h_{3;(n_1,n_2,n_3);1-\alpha}$	
				$\alpha = 0.10$	$\alpha = 0.05$
7	1	2	4	4.50	4.82
	1	3	3	4.57	5.14
	2	2	3	4.50	4.71
8	1	2	5	4.20	5.00
	1	3	4	4.06	5.21
	2	2	4	4.46	5.13
	2	3	3	4.56	5.14
9	1	3	5	4.02	4.87
	1	4	4	4.07	4.87
	2	2	5	4.37	5.04
	2	3	4	4.51	5.40
	3	3	3	4.62	5.60
10	1	4	5	3.96	4.86
	2	3	5	4.49	5.11
	2	4	4	4.55	5.24
	3	3	4	4.70	5.72
11	1	5	5	4.04	4.91
	2	4	5	4.52	5.27
	3	3	5	4.41	5.52
	3	4	4	4.48	5.58
12	2	5	5	4.51	5.25
	3	4	5	4.52	5.63
	4	4	4	4.50	5.65
13	3	5	5	4.55	5.63
	4	4	5	4.62	5.62
14	4	5	5	4.52	5.64
15	5	5	5	4.56	5.66