STATISTIK und WAHRSCHEINLICH-KEITSRECHNUNG

für InformatikerInnen

SS 2009

O.Univ.-Prof. Dipl.-Ing. Dr.techn. Rudolf DUTTER
unter Mitwirkung von Ao.Univ.-Prof. Dipl.-Ing. Dr.techn. Peter FILZMOSER

7. Februar 2009

Dieses Skriptum dient zur Unterstützung der Vorlesung. Es ist im Prinzip als Hilfestellung und auch als (allerdings sehr knappes) Nachschlagewerk gedacht. Das Stichwortverzeichnis (der Index) sollte das formale Finden von Prüfungsfragen (und möglichst deren Antworten) erleichtern.

Die Neubearbeitung im SS 2009 enthält auch systematische Hinweise auf praktische Bearbeitungen mit dem Computerprogrammsystem R.

Inhaltsverzeichnis

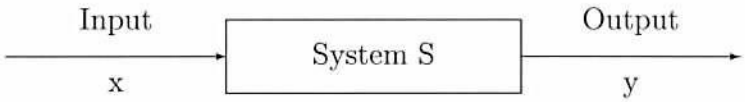
Abbildungsverzeichnis

Tabellenverzeichnis

Kapitel 1

Motivierung mit Anwendungsbeispielen

In der Informatik beschäftigt man sich häufig mit Systemen, die auf *Eingaben*, also *Inputs*, reagieren und nach irgendwelchen Vorschriften *Ausgaben*, also *Outputs*, produzieren. Diese Systeme können z.B. Hard- oder Softwaresysteme sein. Interessiert der konkrete Inhalt nicht, so spricht man von einer *Schwarzen Schachtel (black box)*.



Der Übergang von x nach y wird durch einen *Leistungsparameter* z (*performance parameter*) charakterisiert, der natürlich vom Zustand des Systems abhängt. Beispiele sind

- Ausführungszeit (von Programmen) des Übergangs von x nach y
- Anzahl der Ausführungsschritte, um von x nach y zu kommen
- Korrektheit (z.B. $z = 0$ wenn Ergebnis richtig ist, $= 1$ sonst)
- Betriebsmittelbedarf zur Erlangung des Ergebnisses
- Durchführbarkeit (z.B. $z = 1$ falls bereitgestellte Betriebsmittel nicht ausreichen, $z = 0$ sonst)
- Eintritt/Nichteintritt von speziellen Systemzuständen (z.B. Überlauf, Systemverklemmung - „deadlock“)
- Anzahl von Prozessorbelegungen je Zeiteinheit.

Weitere Beispiele sind in der folgenden Tabelle 1.1, abhängig von der Art des Systems, dargestellt.

Typische Analysen eines solchen Leistungsparameters sind

- „worst-case“ - Analyse
- „average-case“ - Analyse
- „Median-Fall“ - Analyse

Es gibt natürlich noch viele andere Möglichkeiten der Untersuchung von Systemen, die irgendwie auch vom *Zufall* beeinflusst sind, und von denen werden einige in dieser Vorlesung diskutiert.

Tabelle 1.1: Beispiele von Systemen mit Input, Output und Leistungsparameter (Pflug, 1986).

System	Input	Output	Leistungsparameter
Anwendungsprogramme	Daten	Ergebnis	Laufzeit, Korrektheit, Speicherbedarf
Compiler	Quellenprogramm	Objektprogramm	Compilationszeit, Korrektheit
Dateizugriffsroutine	Anforderung	Daten	Zugriffszeit
Information-Retrieval-System	Anfrage	Antwort	Zugriffszeit, Korrektheit
Mustererkennungsprogramm	Muster	Klassifikation	Korrekte Klassifikation
Terminal-Betriebssystem	Kommando	Rückmeldung	Response-Zeit
Batch-Betriebssystem	Benutzerjob	Ergebnis	Turnaround-Zeit

Kapitel 2

Einleitung - Was ist Statistik?

2.1 Überblick und Definition

Beim Wort „Statistik" denkt man zunächst meist an umfangreiche Tabellen und an grafische Darstellungen, die irgendwelche Sachverhalte verdeutlichen (oder auch verschleiern) sollen. Damit erfasst man jedoch nur einen Teil der Statistik, die *Beschreibende Statistik* oder *Deskriptive Statistik*. Diese dient dazu, umfangreiche Datensätze einerseits möglichst übersichtlich und anschaulich darzustellen und zum anderen durch möglichst wenige, einfache Maßzahlen (wie Mittelwert und Streuung) zu ersetzen („Datenreduktion").

Oft fasst man die Daten jedoch nur als *Stichprobe* aus einer größeren (wirklichen oder hypothetischen) *Population* auf und möchte ihre Aussage auf die gesamte Population *verallgemeinern*. Dabei muss man die zufälligen Schwankungen der Stichprobenwerte um die entsprechenden Werte in der Gesamtpopulation berücksichtigen und quantitativ erfassen. Hierzu dient die *Analytische Statistik* (*Beurteilende*, *Schließende Statistik*, auch *Mathematische Statistik* im weiteren Sinne genannt). Sie baut auf den *Zufallsmodellen* und *Zufallsgesetzen* auf, die in der *Wahrscheinlichkeitstheorie* hergeleitet werden, und versucht unter anderem, möglichst einfache und begründete oder bewährte statistische Modelle an die Daten anzupassen (*Schätzungen*) und die Güte der Anpassung zu prüfen (*Tests*).

Die Analytische Statistik ist gleichsam die Umkehrung der Wahrscheinlichkeitstheorie: diese geht von Wahrscheinlichkeitsmodellen aus und berechnet die Wahrscheinlichkeiten zukünftiger Beobachtungen; die Analytische Statistik geht von den Beobachtungen aus und versucht, Schlüsse auf das zugrundeliegende Wahrscheinlichkeitsmodell zu ziehen. Wegen ihres engen Zusammenhanges werden beide Gebiete auch unter dem Namen *Stochastik* (Lehre vom Zufall) zusammengefasst.

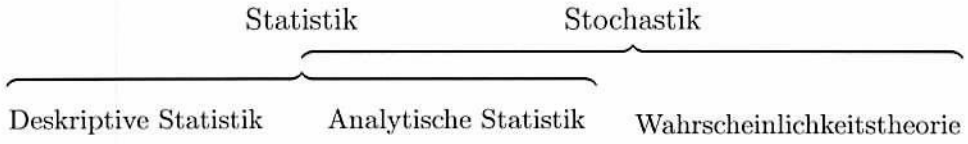
Innerhalb der Analytischen Statistik leitet die *Mathematische Statistik* (im engeren Sinne) die mathematischen Eigenschaften verschiedener Schätz- und Testmethoden unter wohldefinierten Voraussetzungen her, und die *Angewandte Statistik* oder *Statistische Datenanalyse* benutzt diese Erkenntnisse bei der Auswertung konkreter Daten. Für diese Auswertung werden auch oft Methoden der Deskriptiven

Statistik angewandt, besonders dann, wenn man erst nach plausiblen Modellen suchen muss („forschende" oder *explorative Datenanalyse*). Demgegenüber bezeichnet man die Situation, in der man nur ein oder einige wenige gegebene Modelle statistisch prüft, auch als „bestätigende" oder „konfirmatorische Datenanalyse" (etwas unglücklich, da Hypothesen meist nur verworfen und nicht bewiesen werden können).

Nun sind allerdings der rein schematischen, mechanischen Anwendung statistischer Methoden gewisse Schranken gesetzt. Bekanntlich stimmen Modell und Wirklichkeit stets nur mehr oder weniger genau überein. Außerdem stehen oft mehrere, nahezu gleichwertige Auswertungsmethoden zur freien Auswahl. Andererseits muss man sich sehr oft mit künstlich vereinfachten Modellen begnügen, weil das realistischere, kompliziertere Modell entweder noch gar nicht mathematisch erforscht ist oder unter den gegebenen Umständen zu aufwendig wäre; und es ist eine Ermessensfrage zu beurteilen, ob das einfache Modell für die vorgesehenen Zwecke hinreichend genau ist. Aus all dem ergibt sich, dass in der fortgeschrittenen Datenanalyse (wie natürlich in guter Forschung überall) auch die Intuition eine gewisse Rolle spielt: nicht umsonst spricht man von der „Kunst der Datenanalyse".

Zusammenfassend können wir definieren: Die *Statistik* ist die Wissenschaft und Kunst der Erfassung und Analyse von Datenstrukturen unter Berücksichtigung der unvermeidlichen Unschärfe, die durch zufällige Schwankungen und Fehler verursacht wird.

Die Gliederung der Statistik können wir in folgendem Schema darstellen:



2.2 Was leistet die Statistik?

Sie gibt Wege an, um auch umfangreiche und komplizierte Datensätze möglichst anschaulich darzustellen.

Sie zeigt, wie man große Datenmengen durch einige wenige Zahlen und eventuelle Zusatzangaben (z.B. „ungefähre Exponentialverteilung") mit minimalem Informationsverlust zusammenfassen kann.

Sie beschreibt einfache und wichtige Modelle zur Erfassung der Natur (z.B. linearer oder quadratischer Zusammenhang zweier Größen oder ihrer Logarithmen; additives oder multiplikatives Zusammenwirken mehrerer Faktoren).

Sie bietet häufig benutzbare Modelle für die zufälligen Schwankungen und zufälligen Fehler, die in Daten beobachtet werden können (z.B. Binomial- und Poisson-Verteilung für die Anzahl der Fälle, in denen ein bestimmtes Ereignis zufällig eintritt; Normalverteilung - *Gauß'sche Glockenkurve* - für die Größe von Messfehlern).

Sie untersucht und vergleicht verschiedene Versuchspläne gleichen Umfangs zur Messung mehrerer Effekte, oder zur Prüfung einer Hypothese, oder zur schrittweisen Suche eines Optimums oder zum Ziehen einer Stichprobe aus einer strukturierten Grundgesamtheit. Dadurch kann, bei gleicher Genauigkeit, der Versuchsaufwand oft stark reduziert werden.

Sie prüft, inwieweit beobachtete Abweichungen von einem Modell dem Zufall zugeschrieben werden können, also ob Daten und Modell oder Hypothese im Rahmen der zufälligen Fehler miteinander vereinbar sind oder nicht (*Tests*).

Sie liefert eine möglichst gute Anpassung der unbekannten Konstanten (*Parameter*) eines Modells an die Daten, unter Berücksichtigung des Vorhandenseins von zufälligen (eventuell auch groben) Fehlern (*Schätzungen*, genauer *Punktschätzungen*); gleichzeitig gibt sie die ungefähre Genauigkeit dieser Anpassung an (*Standardfehler*, *Vertrauensbereiche*).

Ferner, auf einer höheren Stufe der Theorie:

Sie gibt die Grundlagen für möglichst vernünftige (rationale) Entscheidungen, besonders Routine-Entscheidungen, und studiert die damit verbundenen langfristigen Risken (*Entscheidungstheorie*).

Sie versucht, möglichst genau den Zustand unseres unvollkommenen und unsicheren Teilwissens zu beschreiben (*Likelihood-Theorie*).

Sie bietet Formalismen, um (subjektive oder objektive) Vorkenntnisse explizit und zwangsläufig in die Datenanalyse einzubauen (*Bayes-Theorie*).

2.3 Einige typische Fragestellungen („Kleines Einmaleins" der Statistik)

Gegeben sei ein Haufen von Zahlen, die alle dieselbe Größe oder Variable messen (z.B. die Güte eines Erzvorkommens an verschiedenen Orten einer ungefähr homogenen Lagerstätte).

Wie sieht die *Verteilung* dieser Zahlen aus? Ein- oder mehrgipfelig, symmetrisch oder schief, mit oder ohne Ausreisser? (Darstellung und Beschreibung einer *Stichprobe.*)

Wo liegen Maximum, Minimum, Mittelwert (und andere Kenngrößen); wie stark streuen die Zahlen um den Mittelwert? („Datenreduktion", Kurzbeschreibung der Stichprobe durch *Statistiken*; diese dienen oft gleichzeitig als *Schätzungen* für unbekannte *Parameter.*)

Stimmt die Verteilung im Rahmen der Beobachtungsgenauigkeit mit einer *Gaußschen Glockenkurve* überein? (*Test einer Verteilung.*)

Ist der Mittelwert mit dem (sehr genau bekannten) Mittelwert einer Theorie (oder einer viel größeren Population, z.B. Bevölkerung eines Landes) vereinbar? (*Test eines Parameters.*)

Dazu: Wie groß ist die Wahrscheinlichkeit, eine Änderung des Populationsmittels (Parameters) um z.B. +20% durch den Test zu entdecken? (*Macht des Tests.*)

Wieviele Beobachtungen benötigt man, um eine Änderung des Parameters um +20% mit genügend großer Wahrscheinlichkeit nachzuweisen, wenn die Genauigkeit der Daten bekannt ist? (Bestimmung des benötigten *Stichprobenumfangs.*)

Wie genau ist durch den Stichprobenmittelwert der Mittelwert in der ganzen Population bestimmt? (Angabe durch den *Standardfehler* des Mittels, oder durch einen *Vertrauensbereich* für den Populationsmittelwert.)

Gegeben sei eine einzige, beobachtete Anzahl (z.B. die Anzahl Bakterienkolonien auf einem Nährmedium, oder die Anzahl Personen, die bei einer Meinungsumfrage für etwas sind, oder die Anzahl von Benutzern eines Computers pro Tag, von Zugriffen auf eine Datenbank pro Zeiteinheit - *Zähldaten* im Gegensatz zu *Messdaten*).

Dann gibt in vielen Fällen diese Anzahl ihre eigene Genauigkeit an. Man braucht also keine wiederholten Beobachtungen und kann trotzdem alle genannten Fragen über die mittlere Anzahl bzw. die Gesamtzahl in der untersuchten Population beantworten (Test eines Parameters, Macht gegenüber anderen hypothetischen Parametern, benötigter Stichprobenumfang, Schätzung des zufälligen Fehlers und Angabe von Vertrauensbereichen).

Gegeben seien 2 Stichproben von Messdaten aus 2 verschiedenen Populationen (z.B. Gewichte von Männern und Gewichte von Frauen, SO_2-Ausstoß vor und nach Einbau eines Filters in einer Verbrennungsanlage, Korngrößen von zwei verschiedenen Lagern).

Ist es möglich, dass die Populationsmittel trotz des Unterschieds der beiden Stichprobenmittel gleich sind? (Test der Gleichheit zweier Parameter.)

Ist es möglich, dass sogar die gesamten Verteilungen (also auch Streuung und Gestalt der Verteilungen) in beiden Populationen gleich sind? (Test der Gleichheit zweier Verteilungen.)

Gegeben seien Beobachtungen, die nach 2 Merkmalen gleichzeitig klassifiziert werden (z.B. Personen nach Raucher und Lungenkrebs: jeweils ja oder nein; Korngröße und Festigkeit).

Ist es möglich, dass die beiden Merkmale in der untersuchten Population (statistisch) unabhängig sind, oder besteht ein gesicherter statistischer (!) (nicht notwendig kausaler!) Zusammenhang? (Test einer Verteilung, die sich aus dem speziellen „Modell" der „Unabhängigkeit" ergibt.)

Gegeben seien Objekte, bei denen 2 Größen gleichzeitig gemessen wurden (z.B. Körpergröße vom Vater und vom Sohn bei Vater/Sohn-Paaren; Studentenzahl und durchschnittlicher Erfolg, Maschinenbelegung und Zugriffszeit).

Besteht ein gesicherter statistischer (!) Zusammenhang, eine *Korrelation*, zwischen den beiden Größen in der Gesamtpopulation? (Test eines Parameters: Prüfung, ob die Korrelation in der Population gleich Null sein kann.)

Diese Fragestellungen und ihre Beantwortung in den genannten einfachen Fällen sind gewissermaßen das kleine Einmaleins der Statistik. Es genügt in erster Linie oft auch für kompliziertere Situationen: diese kann man gleichsam in ihre Elementarbausteine zerlegen und sie mit etwas Überlegung und etwas Grundver-

ständnis für Zufallsgesetze wieder zusammenkitten. Dieselben Fragestellungen und dieselben Lösungsprinzipien treten auch in den höheren Teilgebieten der Statistik auf (z.B. in der Varianzanalyse, Regression, Multivariate Statistik, Zeitreihenanalyse, Analyse räumlich abhängiger Daten), und das Verständnis der elementaren Probleme ermöglicht zwar nicht das Verständnis komplizierter mathematischer Beweise, wohl aber eventuell das Verständnis ihrer Resultate, das für den Anwender der Statistik entscheidend ist. Zum Begreifen der meisten elementaren Lösungen (nämlich aller, die Schlüsse über die Stichprobe hinaus auf eine umfassendere Population ziehen) sind natürlich bereits Kenntnisse der einfachsten und wichtigsten Zufallsgesetze, also Grundkenntnisse der Wahrscheinlichkeitstheorie nötig.

2.4 Es stimmt nicht, dass ...

... man mit Statistik alles beweisen kann. (In einem engeren Sinn kann man sogar behaupten, dass sich Hypothesen nur statistisch widerlegen, niemals aber beweisen lassen.) Allerdings lässt sich bekanntlich Statistik leicht und stark zu „Lügen" verschiedenster Art missbrauchen.

... der Statistiker einem sagt, man müsse alles 10 mal tun. Bei Zählungen genügt meist eine einzige Zahl, und bei Messungen unter verschiedenen Versuchsbedingungen reicht oft 1/4 oder 1/8 des vollständigen Versuchsplans vollauf.

... die Befragung von 1000 Menschen aus einer Million genauso dumm und nichtssagend ist wie die von einem Menschen aus 1000. Sie ist etwa so informativ wie die Befragung von 500 aus 1000, oder von 1001 aus einer unendlichen Grundgesamtheit.

... jemand mit den Füßen im Eis und dem Kopf im Feuer sich im statistischen Mittel wohl fühlt. Die Streuung ist meist genauso wichtig wie das Mittel.

... der Fehler des arithmetischen Mittels bei wachsender Beobachtungszahl gegen Null strebt. Er strebt gegen den systematischen Fehler.

... man bei der Durchführung eines Experiments immer alle Versuchsbedingungen außer einer konstant halten soll. Wenn man (sofern möglich) *planmäßig* viele Versuchsbedingungen gleichzeitig variiert, so erhält man bei gleichem Versuchsumfang genauere Ergebnisse, die sich zudem leichter verallgemeinern lassen.

... man Wiederholungen eines Experiments stets unter möglichst gleichartigen Bedingungen durchführen soll. Man soll sie in der Regel unter möglichst verschiedenartigen Bedingungen durchführen, um die Verallgemeinerungsfähigkeit zu erhöhen.

2.5 Wie man mit Statistik lügt

Missbräuche der Statistik sind so zahlreich, dass es sogar ganze Bücher darüber gibt. Ihre Ursachen reichen von Dummheit und Naivität über zwielichtiges geschicktes Manövrieren „an den Grenzen der Legalität" bis hin zu glattem Betrug.

Man denke nur an Werbung und Politik; aber selbst in der Wissenschaft muss man auf der Hut sein. Zu den Quellen, die die sprichwörtliche „Verdrehbarkeit“ der Statistik ermöglichen, gehören die folgenden:

1. Unkenntnis der statistischen Methoden und ihrer vernünftigen Anwendung.

2. Das Odium der Exaktheit, das einer Zahl anhaftet. Im Gegensatz zu Zahlen in der Reinen Mathematik ist eine Zahl in der Angewandten Statistik stets etwas Verschwommenes, mit einem zufälligen Fehler behaftet; und wenn die Größenordnung dieses Fehlers (z.B. durch Angabe eines „Standardfehlers“ oder Angabe der geltenden Ziffern) nicht klar aus dem Kontext hervorgeht, so ist die Zahl wertlos, da sie alles bedeuten kann.

3. Oft soll die Statistik zur „Sanktionierung“ (wissenschaftliche „Heiligsprechung“) von Versuchsergebnissen dienen. Die Statistik ist jedoch niemals Selbstzweck, sondern muss eingebettet bleiben in dem gesamten Denk- und Forschungsprozess. Sie ist ein Werkzeug, das zur Erweiterung und Schärfung des „gesunden Menschenverstandes“ dient, aber kein Ersatz für ihn.

4. Viele Zahlen und Diagramme bedeuten bei genauem Hinsehen etwas ganz anderes, als suggeriert wird; oder es lässt sich überhaupt nicht eruieren, was sie eigentlich bedeuten (Reklame, Politik, Polemiken, ...).

5. Oft wird wichtige Zusatzinformation (versehentlich oder absichtlich) unterschlagen.

6. In vielen Untersuchungen stecken systematische Fehler, die oft nur sehr schwer erkannt und vermieden werden können. (Berühmt sind dafür Meinungsumfragen; aber auch Naturkonstanten sind oft weit über den angeblichen Messfehler hinaus systematisch falsch.)

7. Aus einem statistischen Zusammenhang, einer „signifikanten Korrelation“, folgt noch lange kein Kausalzusammenhang. (Ein bekanntes Beispiel ist die hohe Korrelation zwischen Zahl der Störche und Zahl der Geburten; damit ließe sich sonst „statistisch beweisen“, dass der Klapperstorch die Babies bringt.)

8. Es ist oft erstaunlich schwer zu beurteilen, auf welche Populationen sich ein Ergebnis verallgemeinern lässt. Das gilt insbesondere auch für zeitliche Extrapolation (Prognosen, z.B. Wirtschafts- oder Bevölkerungsprognosen).

Einige Standardfragen an zweifelhafte Statistiken sind: „Na und, was soll's?“ „Was für eine Absicht steckt dahinter?“ „Was fehlt? Was wurde vergessen oder verschwiegen?“ „Woher weiß man, dass?“ „Was besagt das wirklich?“

2.6 Einige besondere Anwendungsbeispiele

Viele wertvolle wissenschaftliche Erkenntnisse können bereits ohne Analytische Statistik, durch das bloße sorgfältige Anschauen von Daten gewonnen werden. So fiel einem australischen Augenarzt auf, dass nach einer schweren Rötelepidemie die Anzahl Neugeborener mit Augenschäden gestiegen war. Er ging der Sache nach und entdeckte so den Einfluss von Röteln während der Schwangerschaft auf Missbildungen bei Neugeborenen. Bei Atomgewichtsbestimmungen von Stickstoff bemerkte Rutherford eine Diskrepanz zwischen den Werten für Stickstoff aus verschiedenen chemischen Verbindungen und denen für „Stickstoff" aus Luft; das führte ihn zur Entdeckung des ersten Edelgases Argon.

Manchmal erfordert schon die Aufbereitung von Daten zum Anschauen komplizierte mathematische Hilfsmittel. Bei der Fourieranalyse von Meereswellen vor der mexikanischen Pazifikküste traten ganz schwach Wellen zutage, die 1 km lang und 1 mm hoch waren, um die halbe Erdkugel gewandert waren und, wie Wetterkarten nachträglich ergaben, von Stürmen im südlichen Indischen Ozean stammten.

Zuweilen ergeben sich wertvolle Konsequenzen aus einem einfachen statistischen Argument. In der ersten Hälfte des vorigen Jahrhunderts gelang dem Geologen Lyell die erste, bahnbrechende relative Chronologie von tertiären Gesteinen aus verschiedenen Teilen Europas, indem er den Prozentsatz an Versteinerungen heute noch überlebender Muschelarten bestimmte: je mehr, desto jünger die Schicht. Die Namen, die Lyell den Schichten gab, vom Pleistozän bis zum Eozän, sind heute noch gebräuchlich. – Zu Anfang dieses Jahrhunderts wiesen Biometriker durch umfangreiche Messungen allerlei lokale Unterschiede zwischen Lebewesen derselben Art nach, oft sogar zwischen Fischen einer Art aus demselben Fjord. Die europäischen Aale jedoch, von Island bis zum Nil, zeigten keinerlei lokale Variation; das legte die überraschende Hypothese eines gemeinsamen Laichplatzes nahe, der dann auch in der Saragossasee gefunden wurde.

Gelegentlich kommt die Erfindung einer neuen statistischen Methode gerade zur rechten Zeit. Vom ersten, um 1800 entdeckten Kleinplaneten Ceres gelangen zunächst nur wenige Beobachtungen, bevor er verschwand, und die baldige Wiederauffindung wurde nur dadurch möglich, dass Gauß mit seiner neuen Methode der Kleinsten Quadrate erstmals aus so wenigen Daten eine genügend genaue Bahn berechnen konnte.

Als Beispiel aus der Gegenwart mag die Methode der „Multidimensionalen Skalierung" dienen, die aus vagen Ähnlichkeitsangaben für je 2 Objekte räumliche Muster in beliebiger Dimension konstruiert. So kann sie die zeitliche Reihenfolge von Gräbern auf Grund der Ähnlichkeit von Grabbeigaben, oder Landkarten auf Grund von Nachbarschaftsangaben darstellen. Beim Sehvorgang wird offenbar, vereinfacht ausgedrückt, das Netzhautbild punktweise durch Nervenfasern auf ein entsprechendes Bild im Gehirn übertragen. Wenn man einem Goldfisch den Sehnerv durchschneidet, so wachsen die Enden der getrennten Nervenfasern zwar wieder zusammen, aber offenbar längst nicht immer zueinandergehörige Enden. Nach einiger

Zeit kann der Goldfisch jedoch wieder sehen: es scheint, dass die falsch verbundenen Nervenenden im Gehirn wandern, bis sie wieder ein ortsgetreues Netzhautbild übertragen, und zwar unter Ausnutzung der lokalen statistischen Ähnlichkeitsinformation, die die oft gleichzeitige Reizung benachbarter Netzhautpunkte übermittelt, wie im mathematischen Modell der Multidimensionalen Skalierung.

2.7 Wozu Statistik lernen?

Grundkenntnisse der Statistik ermöglichen es ...

- ... kleine statistische Anwendungsprobleme mit den eigenen Daten (angefangen bei der Diplomarbeit) selber zu lösen;
- ... bei größeren Problemen sinnvoll mit dem beratenden Statistiker zusammenzuarbeiten;
- ... die Statistik in anderen wissenschaftlichen Arbeiten (wenigstens in den Grundzügen) zu verstehen;
- ... die vielen Missbräuche und Fehler leichter zu durchschauen und selbständig zu beurteilen.

Die für die Statistik benötigte Wahrscheinlichkeitstheorie

- ... gestattet die Lösung kleinerer Aufgaben zur Berechnung von Wahrscheinlichkeiten, wie sie auch in Forschung und Entwicklung immer wieder einmal vorkommen;
- ... öffnet die Tür für das Verständnis der stochastischen Modelle, die in vielen Disziplinen immer bedeutsamer werden;
- ... ermöglicht das Eindringen in Nachbargebiete wie Spieltheorie und Informationstheorie;
- ... bietet manchem eine neue Perspektive des Weltbilds.

Beide Gebiete trainieren das Abstraktionsvermögen, erweitern die Allgemeinbildung und bereiten dem einen oder anderen sogar Freude über neue Fähigkeiten und Erkenntnisse.

Kapitel 3

Beschreibende Statistik

3.1 Messniveau von Daten

Beispiele von Daten:

(a) Blutgruppe, Steuerklassen, Geschlecht, Kraftfahrzeugkennzeichen, erlernter Beruf eines Computerbenutzers.

(b) Schulnoten, sozialer Status, Erdbebenstärke.

(c) Temperatur (in ^oC), Zeitdauer bei verschiedenen Startpunkten.

(d) Gewicht, Länge, Kosten, Korngrößen.

Ad (a) ***Nominalskala*** (klassifikatorische Skala):
Objekte werden nach bestimmten Regeln in Klassen eingeteilt. Diese Klassen werden durch Symbole (auch Ziffern) gekennzeichnet. Man erhält keine Wertung, nur eine Anordnung. Daher würde auch eine 1:1 Transformation der Daten in den folgenden statistischen Kennzahlen im wesentlichen nichts ändern: Relative Häufigkeit, Modalwert etc.

Ad (b) ***Ordinal-(Rang-)skala:***
Es gibt eine Rangordnung, eine Auszeichnung bestimmter Objekte vor anderen. Dagegen haben Differenzen zwischen den Messdaten keine Bedeutung. Die Ordinalskala vermittelt mehr Information als die Nominalskala. Daher ist nur eine *monoton steigende* Transformation zulässig, um die Struktur der Daten nicht zu zerstören. Statistische Kennzahlen: Quantile (Median), Rangstatistiken bleiben unverändert.

Ad (c) ***Intervallskala:***
Hier handelt es sich bereits um reelle Zahlen: Quantitative Daten. Differenzen sind wichtig, aber die absoluten Werte haben keine Bedeutung: „Auf den Bus bis 4 Uhr warten“ sagt nichts über die Wartezeit aus; 10^oC ist nicht doppelt

so warm wie 5^oC. Lineare Transformationen $y = ax + b \ \ (a > 0)$ lassen die Skala invariant.
Z.B. x: ^oC, y: ^oF, Transformation: $y = 1.8x + 32$.

Statistische Kennzahlen, die in dieser Skala gemessen werden, sind: Arithmetisches Mittel, Streuung, Korrelationskoeffizient.

Ad (d) ***Verhältnisskala:***
Es gibt einen festen Nullpunkt. Z.B. Temperatur in ^{o} Kelvin, Gewicht, Länge, Kosten, elektrische Spannung. Bei der Transformation $y = ax$ wird der Nullpunkt nicht verändert. Die relative Erhöhung des Umsatzes kann in Dollar oder in Euro gerechnet werden; das Resultat ist das gleiche.

Meistens erscheint es sinnvoll, eine gröbere Unterscheidung nur zwischen der *topologischen Skala* (= Nominal- und Ordinalskala) und der *Kardinalskala* (= Intervall- und Verhältnisskala) zu treffen. In der topologischen Skala gibt es i.a. nur diskrete Werte, während die Kardinalskala reelle Zahlen verwendet.

3.2 Verteilungen

Gegeben sei eine Datenmenge, i.a. eine Menge von Zahlen. Man möchte sich ein Bild dieser Daten machen, und wie der Wunsch schon anregt, geschieht dies am besten grafisch.

3.2.1 Histogramm

Die Zugriffszeiten auf bestimmte Daten in 5 Computern der gleichen Größe wurden gemessen: 68, 72, 66, 67 und 68 ms. Wie sind diese Zeiten größenmäßig verteilt? Ein Histogramm könnte folgendermaßen aussehen (Abbildung 3.1):

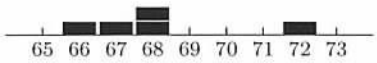


Abbildung 3.1: Zugriffszeiten in $[ms]$.

Interessanter wird es, wenn mehr Zahlenmaterial vorliegt; z.B. 80 Zugriffszeiten auf Daten. Diese Werte sind in der Abbildung 3.2 als Histogramm zusammengefasst. Vermutlich liegt die durchschnittliche Zeit bei 68 oder 69 Millisekunden.

Die Messgröße (Zugriffszeit) ist zweifellos eine stetige Variable. Allerdings wird sie wegen der natürlichen Gegebenheiten nur diskret registriert. Um die *Häufigkeiten* in einem Histogramm auftragen zu können, müssen aber auf alle Fälle die Daten diskretisiert werden. Die Frage der Klassenbreite bleibt aber offen. Wird

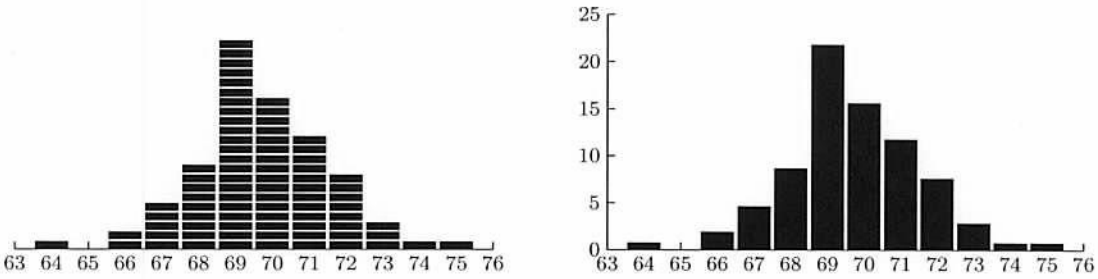


Abbildung 3.2: Einfache Histogramme.

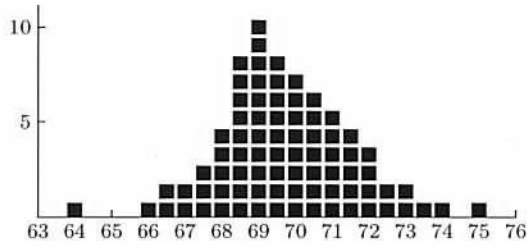


Abbildung 3.3: Histogramme mit Werten in 1/2 $[ms]$.

die Zeit auf $\frac{1}{2}$ ms genau angegeben, dann könnte man ein Histogramm wie in der Abbildung 3.3 zeichnen.

Theoretisch könnte man die Einteilung immer feiner machen, genauer messen und öfter Zugriffszeiten untersuchen, sodass die Kontur des Histogramms gegen eine glatte Kurve strebt, z.B. gegen jene wie die Abbildung 3.4 dargestellt.

Die Häufigkeiten wurden in absoluten Werten eingetragen (absolute Häufigkeiten). Genauso hätte man die Häufigkeiten auf die Gesamtanzahl (hier 80) beziehen können (relative Häufigkeiten).

Manchmal fällt ein Datenpunkt gerade auf das Ende eines Intervalls, und man muss entscheiden, in welche Klasse der Wert gegeben wird. In unserem Beispiel wurde der „Beobachter“ (die Person, die Zugriffszeiten feststellte) beauftragt, die Zeiten auf ein Zehntel ms aufzunehmen und für den Fall, dass eine Größe nahe einer halben ms liegt, ein + bzw. − Zeichen zur entsprechenden Zahl zu schreiben, wenn der Wert knapp darüber bzw. darunter liegt. Die gemessenen Werte werden in der Tabelle 3.1 angegeben.

Das bisher Besprochene entspricht der Konstruktion eines Stabdiagramms. Dabei werden einfach die Häufigkeiten der verschiedenen Werte gezählt und diese Anzahl vertikal visualisiert.

Ähnlich, aber doch anders, wird man vorgehen, wenn allgemein reelle Zahlen (mit beliebig vielen Dezimalstellen) gegeben sind. Hier kann es ja auch passieren, dass alle Werte verschieden sind. In diesem Fall wird das „Histogramm“ anders konstruiert:

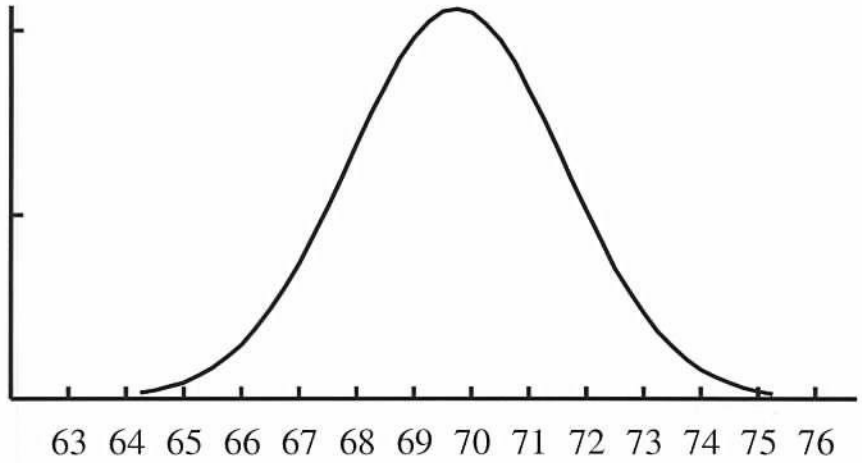


Abbildung 3.4: Stilisiertes, geglättetes Histogramm.

Tabelle 3.1: 80 gemessene Werte von Zugriffszeiten.

70.0	68.6	67.9	66.3	71.0	64.2	69.6	71.0
69.7	69.0	73.4	69.0	70.1	69.8	69.0	73.0
70.0	69.1	69.5-	67.9	72.8	72.1	69.5+	70.1
70.2	69.1	68.9	68.3	74.9	68.4	69.1	66.6
71.1	69.2	71.2	68.9	70.9	70.6	69.9	69.9
69.4	69.5+	68.5+	70.9	71.6	68.9	72.0	70.3
68.6	68.5-	67.8	72.2	68.7	70.6	66.9	69.3
71.4	68.7	74.2	68.8	71.4	71.8	67.5-	70.4
71.4	67.4	69.5-	72.4	70.4	69.3	68.2	67.0
71.7	70.5+	72.5-	68.2	67.6	68.6	70.5-	65.8

In der Praxis (typisch mit Computer[1]) werden bei Vorliegen allgemeiner Werte (reelle Zahlen) einer kontinuierlichen Variable die Klassengrenzen nach einem vorgegebenem Algorithmus ermittelt. Nennen wir die gegebenen Zahlen $x_1, x_2, ..., x_n$. Die Anzahl von Klassen k wird auf Grund der Größe n der Datenmenge bestimmt (verschiedene Formeln findet man in der Literatur, z.B. $\sqrt{n}$, was sinnvoll ist für etwa $n = 100$; für sehr große n muss man etwas anderes wählen). Die Klassengrenzen $g_1, g_2, ..., g_{k-1}$ werden nun gleichmäßig zwischen minimalem und maximalem Datenpunkt aufgeteilt; $g_0 = x_{min}$, $g_k = x_{max}$ und gleiche Klassenbreiten. (Alle Grenzen sollten am besten dann so modifiziert werden, dass sie möglichst vernünftig einfache Zahlenwerte annehmen.) Ein Datenwert x_i wird nun der j-ten Klasse zugeordnet, wenn gilt

$$g_{j-1} < x_i \leq g_j$$

[1] ℝ: `hist(Daten)`

mit Ausnahme von x_{min}, das in die 1. Klasse kommt.

Man bemerke die große Schwierigkeit mit der hohen Fehleranfälligkeit der Vorgangsweise ohne Computer. Zweitens reagiert das Histogramm, wie man sich leicht überlegt, sehr empfindlich auf den Anfangswert g_0 und die Klassenbreite $g_j - g_{j-1}$.

3.2.2 Strichliste, Zehnersystem

Bei der grafischen Darstellung von Hand wird die Intervalleinteilung zweckmäßigerweise vertikal (statt horizontal wie im Histogramm) aufgetragen. Üblicherweise wird jeder Wert in die entsprechende Klasse als Schrägstrich (/) eingetragen. Statt des 5. Schrägstriches werden dann die vorhergehenden 4 durchgestrichen: ////. Die Häufigkeiten können nachher abgezählt werden.

Neuerdings hat sich ein „Zehnersystem“ als günstiger (weniger fehleranfällig) erwiesen. Die ersten 4 Werte werden als Punkte in einem Quadrat eingetragen. Die nächsten 4 Werte werden als Seiten des Quadrats dargestellt. Weitere 2 Werte können als Diagonalen repräsentiert werden. Die Zahlen 1 bis 10 werden folgendermaßen dargestellt.

Beispiel 3.1: Baujahre von Autos im Anzeigenteil der Sunday Standard-Times, August 1968:

#	Baujahr	
-	1954	
1	55	•
-	56	
1	57	•
1	58	•
1	59	•
4	60	• •
10	61	⊠
14	62	⊠ •
17	63	⊠ ⌐
25	64	⊠ ⊠ ⁞
34	65	⊠ ⊠ ⊠ :
27	66	⊠ ⊠ ⌐
12	67	⊠ •
4	68	• •

Übung 3.1: Verteilung der Porosität (prozentualer Anteil des Porenraumes eines

Gesteins am Gesamtvolumen) in einem Sandstein: 57 Werte (in %) wurden an einem Teilprofil einer Bohrung gemessen.

22.1	23.5	25.3	26.6	23.9	26.0	22.8	22.3	23.1	23.0	21.0	21.8
22.0	22.2	22.3	22.4	22.4	22.4	22.3	21.6	22.1	22.6	22.1	21.9
22.3	23.9	23.2	22.5	23.7	23.3	24.4	22.6	23.9	24.2	27.6	27.9
25.2	21.7	20.0	19.8	21.5	25.6	25.3	24.1	28.6	23.7	24.0	21.8
24.9	24.2	25.0	23.7	27.3	23.0	23.8	21.2	21.1			

Man untersuche die Verteilung an Hand einer Strichliste, des Zehnersystems und eines Histogramms.

3.2.3 Häufigkeitstabelle und Summenhäufigkeitspolygon

Häufigkeiten (absolute und relative) der Werte pro Klasse werden gerne in einer Tabelle eingetragen. Summenhäufigkeiten können in einer weiteren Spalte daneben geschrieben werden, wobei meistens von der niedrigsten Klasse an durchgezählt wird. Dabei ergeben sich *steigende Ränge* der Daten. *Fallende Ränge* können analog konstruiert werden. (Es gilt: Steigender Rang + fallender Rang = 1 + Gesamtanzahl der Daten)

Die *Tiefe* eines Datums ist immer das Minimum des steigenden und fallenden Ranges, d.h. etwa der kürzeste Abstand zu den Extremen.

Die klassische Häufigkeitstabelle mit den Daten der Zugriffszeiten (aus Tabelle 3.1, auf ms gerundet) wird in Tabelle 3.2 dargestellt.

Ein *Summenhäufigkeitspolygon* wird durch Verbindung der prozentualen Summenhäufigkeit an den rechten Endpunkten der Intervalle konstruiert. Die Daten aus Tabelle 3.2 ergeben Abbildung 3.5.

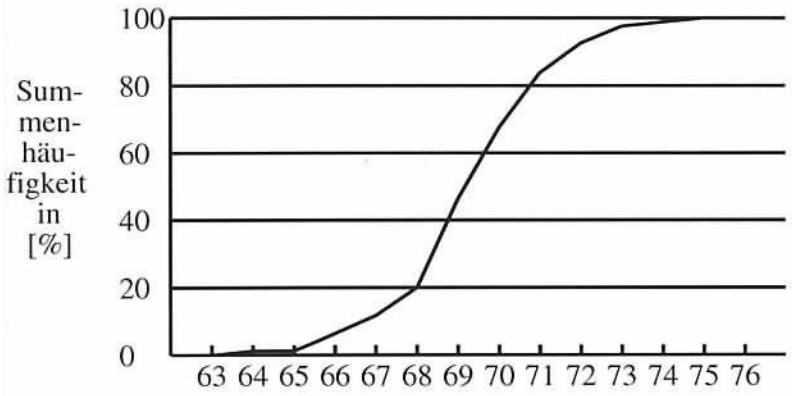


Abbildung 3.5: Summenhäufigkeitspolygon.

Übung 3.2: Man berechne die relativen Häufigkeiten der Daten aus der Übung auf Seite 21. Weiters zeichne man ein Summenhäufigkeitspolygon.

Tabelle 3.2: Häufigkeitstabelle der Zugriffszeiten.

Zeit (mittl. Punkt)	Häufigkeit	%	Kumulierung Häufigkeit	%
76				
75	1	1.25	80	100.00
74	1	1.25	79	98.75
73	3	3.75	78	97.50
72	8	10.00	75	93.75
71	12	15.00	67	83.75
70	16	20.00	55	68.75
69	22	27.50	39	48.75
68	9	11.25	17	21.25
67	5	6.25	8	10.00
66	2	2.50	3	3.75
65			1	1.25
64	1	1.25	1	1.25
63			0	0
Summe	80	100.00		

3.3 Kenngrößen von Verteilungen

Verteilungen können auch (zumindest in einem gewissen Maße) durch bestimmte Größen charakterisiert werden. Dazu gehören Kenngrößen für das Mittel (Ortsparameter), die Streuung (Streuungsparameter) sowie höhere Momente.

3.3.1 Ortsparameter

Die Lokation der Verteilung einer statistischen Variablen kann meist durch eine Größe angegeben werden. Am häufigsten wird wohl das *arithmetische Mittel* verwendet. Es seien n Beobachtungen (bezeichnet mit x_i, $i = 1, \ldots, n$) gegeben. Dann ist das arithmetische Mittel $\bar{x}$ definiert als[2]

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

Interpretiert man die Variablenwerte als gleich große Gewichte, so stellt $\bar{x}$ gerade den *Schwerpunkt* dar.

Bei Vorliegen einer Klasseneinteilung der Daten (wie in einem Histogramm) findet man $\bar{x}$ schneller über die Häufigkeiten. Bezeichnen wir die Klassenmitten mit c_j, die entsprechenden (absoluten) Häufigkeiten mit h_j und die relativen Häufigkeiten

[2] R: `mean(Daten)`

mit f_j. Wenn es k Klassen gibt, berechnet sich $\bar{x}$ approximativ als

$$\bar{x} \cong (\sum_{j=1}^{k} h_j c_j) / \sum_{j=1}^{k} h_j = \sum_{j=1}^{k} f_j c_j \quad .$$

Da das arithmetische Mittel auch den Schwerpunkt der Verteilung angibt, wird es sehr stark von Werten, die von der Masse der Daten weit entfernt liegen, beeinflusst. Ein realistischerer Ortsparameter, der eher das „Zentrum" der Verteilung angibt (und hiermit viel stabiler, *robuster* gegenüber Änderungen von Daten ist), ist der *Median* oder *Zentralwert*. Wir bezeichnen ihn mit $\tilde{x}$, und er ist durch jenen Wert definiert, der die Menge der Beobachtungen in 2 gleiche Teile teilt: höchstens 50% der Werte sind kleiner als $\tilde{x}$, und höchstens 50% größer. In Häufigkeitstabellen findet man den Wert einfach durch Abzählen. Im Summenhäufigkeitspolygon liest man $\tilde{x}$ an jener Stelle der Abszisse ab, wo das Polygon gleich $\frac{1}{2}$ ist.

Andere Kenngrößen für den Ort der Verteilung sind Quantile, der Modalwert[3], die Mitte des Bereichs etc.

Quantile und Perzentile

Das α-*Quantil* Q_α ist definiert durch jenen Wert, für den ein α-Anteil der Daten kleiner oder gleich *und* ein $(1-\alpha)$-Anteil größer oder gleich Q_α ist[4][5]. *Perzentile* P_α sind ähnlich definiert, nur mit Prozent-Angaben. So ist das 10%-Perzentil unserer Verteilung der Zugriffszeiten auf Daten (Tabelle 3.1) gleich 67.5 ms.[6]

.25- und .75-Quantile heißen auch *Quartile*. Das .5-Quantil bezeichnet gleichzeitig den Median. Man sollte sich überlegen, wie Quantile vom Summenhäufigkeitspolygon leicht abgelesen werden können.

Übung 3.3: Man bestimme $\bar{x}$, $\tilde{x}$, Modalwert, Mittelpunkt des Bereichs, $Q_{.25}, Q_{.75}$ der Daten aus Tabelle 3.1 und der Übung auf Seite 21.

3.3.2 Streuungsmaße

Der Ortsparameter allein gibt kaum ein klares Bild der Verteilung. Die Abbildung 3.6 stellt verschiedene Verteilungen dar, die alle das gleiche arithmetische Mittel haben, wobei das Verhalten der Daten um dieses Mittel aber sehr verschieden sein kann.

[3] Häufig ist der Modalwert so definiert: Mitte jener Klasse mit größter Häufigkeit; gibt es mehrere Klassen mit der höchsten Häufigkeit (Nichteindeutigkeit!), so wird die Mitte dieser genommen, soferne sie nebeneinander liegen, sonst ist der Modalwert undefiniert.

[4] Praktisch werden Quantile q_α oft so berechnet:

$$\text{Ist } n\alpha = \begin{cases} \text{ganzzahlig} & \Rightarrow \quad q_\alpha = \frac{x_{(n\alpha)} + x_{(n\alpha+1)}}{2} \\ \text{nicht ganzzahlig} & \Rightarrow \quad q_\alpha = x_{(\lfloor n\alpha \rfloor + 1)} \end{cases} ,$$

wobei $x_{(i)}$ der i-te geordnete Wert und $\lfloor . \rfloor$ Abrundung bedeutet.

[5] ℝ gibt sogar 9 Varianten der Berechnung von q_α als Optionen an!

[6] ℝ: `quantile(Daten,alpha)`

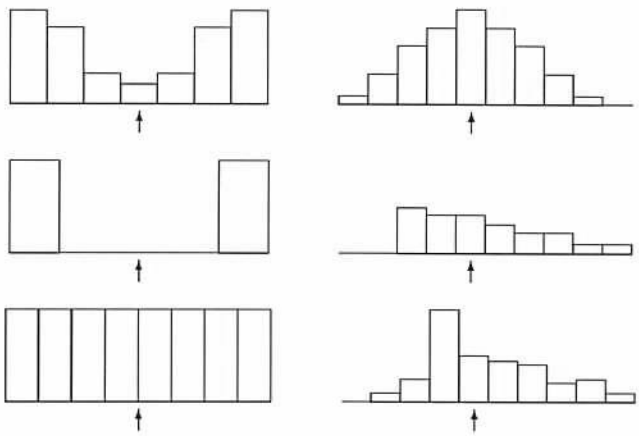


Abbildung 3.6: Verschiedene Verteilungen bei gleichem Mittel.

Eine andere Kenngröße für die Verteilung sollte die Variation der Werte um den Ortsparameter angeben. Man könnte eine Funktion der Abweichungen vom Mittel

$$x_i - \bar{x}$$

nehmen. Allerdings ist – wie man sich leicht überlegt – die Summe und damit das Mittel dieser Abweichungen immer gleich null. Häufig wird daher das Quadrat der Abweichungen betrachtet, und das Mittel dieser Abweichungsquadrate

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

bezeichnet man als *Varianz*[7,8]. Die *Standardabweichung* oder *Streuung* (*standard deviation*[9]) ist definiert durch

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

Zur Berechnung von s ist es leichter, die Formel

$$\sum(x_i - \bar{x})^2 = \sum x_i^2 - n(\bar{x})^2$$

[7]Man bemerke die Division durch $n-1$ statt n: $n-1$ heißt auch *Freiheitsgrad*. Man nimmt an, dass n unabhängige Beobachtungen x_i vorliegen (n Freiheitsgrade). Bei den Differenzen $(x_i - \bar{x})$ geht aber 1 Freiheitsgrad verloren, weil $\bar{x}$ aus den Beobachtungen ausgerechnet wurde und hiermit eine Differenz schon aus den anderen Differenzen ausgerechnet werden kann, weil gilt: $\sum(x_i - \bar{x}) = 0$. Die mathematische Begründung liegt allerdings in der Erwartungstreue (siehe Seite 73).

[8]R: `var(Daten)`

[9]R: `sd(Daten)`

zu verwenden, wobei dieser Ausdruck allerdings zu großen numerischen Instabilitäten führen kann. Wenn man eine Klasseneinteilung der Daten in k Klassen mit c_j als Klassenmitten und h_j als Häufigkeiten zur Verfügung hat, gilt ungefähr

$$\sum_{i=1}^{n} x_i^2 \cong \sum_{j=1}^{k} c_j^2 h_j \quad .$$

Als *Variationskoeffizient* bezeichnet man das Verhältnis

$$v = \frac{s}{\bar{x}} \qquad (\mid \bar{x} \mid > 0) \quad .$$

Lineare Transformationen : Werden die Daten durch

$$y_i = ax_i + b$$

mit $a \neq 0$ transformiert, so errechnet sich das Mittel von y_i als

$$\bar{y} = \frac{1}{n}\sum(ax_i + b) = a\bar{x} + b$$

und die Varianz als

$$s_y^2 = \frac{1}{n-1}\sum(y_i - \bar{y})^2 = \frac{a^2}{n-1}\sum(x_i - \bar{x})^2 = a^2 s_x^2$$

bzw. die Standardabweichung als

$$s_y = a s_x \quad .$$

Übung 3.4: Man berechne das arithmetische Mittel und die Standardabweichung der *standardisierten* Größen

$$y_i = \frac{x_i - \bar{x}}{s_x} \quad .$$

Andere Streuungsmaße : Der *interquartile Abstand* $Q_{.75} - Q_{.25}$ [10] approximiert die Standardabweichung bei Vorliegen der Normalverteilung durch

$$s_{IQR} = \frac{Q_{.75} - Q_{.25}}{1.349} \text{ [11]} \quad ,$$

der *Median der absoluten Abweichungen vom Median* (*Medmed*) durch

$$s_{Medmed} = \frac{1}{.6745} med(\mid x_i - \tilde{x} \mid) \text{ [12,13]} \quad .$$

Weiters kann man sich als Maß für die Streuung allgemein den Abstand zwischen zwei Quantilen vorstellen. Als extreme Statistik dieser Art wird oft die *Spannweite* (Maximal- minus Minimalwert) verwendet.

Übung 3.5: Man berechne eine Schätzung für die Streuung der Daten aus Tabelle 3.1 und der Übung auf Seite 21 durch den interquartilen Abstand und den Medmed.

[10] Man bemerke, dass diese Größe wieder relativ stabil (*robust*) ist, weil sie vom Betrag von bis zu 25% der Beobachtungen nicht wesentlich beeinflusst wird.

[11] R: `IQR(Daten)/1.349`

[12] Bei dieser Berechnung geht die *Robustheit* sogar auf 50% der Beobachtungen!

[13] R: `mad(Daten)`

3.3.3 Höhere Momente

Die Varianz s^2 wird auch als Moment 2. Ordnung bezeichnet. Das 3. (normierte) zentrale Moment

$$g_1 = \frac{1}{n}\sum(x_i - \bar{x})^3/s^3$$

heißt auch *Schiefe*[14] (*skewness*) und das um 3 verminderte (normierte) zentrale 4. Moment

$$g_2 = \frac{1}{n}\sum(x_i - \bar{x})^4/s^4 - 3$$

Kurtosis[15] (*Exzess, Wölbung*).

3.3.4 Daten-Zusammenfassungen

Zusammenfassungen von Daten werden mit guter Darstellung von verschiedenen Kenngrößen präsentiert. Unsere Zugriffsdaten könnten mit den schon bekannten Größen so zusammengefasst werden. Dabei geben wir auch gleich die entsprechenden R-Befehle dazu:

Kenngröße	Resultat	R
Anzahl n	80	`length(Zugriffsz)`
Minimum	64.2	`min(Zugriffsz)`
Quantil $q_{.05}$	66.885	`quantile(Zugriffsz,.05)`
1. Quartil $q_{.25}$	68.6	`quantile(Zugriffsz,.25)`
Mittel $\bar{x}$	69.734	`mean(Zugriffsz)`
Median $\tilde{x}$	69.5	`median(Zugriffsz)`
3. Quartil $q_{.75}$	70.925	`quantile(Zugriffsz,.75)`
Quantil $q_{.95}$	72.81	`quantile(Zugriffsz,.95)`
Maximum	74.9	`max(Zugriffsz)`
Varianz s^2	3.534	`var(Zugriffsz)`
Streuung s	1.88	`sd(Zugriffsz)`
s_{IQR}	1.723	`IQR(Zugriffsz)/1.349`
s_{Medmed}	1.557	`mad(Zugriffsz)`
Schiefe g_1	0.089	`skewness(Zugriffsz)`
Kurtosis g_2	0.466	`kurtosis(Zugriffsz)`

Tukey (1977) schlägt eine kurze 5-Zahlen-Zusammenfassung vor, wobei `M40h` für dem Median mit entsprechender Tiefe, `H20h` für die „Hinges“ (*Gelenke*) und „1“ für die Extremwerte (Tiefe 1) stehen. Er definiert dabei die Tiefe des Medians mit $t_m = \frac{n+1}{2}$, wobei n für die Anzahl der Daten steht. Die Tiefe der Gelenke ist dann einfach $\frac{\lfloor t_m \rfloor + 1}{2}$, wobei $\lfloor . \rfloor$ abrunden bedeutet:

[14] R: `skewness(Daten)` (Package e1071).
[15] R: `kurtosis(Daten)` (Package e1071).

Zugriffszeiten in [ms]		
M40h	69.5	
H20h	68.6	70.95
1	64.2	74.9

3.3.5 Boxplots

Boxplots sind grafische Zusammenfassungen von Daten. Dabei entsprechen die Enden des Rechteckes den *Hinges* und die Unterteilung dem Median. Extreme Werte (Ausreißer) können leicht eingezeichnet werden. Dazwischen werden sogenannte „Barthaare" (*whiskers*) bis zum letzten nicht ausreissenden Wert, das heißt, einem Wert mit Abstand von bis zu 1.5 mal dem Interquartilabstand von der Schachtel, durchgehend gezeichnet, und dies in beide Richtungen.

Wir illustrieren dies in Abbildung 3.7 mit geochemischen Daten Nickel und Chrom.

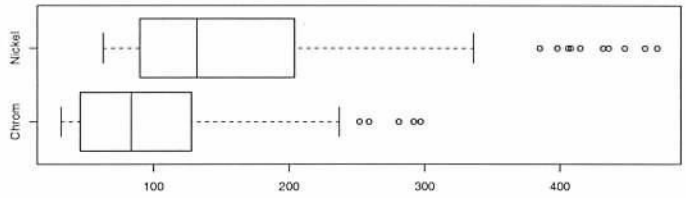


Abbildung 3.7: Boxplot von Nickel und Chrom.

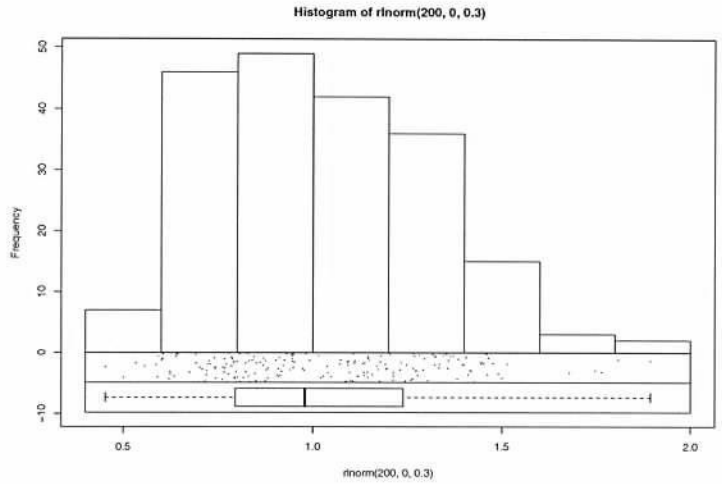
3.3.6 Illustratives Beispiel

Ein Beispiel mit 200 künstlich erzeugten Werten aus einer Lognormalverteilung (wobei für die log-transformierten Werten als Mittel = 0 und Streuung = .3 genommen wurde) sieht man in der folgenden Grafik.

3.4 Stichproben und Population

Unsere Datenmengen wurden immer als Teilmenge einer (großen oder unendlichen) übergeordneten Menge, der sogenannten *Population* oder *Grundgesamtheit*, betrachtet. Eine Population kann aus einer beliebigen Menge von Objekten oder Individuen bestehen, die irgendwelche gemeinsame, beobachtbare Charakteristiken, sogenannte *Merkmale* aufweisen. Eine Untermenge einer Population wird auch als *Stichprobe* oder kurz *Probe* bezeichnet.

In der Statistik werden meistens zufällig ausgewählte Proben oder *Zufallsstichproben* betrachtet. Die Elemente der Population haben dabei gleiche Chance, in



die Probe aufgenommen zu werden. Wenn ein ausgewähltes Element in die Population zurückgegeben wird, bevor das nächste „gezogen" wird, so spricht man vom „Stichprobenziehen mit Zurücklegen".

Der Begriff Grundgesamtheit oder Population kann sich auf die Objekte oder auch auf die Merkmale dieser selbst beziehen. Ein Element der Stichprobe wird auch als *Realisation* des Merkmals bezeichnet. Man unterscheidet zwischen der Verteilung der Werte der Stichprobe (der gemessenen, wirklich beobachteten oder empirischen Werte, daher *empirische* Verteilung) und der Verteilung der Werte der Gesamtheit (der theoretischen, aber normalerweise unbekannten und nicht beobachtbaren Verteilung). Daher ist eines der wichtigsten Probleme der Statistik, inwieweit man aus der Verteilung der Stichprobe auf die Verteilung der Grundgesamtheit schließen kann.

Eine dritte Klasse von Verteilungen wird durch die Verteilung einer Größe, die sich aus einer möglichen Stichprobe (eines bestimmten Umfangs) aus der Population errechnet, bestimmt. Zum Beispiel, wenn wir alle möglichen Stichproben von 10 Schülern einer bestimmten Schule nehmen und jeweils die mittlere Körpergröße aus jeder Stichprobe berechnen, bekommen wir eine große Anzahl von Mitteln, die eine Verteilung bestimmen, die man mit Verteilung des Mittels von Stichproben des Umfangs 10 bezeichnet („Verteilung einer Statistik").

Beispiele: (a) Eine Studie der Körpergrößen der Männer einer Stadt wird durchgeführt. Eine Stichprobe von 200 Männern wird erhoben und die Körpergrößen werden gemessen. Die Stichprobe besteht aus den 200 (zufällig) ausgewählten Männern. Die gemessene Charakteristik (das Merkmal) ist die Körpergröße. Wir können die (empirische) Verteilung der Körpergröße der 200 Männer feststellen, aber die Verteilung der gesamten Population wird nicht zu bestimmen sein. Das ausgerechnete Stichprobenmittel aus den Körpergrößen der betrachteten 200 Männer könnte als „Einzelbeobachtung" der Mittel aller möglichen Stichproben mit dem Umfang

200 interpretiert werden.

(b) Ein Forschungsprojekt zur Untersuchung der Wirkung eines bestimmten Medikamentes auf Tetanus wird durchgeführt. Eine Gruppe von Ratten, die mit Tetanus infiziert wurde, wird mit dem Medikament behandelt und der Anteil der überlebenden Ratten nach einer bestimmten Zeitperiode festgestellt. Die Population besteht aus allen Ratten, die infiziert und behandelt werden oder werden könnten. Die Stichprobe besteht aus der Gruppe der Ratten, die wirklich in diesem Projekt verwendet werden, und das Merkmal besteht (abstrakt) aus dem Überleben nach der Zeitperiode: ja oder nein. Die Population kann nicht beobachtet werden, weil man den Versuch nicht mit allen Ratten durchführen kann.

(c) Die Korngrößen eines dispersen Stoffes sollen analysiert werden. Eine Stichprobe („repräsentativ") einer bestimmten Größe (Durchmesser oder Gewicht) wird genommen und die Verteilung der Korngröße in der Probe festgestellt. Für eine *festgelegte* Korngröße könnte der Anteil der kleineren Korngrößen (gemessen als relative Anzahl, in Gewichts- oder Volumseinheiten) das Merkmal sein, der festgestellte Anteil in der Stichprobe die Realisation, und für *dieses* Merkmal liefert die Stichprobe nur einen einzigen Wert.

Abschließend sei vermerkt, dass viele Kenngrößen wie Mittelwerte und Varianz sich auch auf die Verteilungen der Grundgesamtheiten anwenden lassen, und auf die wir mit der Information aus den Stichproben schließen wollen. Dies wird aber später noch ausführlich behandelt.

Kapitel 4

Wahrscheinlichkeitstheorie

4.1 Ereignisse

Ereignisse stellen gewissermaßen Elemente der Wahrscheinlichkeitstheorie und Mathematischen Statistik dar. Man denke an das Ereignis „Verbesserung des Gesundheitszustandes“ eines Versuchstieres beim Testen eines neuen Medikamentes, an „besonders günstige Responsezeit“ (unter einem bestimmten festgesetzten Wert liegend) eines Mehrplatzsystemes.

Als klassisches Beispiel betrachten wir das Werfen von 2 Würfeln. Jemand könnte am Ereignis „2 mal eine Eins“ interessiert sein; bezeichnen wir dieses Ereignis mit D. Ein anderes interessantes Ereignis wäre „die Würfel bleiben in einem Abstand von höchstens 10 cm voneinander liegen“, welches wir mit L bezeichnen. Ein drittes Ereignis könnte sein „ein bestimmter Würfel bleibt mit nicht mehr als 10^o zum rechteckigen Tisch liegen“. Dieses Ereignis bezeichnen wir mit W.

Ein Ereignis kann also als Beschreibung eines Resultats eines Experiments, als ein bestimmtes Verhalten eines Systems oder als der übliche Begriff eines Ereignisses gelten. Ereignisse können eintreten oder nicht; eine Beschreibung kann stimmen oder nicht; das Verhalten eines Systems kann wie ein vorgegebenes sein oder nicht.

Sei x ein *Versuchsausgang*. Dann kann ein Ereignis immer in der Form „x fällt in eine bestimmte Menge A“ interpretiert werden. Das Ereignis D = „mit 2 Würfeln zwei Einser werfen“ kann *identifiziert* werden mit der Menge, die aus dem Zahlenpaar (1,1) besteht. Wir können nun alle möglichen Versuchsausgänge (hier Augenpaare) beim Werfen von 2 Würfeln betrachten und diese auf natürliche Weise zusammenfassen zur Menge $\Omega = \{(1,1),(1,2),(1,3),\ldots,(2,1),(2,2),\ldots,(6,6)\}$. Jeder Versuchsausgang kann mittels der beiden „Augenzahlen“ beschrieben und damit als Teilmenge von Ω identifiziert werden.

$\Omega = \{\ldots,\omega,\ldots\}$ heißt *Menge aller möglichen Versuchsausgänge*, und ihre einpunktigen Teilmengen $\{\omega\}$ bezeichnet man als *Elementarereignisse*. *Ereignisse* sind Teilmengen von Ω. Im obigen Beispiel gibt es die 36 Elementarereignisse $\{\omega_1\} = \{(1,1)\}, \{\omega_2\} = \{(1,2)\}, \ldots, \{\omega_{36}\} = \{(6,6)\}$. Ein Ereignis wäre z.B. die Menge $\{(1,1),(3,1),(2,5)\}$.

Zu einer anderen Fragestellung gehört das Ereignis L. Man interessiert sich nicht für die Augenzahlen der beiden Würfel, sondern für den Abstand zwischen ihnen. Das Ereignis L kann identifiziert werden mit dem Intervall [0,10]. Ω besteht hier aus allen möglichen Abständen der beiden Würfeln auf dem Tisch und kann zusammengefasst werden zum Intervall [0,d], wobei d die Länge der Diagonale des Tisches in cm bezeichnet.

Analog den Mengenoperationen spricht man von *Ereignisoperationen*, die im folgenden zusammengefasst sind:

(a) Durchschnittsbildung: Das Ereignis $C =$ „A geschnitten mit B“, bezeichnet mit $A \cap B$ oder kurz AB, bedeutet das Ereignis „beide, A *und* B“, bzw. beinhaltet alle Werte, die sowohl in A als auch in B liegen. $D \cap L$ beschreibt z.B. das Ereignis „2 mal die Augenzahl 1 *und* der Abstand der beiden Würfel beträgt höchstens 10 cm“.

(b) Vereinigungsbildung: Das Ereignis $C =$ „A vereinigt mit B“, bezeichnet mit $A \cup B$, bedeutet das Ereignis „entweder A *oder* B (oder beide)“ bzw. beinhaltet alle Werte, die in A oder in B liegen. Zum Beispiel beschreibt $D \cup L$ das Ereignis „2 mal die Augenzahl 1 *oder* der Abstand der beiden Würfel beträgt höchstens 10 cm“.

(c) Komplementbildung: Das Ereignis „Komplement von A“, bezeichnet mit A^C, bedeutet das Ereignis „*nicht* A“ bzw. beinhaltet alle Werte aus Ω, die nicht in A liegen. Zum Beispiel bedeutet L^C „der Abstand beider Würfel beträgt mehr als 10 cm“.

Als besonderes Ereignis gilt das *unmögliche* Ereignis (oder *Nullereignis*), bezeichnet mit $\varnothing$, das nie eintreten kann. Z.B. das Ereignis $A \cap A^C$ kann nie verwirklicht werden. Weiters bezeichnen wir mit Ω das *sichere (umfassende)* Ereignis, das eintreten muss, z.B. $A \cup A^C = \Omega$.

Einige weitere Begriffe:

(a) Ereignisse A und B sind *disjunkt*, wenn gilt $A \cap B = \varnothing$, wenn sie also niemals gleichzeitig eintreten können. Z.B. A und A^C sind disjunkt. Sie „schließen einander aus“, sie sind „unvereinbar“.

(b) Ein Ereignis A ist in einem anderen Ereignis B enthalten, wenn $A \cap B = A$. Wir bezeichnen dies mit $A \subset B$. D.h. das Eintreten von A impliziert das Eintreten von B.

(c) Eine Menge von Ereignissen $\{A_1, A_2, \ldots\}$ bezeichnet man als *Zerlegung* (Partition) des Ereignisses A, wenn gilt $A = A_1 \cup A_2 \ldots$ und $A_i \cap A_j = \varnothing$ für alle $i \neq j$. Zum Beispiel ist $\{DL, DL^C\}$ eine Zerlegung von D.

(d) Die Menge der betrachteten Ereignisse (Teilmengen von Ω) bezeichnet man als *Ereignisraum* $\mathfrak{A}$. Um eine bestimmte Vollständigkeit zu haben, sollte dieser

auch Durchschnitte und Vereinigungen von Ereignissen, so wie das unmögliche und das sichere Ereignis beinhalten. So ein System heißt auch *Ereignis-Algebra*.

(e) Durchschnitt kann durch Vereinigung und Komplement gebildet werden: A und B seien Ereignisse; dann gilt

$$A \cap B = (A^C \cup B^C)^C \quad \text{(Regel von De Morgan)} \ .$$

Analog gilt

$$A \cup B = (A^C \cap B^C)^C \ .$$

Beispiel 4.1: Bei der Konstruktion von Histogrammen wird zunächst eine Intervalleinteilung durchgeführt: $a_0, a_1, a_2, \ldots, a_k$ bezeichnen die Intervallendpunkte. Die Menge der Versuchsausgänge Ω wäre die Menge der reellen Zahlen $\mathbb{R}$ mit den Elementarereignissen $\{r\}$. „Ein Datenpunkt x fällt in ein Intervall $(a_{i-1}, a_i]$" ist ein Ereignis, das wir mit dem Intervall $(a_{i-1}, a_i]$, einer Teilmenge von $\mathbb{R}$, identifizieren: $(a_{i-1}, a_i] = A_i$, $i \in \{1, \ldots, k\}$. Nehmen wir die Intervalle $(\infty, a_0] = A_0$ und $(a_k, \infty) = A_{k+1}$ hinzu, so erhalten wir eine Zerlegung $\mathfrak{Z} = \{A_0, A_1, \ldots, A_{k+1}\}$ von $\mathbb{R} = A_0 \cup A_1 \cup \ldots \cup A_{k+1}$. Die Elemente aus $\mathfrak{Z}$, das sichere Ereignis $\mathbb{R}$, das unmögliche $\varnothing$, so wie alle Vereinigungen aus Elementen aus $\mathfrak{Z}$ formen eine Ereignisalgebra.

Beispiel 4.2: Ein Physiker zählt α-Teilchen von einer radioaktiven Quelle etwa in einer 7.5 Sekunden-Periode. E_i bezeichnet das Ereignis „i Teilchen in dieser Zeit-Periode". Der Physiker interessiert sich für das Ereignis „mehr als 5 Teilchen in 7.5 Sekunden", was man symbolisch als $E_6 \cup E_7 \cup E_8 \cup \ldots$ schreibt. Wir müssen also die Operation unendlich oft anwenden. Man spricht dann von einer Ereignis-σ-Algebra.

Sei $\Omega = \{\ldots, \omega, \ldots\}$ eine Menge von möglichen Versuchsausgängen. Dann bezeichnet man ein Mengensystem $\mathfrak{A}$ von Teilmengen aus Ω als *Ereignis-σ-Algebra*, wenn es die Gesamtmenge Ω enthält und wenn gilt:

(a) Abgeschlossenheit unter Vereinigung: $A_1, A_2, \ldots$ sei eine beliebige Folge von Elementen aus $\mathfrak{A}$. Dann gibt es ein Element C aus $\mathfrak{A}$, das sich aus der Vereinigung dieser A_i ergibt, d.h. $C = \bigcup_{i=1}^{\infty} A_i$. C enthält jedes A_i; und jede andere Menge, die alle A_i enthält, enthält auch C.

(b) Abgeschlossenheit unter Komplementbildung: Sei A aus $\mathfrak{A}$, dann ist auch A^C aus $\mathfrak{A}$.

Beispiel 4.3: Zwei Würfel werden geworfen. x_1 bezeichne die Augenzahl des 1. Würfels und x_2 die des 2. Die Menge aller möglichen Versuchsausgänge oder der *Stichprobenraum* Ω ist daher

$$\Omega = \{(x_1, x_2) : x_i = 1, 2, \ldots, 6\} \ .$$

Als Ereignisraum können wir die *Potenzmenge*, die Menge aller Untermengen aus Ω, nehmen. Das Ereignis „mindestens eine 6" ist die Menge

$$M = \{(1,6), (2,6), \ldots, (6,6), (6,5), (6,4), \ldots, (6,1)\} \ .$$

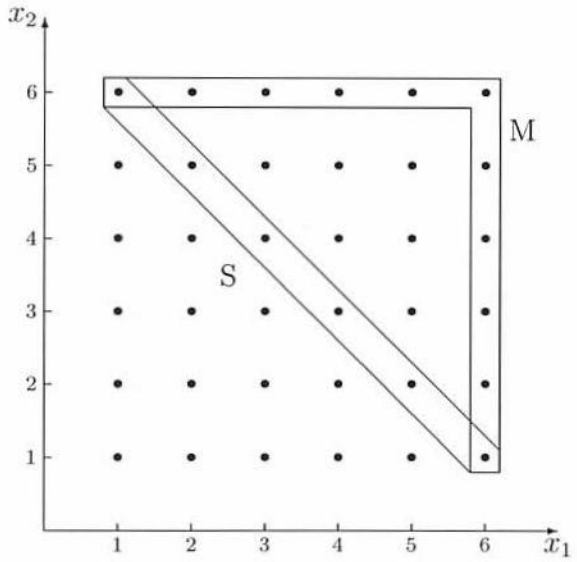


Abbildung 4.1: Elementarereignisse beim Spiel mit zwei Würfel.

Das Ereignis $S =$ „Summe der Augenzahlen gleich 7“ wäre

$$S = \{(1,6),(2,5),(3,4),(4,3),(5,2),(6,1)\} \quad .$$

Die Anzahl von Elementarereignissen in M und S sieht man am besten in der grafischen Darstellung Abbildung 4.1.

Bei endlich vielen möglichen Versuchsausgängen ist es noch sinnvoll mit allen Teilmengen von Ω zu operieren. Häufig kann man sich jedoch mit einer kleineren Klasse begnügen.

Beispiel 4.4: Ein Physiker misst elektrischen Strom. Die Menge aller möglichen Ausgänge ist

$$\Omega = \{y : y \in \mathbb{R}\} = \mathbb{R} \quad ,$$

die Zahlengerade. Den Physiker interessieren Ereignisse wie z.B. „y fällt in das Intervall (1,2]“, oder „in die Vereinigung von Intervallen, etwa (-2,-1] $\cup$ (1,2]“. Es wäre daher natürlich, für $\mathcal{A}$ die Menge aller links-halboffenen Intervalle (a,b] und alle Vereinigungen und Komplimente solcher zu nehmen. Die „kleinste“ σ-Algebra, die alle diese Mengen enthält, bezeichnet man als *Borel'sche σ-Algebra* $\mathfrak{B}$ und ihre Elemente als *Borel-Mengen.*

Betrachten wir den Fall, dass der Physiker k Messungen $y_1, \ldots, y_k$ des elektrischen Stroms durchführt. Dann ist

$$\Omega = \{(y_1, \ldots, y_k) : y_i \in \mathbb{R}\} = \mathbb{R}^k \quad ,$$

der k-dimensionale euklidische Raum der reellen Zahlen. Die interessierenden Ereignisse sind jetzt k-dimensionale Rechtecke sowie Vereinigungen und Komplimente

dieser. Wir fordern wieder Abgeschlossenheit unter Vereinigung und Komplementbildung und bezeichnen die kleinste σ-Algebra als Borel'sche σ-Algebra $\mathfrak{B}^k$.

4.2 Wahrscheinlichkeiten

Betrachten wir das Werfen von 2 Münzen. Dann ist die Menge aller möglichen Versuchsausgänge

$$\Omega = \{(x_1, x_2)\} = \{(0,0), (0,1), (1,0), (1,1)\} \quad ,$$

wobei 1 *Kopf* bedeutet und 0 das *Wappen* der Münze. Um zu erfahren, wie wahrscheinlich es ist, gleichzeitig einen Kopf und ein Wappen zu erhalten (sagen wir, Ereignis A), könnte man 100 mal die 2 Münzen werfen und die Anzahl des Eintreffens von A zählen. Wenn dies in 47 der Versuche der Fall ist, könnte man dem Eintritt dieses Ereignisses A die Wahrscheinlichkeit $P(A) = 47/100 = .47$ zuordnen, mit dem es beim nächsten Wurf zu erwarten ist (*a posteriori-Wahrscheinlichkeiten*).

Andererseits kann man ein theoretisches Argument anwenden: Wenn kein Grund zur Annahme einer Asymmetrie der Münze besteht, müsste eine Münze ungefähr gleich oft auf der einen Seite wie auf der anderen zu liegen kommen. Die Wahrscheinlichkeit des Auftretens einer bestimmten Seite wäre demnach $\frac{1}{2}$. In unserer Menge Ω gibt es 4 Elemente, von denen aus „Symmetriegründen" keines ausgezeichnet ist, so dass jedem die gleiche Wahrscheinlichkeit, nämlich $\frac{1}{4}$, zugeschrieben werden muss. Das Ereignis oder die Menge A beinhaltet aber 2 Punkte, nämlich (0,1) und (1,0), sodass der Eintritt von A mit Wahrscheinlichkeit $\frac{1}{4} + \frac{1}{4} = \frac{1}{2}$ zu erwarten ist (*a priori-Wahrscheinlichkeiten*).

Man könnte sich auch jedes Element von Ω mit einem Gewicht (einer Wichtigkeit) versehen vorstellen, von dem dann die Wahrscheinlichkeit abgeleitet wird. Bei gleichen Gewichten kann man oft die Wahrscheinlichkeiten der Ereignisse durch einfaches Abzählen finden (siehe die Zeichnung im Beispiel auf Seite 34). Allgemein spricht man von einem *Maß*, das jedem Ereignis A aus $\mathcal{A}$ eine nicht-negative, reelle Zahl μ(A) zuordnet und das bestimmte Eigenschaften aufweist. Formal heißt das:

Ein *Maß* μ ist eine Funktion vom Ereignisraum in $[0, \infty]$, wobei, wenn $\{A_1, A_2, \ldots\}$ eine Zerlegung von A darstellt, gilt (σ-Additivität)

$$\mu(A) = \sum_{i=1}^{\infty} \mu(A_i) \quad .$$

Gilt außerdem $\mu(\Omega) = 1$, dann spricht man von einem *Wahrscheinlichkeitsmaß*, das wir mit P bezeichnen.

Aus der σ-Additivität folgt natürlich auch die endliche Additivität. Weiters gilt trivialerweise $P(\varnothing) = 0$.

Wenn wir – wie im einführenden Beispiel – jeden Punkt ω aus Ω mit einem Gewicht p(ω) versehen, wobei das totale Gewicht gleich 1 ist, d.h. $\sum p(\omega) = 1$,

dann gilt

$$P(A) = \sum_{\omega \in A} p(\omega) \quad .$$

Sind die Gewichte in einem endlichen Raum $\Omega = \{\omega_1, \ldots, \omega_n\}$ alle gleich (*diskrete Gleichverteilung*: $p(\omega_i) = \frac{1}{n}$), dann gilt für alle $A \in \mathcal{U}$

$$P(A) = P(\bigcup_{\omega_i \in A} \{\omega_i\}) = \sum_{\omega_i \in A} \frac{1}{n} = \frac{1}{n} \#(\omega_i \in A)$$

$$= \frac{\#(\text{günstige Elementarereignisse})}{\#(\text{mögliche Elementarereignisse})} \quad .$$

Die Menge der möglichen Versuchsausgänge Ω, die darauf definierte σ-Algebra von Ereignissen $\mathcal{U}$ und das Wahrscheinlichkeitsmaß P auf $\mathcal{U}$ bilden einen *Wahrscheinlichkeitsraum* $(\Omega, \mathcal{U}, P)$.

Beispiel 4.5: Stetige Gleichverteilung (Rechtecksverteilung) $R(0,1)$: Es sei $\Omega = (0,1]$ ein Intervall aus $\mathbb{R}$ und $\mathcal{Z}$ die Borel'sche-σ-Algebra. Um ein Wahrscheinlichkeitsmaß auf $\mathcal{Z}$ zu definieren, brauchen wir nur Wahrscheinlichkeiten für Intervalle anzugeben. Wir definieren für ein Intervall $I = (a, b]$ mit $0 \leq a < b \leq 1$ die Wahrscheinlichkeit $P(I) = b - a$, also gleich der Länge des Intervalls. Nehmen wir eine Zerlegung von $\Omega = A \cup B$ mit $A = (0, \frac{1}{2}]$ und $B = (\frac{1}{2}, 1]$, so gilt offensichtlich

$$1 = P(\Omega) = P(A) + P(B) \quad .$$

Eine Zerlegung von A wäre $A = A_2 \cup A_3 \cup A_4 \cup \ldots$ mit $A_n = (\frac{1}{n+1}, \frac{1}{n}]$. Wir finden, dass

$$\frac{1}{2} = P(A) = \sum_{n=2}^{\infty} P(A_n) = \sum_{n=2}^{\infty} (\frac{1}{n} - \frac{1}{n+1}) = (\frac{1}{2} - \frac{1}{3}) + (\frac{1}{3} - \frac{1}{4}) + \ldots = \frac{1}{2} \quad .$$

Regeln für das Rechnen mit dem Wahrscheinlichkeitsmaß:

(1) Aus der Additivität folgt

$$P(A) + P(A^c) = P(\Omega) = 1$$

oder

$$P(A) = 1 - P(A^c) \quad .$$

(2) Für $A - B = A \cap B^c$ gilt

$$P(A - B) = P(A) - P(A \cap B) \quad .$$

(3) Aus $A \supset B$ folgt

$$P(A) \geq P(B) \quad .$$

(4) Allgemeiner *Additionssatz*:

$$P(A \cup B) = P(A) + P(B) - P(A \cap B) \quad .$$

Beispiel 4.6: Wir betrachten wieder das Werfen von 2 Würfeln und nehmen an, dass jedes Augenpaar mit gleicher Wahrscheinlichkeit auftritt, nämlich

$$P(\{(x_1, x_2)\}) = 1/36, \qquad \text{wobei} \qquad x_i \in \{1, \ldots, 6\} \ .$$

Daher gilt $P(D) = P(\{(1,1)\}) = 1/36$ und P(„nicht gleichzeitig 1“) $= P(\{(1,1)\}^C)$ $= 1 - 1/36 = 35/36$. Es sei F das Ereignis „die erste Augenzahl ist 1“, d.h. $F = \{(1,1),(1,2),(1,3),(1,4),(1,5),(1,6)\}$, und Z = „die zweite Augenzahl ist 1“. Es gilt offensichtlich $P(F) = P(Z) = 1/6$. Weiters

$$F \cap Z = \{(1,1)\} \ ,$$

$$F - Z = F \cap Z^c = \{(1,2),(1,3),(1,4),(1,5),(1,6)\} \ ,$$

$$P(F - Z) = P(F) - P(F \cap Z) = \frac{1}{6} - \frac{1}{36} = \frac{5}{36} \ ,$$

$$P(M) = P(F \cup Z) = P(F) + P(Z) - P(F \cap Z) = \frac{1}{6} + \frac{1}{6} - \frac{1}{36} = \frac{11}{36} \ .$$

Beispiel 4.7: Bernoulli-Verteilung $Bi(1,p)$. Ω besteht nur aus 2 Elementen $\omega_1 = 0$, $\omega_2 = 1$. Das Wahrscheinlichkeitsmaß ist gegeben durch

$$P(\{\omega = 1\}) = p \ .$$

Daraus folgt $P(\{\omega = 0\}) = 1-p$. Praktische Anwendung: Überleben eines Versuchstieres bei einem pharmazeutischen Experiment. Anteil der Großkunden größer als ein bestimmter Wert.

4.3 Bedingte Wahrscheinlichkeit und Unabhängigkeit

Bis jetzt wurden Wahrscheinlichkeiten nur auf Grund der Kenntnis des Wahrscheinlichkeitraumes definiert. Manchmal ist es aber leichter die Wahrscheinlichkeit des Eintretens eines Ereignisses zu definieren, wenn man weiß, dass ein anderes bereits eingetreten ist.

Beispiel 4.8: Betrachten wir wieder das Würfeln mit 2 Würfeln, wobei $\Omega = \{(x_1, x_2) \ : x_i = 1, \ldots, 6\}$ und x_i bezeichnet die Augenzahl des i-ten Würfels. Die Ereignisse A und B seien definiert durch $A =$ „$x_1 = 2$“ und $B =$ „$x_1 + x_2 \geq 7$“.

Es gilt $P(A) = 1/6, P(B) = 21/36$ und $P(A \cap B) = 2/36$. Was ist die Wahrscheinlichkeit des Eintritts von B, wenn wir schon wissen, dass A eintritt? Wir schreiben es als $P(B \mid A)$ (B bedingt durch A). Diese Wahrscheinlichkeit ist $P(A \cap B)$, allerdings normiert auf unseren jetzt eingeschränkten Raum der möglichen Versuchsausgänge, nämlich $(\Omega \mid A) = A = \{(x_1, x_2) : x_1 = 2, \ x_2 = 1, \ldots, 6\}$. Wir erhalten

$$P(B \mid A) = \frac{P(A \cap B)}{P(A)} = \frac{2/36}{1/6} = \frac{2}{6} = \frac{1}{3} \ .$$

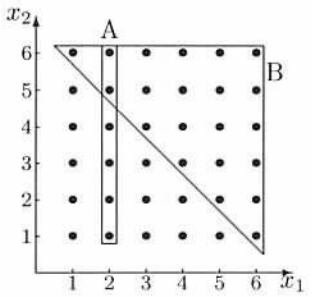


Abbildung 4.2: Bedingte Ereignisse.

Allgemein definieren wir für 2 Ereignisse A und H aus $\mathcal{A}$ mit $P(H) > 0$ die *bedingte Wahrscheinlichkeit von A unter H*

$$P(A \mid H) = \frac{P(A \cap H)}{P(H)} \quad .$$

$P(. \mid H)$ ist natürlich wieder ein Wahrscheinlichkeitsmaß.

Beispiel 4.9: Man wirft mit 2 Münzen. A sei das Ereignis „beide Münzen zeigen die gleiche Seite"; $B =$ „*mindestens* eine Münze zeigt Wappen". Es gilt offensichtlich

$$P(A) = 1/2, P(B) = 3/4, P(A \cap B) = 1/4$$

und

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)} = \frac{1/4}{3/4} = \frac{1}{3} \quad .$$

Die Definition der bedingten Wahrscheinlichkeit lässt sich für eine *Multiplikationsregel* für Wahrscheinlichkeiten verwenden. Es gilt, wenn $P(A)$ und $P(B)$ größer 0 sind,

$$P(A \cap B) = P(A)P(B \mid A) = P(B)P(A \mid B) \quad .$$

Diese Regel lässt sich leicht auf endlich viele Ereignisse verallgemeinern. Sei

$$A_1, A_2, \ldots, A_n \in \mathcal{A} \text{ mit } P(\bigcap_{i=1}^{n-1} A_i) > 0 \quad .$$

Dann gilt

$$P(\bigcap_{i=1}^{n} A_i) = P(A_1)P(A_2 \mid A_1)P(A_3 \mid A_1A_2) \ldots P(A_n \mid \bigcap_{i=1}^{n-1} A_i) \quad .$$

Beispiel 4.10: Es werden 6 Würfel geworfen. Wie groß ist die Wahrscheinlichkeit, dass 6 verschiedene Augenzahlen auftreten? $A_1, \ldots, A_6$ bezeichnen Ereignisse,

wobei A_1 = „der erste Würfel zeigt irgendeine Zahl“, A_i = „der i-te Würfel zeigt eine andere Zahl als die anderen davor“ ($i = 2, \ldots, 6$). Es gilt:

$$P(A_1) = 1 \text{ (trivial)}, \quad P(A_2 \mid A_1) = \frac{5}{6} \quad ,$$

$$P(A_3 \mid A_1 \cap A_2) = \frac{4}{6}, \quad \ldots, \quad P(A_6 \mid \bigcap_{i=1}^{5} A_i) = \frac{1}{6} \quad .$$

Daraus folgt die Wahrscheinlichkeit unseres Ereignisses

$$P(\bigcap_{i=1}^{6} A_i) = \frac{6}{6}.\frac{5}{6}.\frac{4}{6}.\frac{3}{6}.\frac{2}{6}.\frac{1}{6} = \frac{6!}{6^6} = .015 \quad .$$

Wenn für zwei Ereignisse A und B ($P(B) > 0$) gilt

$$P(A \mid B) = P(A) \quad ,$$

also, dass das bedingte Ereignis die gleiche Wahrscheinlichkeit wie das nichtbedingte aufweist, dann sind A und B *unabhängig*. Eine häufig verwendete Regel leitet sich davon ab: Bei A unabhängig von B gilt

$$P(A \cap B) = P(A)P(B) \quad .$$

Beispiel 4.11: Es werden 2 Würfel geworfen. Wie groß ist die Wahrscheinlichkeit, dass der erste Würfel eine gerade (Ereignis A) *und* der zweite eine ungerade Zahl liefert (Ereignis B)? Unter der Annahme der Unabhängigkeit erhalten wir

$$P(A \cap B) = P(A)P(B) = \frac{1}{2}.\frac{1}{2} = \frac{1}{4} \quad .$$

Beispiel 4.12: Binomialverteilung $Bi(n, p)$. Es werden n unabhängige Versuche durchgeführt, von denen jeder Versuchsausgang ähnlich wie in einem Beispiel auf Seite 37 mit der Wahrscheinlichkeit p gut (1) und mit $1 - p$ schlecht (0) ist. Für alle n Versuche stellt sich Ω als kartesisches Produkt dar:

$$\Omega = \{0, 1\}^n = \{(x_1, \ldots, x_n) : x_i \in \{0, 1\}\} \quad .$$

Wir interessieren uns für das Ereignis E_i = „# der guten Versuchsausgänge gleich i“ für $i = 0, \ldots, n$. Nachdem die n Versuche *unabhängig* voneinander durchgeführt werden, ist die Wahrscheinlichkeit, dass die ersten i Versuche gut und die restlichen $n - i$ Versuche schlecht ausgehen, gleich

$$\left(\prod_{j=1}^{i} P(x_j = 1)\right)\left(\prod_{j=i+1}^{n} P(x_j = 0)\right) = p^i(1-p)^{n-i} \quad .$$

Es gibt aber $\binom{n}{i}$ Möglichkeiten der Reihenfolge der Versuchsausgänge, bei denen das gleiche Ereignis E_i eintritt. Daher bekommen wir

$$P(E_i) = \binom{n}{i} p^i (1-p)^{n-i} \quad .$$

Eine andere Vorgangsweise wäre die Reduzierung der Menge der möglichen Versuchsausgänge auf

$$\Omega_1 = \{0, 1, 2, 3, \ldots, n\} \quad ,$$

in der als Elementarereignisse nur die Summen der guten Versuchsausgänge angegeben sind. Die Wahrscheinlichkeit des Eintritts eines Elementarereignisses $\{\omega = i\}$ für $i = 0, 1, \ldots, n$ ist dann

$$P_1(\{\omega = i\}) = \binom{n}{i} p^i (1-p)^{n-i} \quad .$$

Der neue Wahrscheinlichkeitsraum besteht aus Ω_1, der Potenzmenge $\mathcal{A}_1$ auf Ω_1 und dem entsprechenden Wahrscheinlichkeitsmaß P_1. Man sieht, dass eine Änderung der (Grund-) Menge Ω eine große Vereinfachung bringen kann. Der funktionelle Zusammenhang besteht über eine *Zufallsvariable*, die Thema des nächsten Abschnitts ist.

Beispiel 4.13: Anwendung der Binomialverteilung in der Korngrößenanalyse bei der „vollkommenen Trennung" eines Kornguts in Fein- und Grobgut. Sei p der relative Anteil des Feinguts (Durchmesser kleiner als d) am gesamten Kornkollektiv. Wird eine Stichprobe der Größe n (n Körner) gezogen, so ist

$$\begin{aligned} & P(\text{„}i \text{ Körner mit Durchmesser kleiner als } d\text{"}) \\ = \; & P(\text{„der Anteil Feingut in der Stichprobe ist } \tfrac{i}{n}\text{"}) \\ = \; & \binom{n}{i} p^i (1-p)^{n-i}. \end{aligned}$$

Z.B. $n = 100, \; p = .75$

$$\begin{aligned} \text{P(„0 Körner")} &= 0.6 \text{ x } 10^{-60} \\ 1 &= 1.9 \text{ x } 10^{-58} \\ 2 &= 2.8 \text{ x } 10^{-56} \\ . &= . \\ . &= . \\ . &= . \\ 75 &= 0.092 \\ . &= . \\ . &= . \\ . &= . \\ 98 &= 1.8 \text{ x } 10^{-10} \\ 99 &= 1.1 \text{ x } 10^{-11} \\ 100 &= 3.2 \text{ x } 10^{-13} \end{aligned}$$

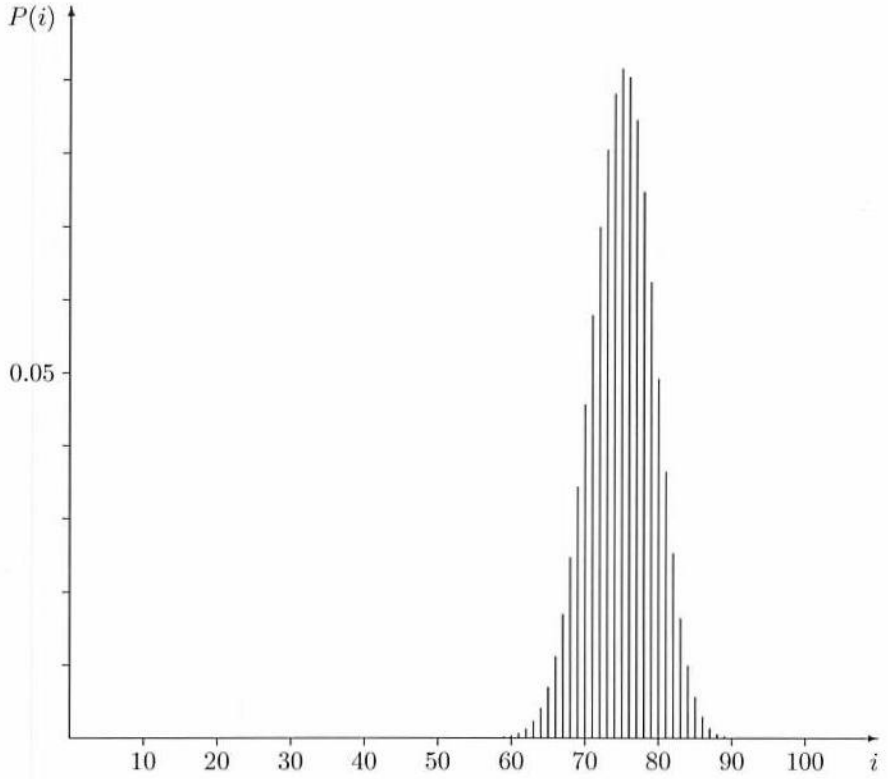


Abbildung 4.3: Wahrscheinlichkeitsfunktion der Verteilung $Bi(100, .75)$.

[1]

4.4 Zufallsvariable

Im Beispiel auf Seite 39 haben wir gesehen, dass es von Vorteil sein kann, von einem Wahrscheinlichkeitsraum $(\Omega, \mathcal{A}, P)$ in einen anderen überzugehen, in dem man leichter arbeiten kann. Insbesondere versucht man, die reellen Zahlen $\mathbb{R}$ zu verwenden, die dann die Menge der möglichen Versuchsausgänge darstellen. Die Borel'schen Mengen aus $\mathbb{R}$ bilden die Ereignisalgebra und der funktionelle Zusammenhang wird durch eine *Zufallsvariable* X gebildet, wenn jede Menge B aus $\mathfrak{B}$ ein Urbild $A = X^{-1}(B)$ als Element aus $\mathcal{A}$ in $(\Omega, \mathcal{A}, P)$ besitzt. Formal heißt das
$X : (\Omega, \mathcal{A}, P) \to (\mathbb{R}, \mathfrak{B}, P_1)$,
eine Zufallsvariable X stellt eine Abbildung von $(\Omega, \mathcal{A})$ in $(\mathbb{R}, \mathfrak{B})$ dar, wobei für jedes $B \in \mathfrak{B}$ gilt $X^{-1}(B) = \{\omega \mid X(\omega) \in B\} = A \in \mathcal{A}$.

[1] R: `plot(0:100,dbinom(0:100,100,.75),type='h')`

Beispiel 4.14: Eine einfache Zufallsvariable wird durch die Indikatorfunktion $X = I_A$ für ein $A \in \mathcal{U}$ definiert:

$$I_A(\omega) = \begin{cases} 0 & \text{für} \quad \omega \notin A \\ 1 & \text{für} \quad \omega \in A. \end{cases}$$

Es genügt, die inverse Funktion für die Intervalle $(-\infty, x]$ zu überprüfen:

$$I_A^{-1}(-\infty, x] = \begin{cases} \varnothing & \text{für} \quad x < 0 \\ A^c & \text{für} \quad 0 \leq x < 1 \\ \Omega & \text{für} \quad 1 \leq x. \end{cases}$$

Beispiel 4.15: Wie im Beispiel auf Seite 39 werden n unabhängige Versuche betrachtet, wobei für $i = 1, \ldots, n$ die Wahrscheinlichkeit des Eintrittes des Ereignisses E_i, dass der i-te Versuch positiv ausgeht, gleich $P(E_i) = p$ ist. Die Zufallsvariable, die die Anzahl der positiven Versuchsausgänge angibt, kann durch die Summe der Indikatorvariablen angegeben werden, nämlich,

$$X(\omega) = \sum_{i=1}^{n} I_{E_i}(\omega) \quad .$$

(ω stellt das n-dimensionale Ergebnis aller Versuchsausgänge dar.)

Jede Borel'sche Menge kann durch Intervalle der Form $(-\infty, x]$ gebildet werden. Deshalb definiert die *Verteilungsfunktion*

$$F(x) = P(X \leq x) = P(\omega \in X^{-1}(-\infty, x])$$

vollständig das Wahrscheinlichkeitsmaß P auf $(\mathbb{R}, \mathcal{B})$. Es gilt für jedes Intervall mit $a < b$

$$P((a, b]) = F(b) - F(a) \quad .$$

Beispiel 4.16: $Bi(10, .5)$. Die Verteilungsfunktion wird in Abbildung 4.4 dargestellt.

Eigenschaften der Verteilungsfunktion:

(a) F ist monoton wachsend (nicht fallend).

(b) F ist rechtsseitig stetig.

(c) $\lim_{x \to -\infty} F(x) = F(-\infty) = 0$,
$\lim_{x \to \infty} F(x) = F(\infty) = 1$.

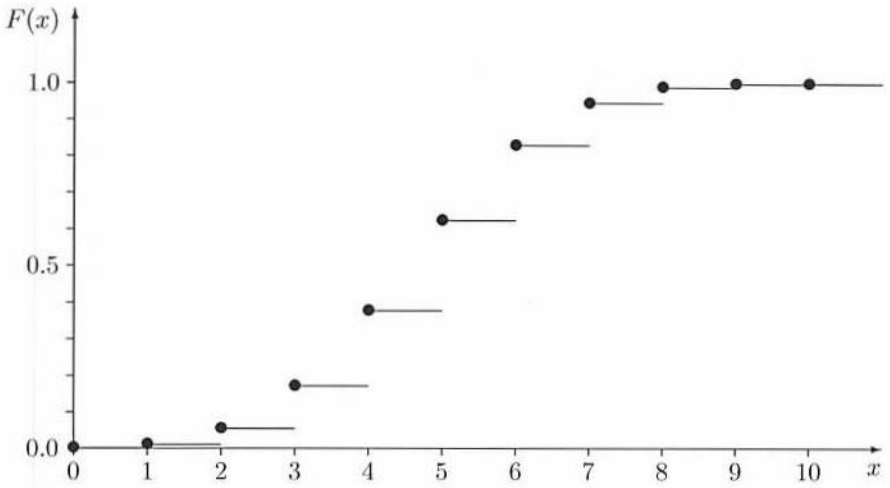


Abbildung 4.4: Verteilungsfunktion der Verteilung $Bi(10, .5)$.

4.4.1 Diskrete Zufallsvariable

Eine Zufallsvariable heißt *diskret*, wenn sie höchstens abzählbar viele verschiedene Werte annehmen kann, d.h.

$$X(\omega) = x_i,\ x_i \in \mathbb{R},\ i = 1, 2, \dots \ .$$

Bezeichnen wir

$$p_i = P(X(\omega) = x_i),\ i = 1, 2, \dots \ ,$$

dann gilt

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

und die Verteilungsfunktion ist

$$F(x) = P(X \leq x) = \sum_{x_i \leq x} P(X = x_i) = \sum_{x_i \leq x} p_i \ .$$

p_i bezeichnet man als *Wahrscheinlichkeitsfunktion.*[2]

Beispiel 4.17: Werfen eines Würfels. ω_i bezeichne den Versuchsausgang „Augenzahl i". Dann definieren wir die Zufallsvariable

$$X : \omega_i \to i \qquad \text{mit} \quad P(\omega_i) = P(X = i) = p_i = \frac{1}{6} \ .$$

Die Verteilungsfunktion wird daher

[2] R: `plot(0:10, pbinom(0:10,10,.5))`

$$F(x) = \sum_{i \le x} P(X = i) = \begin{cases} 0 & \text{für} \quad x < 1 \\ \frac{i}{6} & \text{für} \quad i \le x < i+1, \quad i = 1, \ldots, 5 \\ 1 & \text{für} \quad x \ge 6 \ . \end{cases}$$

Beispiel 4.18: Poissonverteilung $P(\lambda)$. Die möglichen Werte der Zufallsvariablen X sind $x_i = 0, 1, 2, \ldots$. Die Wahrscheinlichkeitsfunktion ist durch

$$p_i = \frac{\lambda^i}{i!} e^{-\lambda}, \ i = 0, 1, 2, \ldots$$

für ein gewisses $\lambda > 0$ definiert. Natürlich muss

$$\sum_{i=0}^{\infty} p_i = \sum_{i=0}^{\infty} \frac{\lambda^i}{i!} e^{-\lambda} = e^{-\lambda} e^{\lambda} = 1$$

gelten.

Die Poissonverteilung wird auch Verteilung der seltenen Ereignisse genannt. Für großes n und kleines p lässt sich nämlich die Binomialverteilung gut durch die Poissonverteilung annähern ($\lambda = np$) (Faustregel: $p < .1$, $n > 50$).

Beispiel 4.19: Es sei bekannt, dass pro Jahr 0.005% einer Bevölkerungsgruppe durch einen gewissen Unfall getötet wird. Bei einer Versicherung sind 10 000 Personen aus obiger Gruppe gegen diesen Unfall versichert. Wie groß ist die Wahrscheinlichkeit, dass in einem gegebenen Jahr mindestens 3 dieser Versicherten durch den genannten Unfall umkommen?[3]

X = Anzahl der versicherten Leute, die pro Jahr durch diesen Unfall getötet werden.

$n = 10000$, $p = .00005$, $\lambda = np = .5$.

$$\begin{aligned} P(X \ge 3) &= 1 - P(X \le 2) \\ &= 1 - \sum_{i=0}^{2} \frac{\lambda^i}{i!} e^{-\lambda} = 1 - e^{-.5}(1 + \frac{1}{2} + \frac{1}{8}) = .0144 \approx 1.4\%. \end{aligned}$$

Wenn wir das Modell der Binomialverteilung verwenden, so finden wir

$$\begin{aligned} P(X \ge 3) &= 1 - \sum_{i=0}^{2} \binom{n}{i} p^i (1-p)^{n-i} \\ &= 1 - (\binom{10000}{0} .00005^0 .99995^{10000} + \binom{10000}{1} .00005^1 .99995^{9999} \\ &\quad + \binom{10000}{2} .00005^2 .99995^{9998}) \\ &= 1 - (.607 + .303 + .076) = .0144. \end{aligned}$$

Weitere Anwendungen: Anzahl der Telefonteilnehmer, die in einer Zentrale pro Zeiteinheit anrufen. Anzahl der Druckfehler pro Seite eines Buches. Anzahl der Rosinen in einem Stück Kuchen. Anzahl von tektonischen Störungen pro Flächeneinheit.

[3] ℝ: `1-ppois(2,.5)`

4.4.2 Stetige Zufallsvariable

Eine Wahrscheinlichkeitsverteilung heißt *absolut stetig* , wenn es eine nicht negative Funktion $f(x)$ gibt, sodass sich die Verteilungsfunktion $F(x)$ für alle $x \in \mathbb{R}$ als Integral über f darstellen lässt, nämlich

$$F(x) = \int_{-\infty}^{x} f(t)dt \quad .$$

f wird *(Wahrscheinlichkeits-) Dichte* genannt. X ist eine *stetige Zufallsvariable*, wenn ihre Verteilung absolut stetig ist. Wenn F differenzierbar ist, gilt

$$f(x) = F'(x) \quad .$$

Ein Beispiel einer absolut stetigen Verteilung wäre die Gleichverteilung (Beispiel auf Seite 36).

Beispiel 4.20: Dreiecksverteilung. Diese entsteht bei der Summe von 2 unabhängigen, gleichverteilten $R(0,1)$ Zufallsvariablen (siehe später). Die Dichte f ist gegeben durch

$$f(x) = \begin{cases} x & 0 \leq x \leq 1 \\ 2 - x & 1 \leq x \leq 2 \\ 0 & \text{sonst.} \end{cases}$$

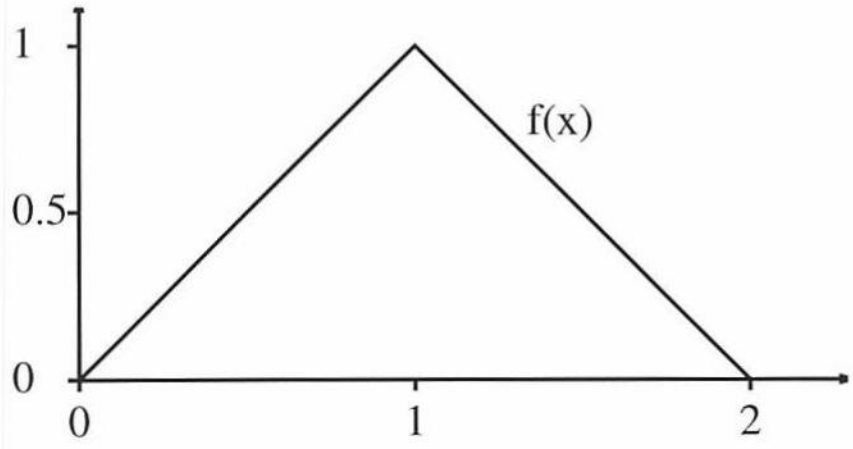


Abbildung 4.5: Dichte der Dreiecksverteilung.

Die Verteilungsfunktion errechnet sich durch Integration (Übung).

Beispiel 4.21: Normalverteilung $N(\mu, \sigma^2)$. Viele quantitative Größen konzentrieren sich oft um einen bestimmten Wert und größere Abweichungen sind eher selten. Diese Konzentration (besser *Dichte*) könnte man sich wie in der Abbildung 4.7 unter einer Glockenkurve vorstellen. Traditionsgemäß dient die *Normalverteilung* als Approximation eines solchen Verhaltens. Unter bestimmten Voraussetzungen kann

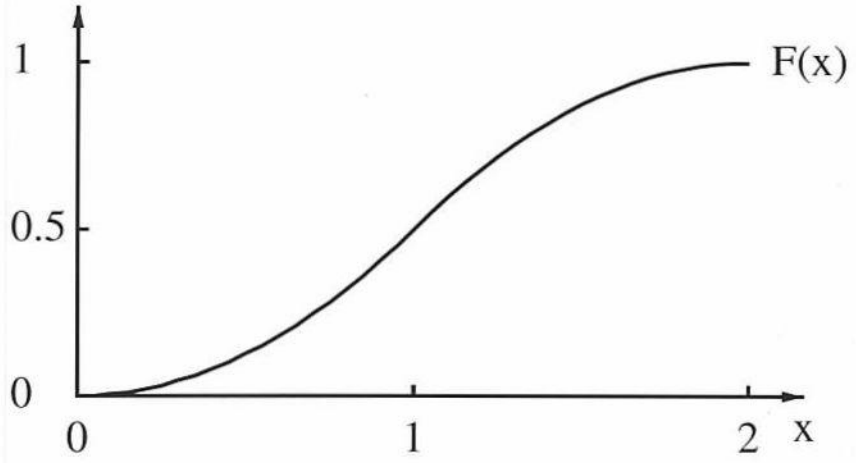


Abbildung 4.6: Verteilungsfunktion der Dreiecksverteilung.

man auch theoretisch zeigen, dass die Verteilung einer Summe von „unabhängigen" Zufallsvariablen gegen die Normalverteilung strebt (im besonderen gilt dies für die Binomialverteilung.) Die Dichte der Normalverteilung (bezeichnet mit $N(\mu, \sigma^2)$) ist gegeben durch

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}, \qquad \mu, x \in (-\infty, \infty),\ \sigma > 0 \ ,$$

wobei μ einen Ortsparameter (das Mittel) und σ einen Skalierungsparameter darstellen.[4]

Die Verteilungsfunktion ist natürlich das Integral über f, d.h.

$$F(x) = \int_{-\infty}^{x} f(t)dt = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(t-\mu)^2}{2\sigma^2}} dt \ .$$

Ihre Berechnung ist nicht ganz einfach, und entsprechende Werte werden üblicherweise Tabellen (siehe z.B. Tabelle A) entnommen.

Betrachten wir die transformierte Zufallsvariable $Z = (X-\mu)/\sigma$, wobei X $N(\mu, \sigma^2)$ verteilt ist. Die Verteilungsfunktion $G(z)$ wird zu

$$\begin{aligned} G(z) &= P(Z \le z) = P(\frac{X-\mu}{\sigma} \le z) = P(X \le z\sigma + \mu) \\ &= F(z\sigma + \mu) = \int_{-\infty}^{z\sigma+\mu} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(t-\mu)^2}{2\sigma^2}} dt. \end{aligned}$$

Die Transformation der Integrationsvariablen $t \to s\sigma + \mu$ liefert

$$G(z) = \int_{-\infty}^{z} \frac{1}{\sqrt{2\pi}} e^{-\frac{s^2}{2}} ds = \int_{-\infty}^{z} g(s)ds \ ,$$

[4] R: `plot(x<-seq(-3,3,length=100),dnorm(x),type='l')`

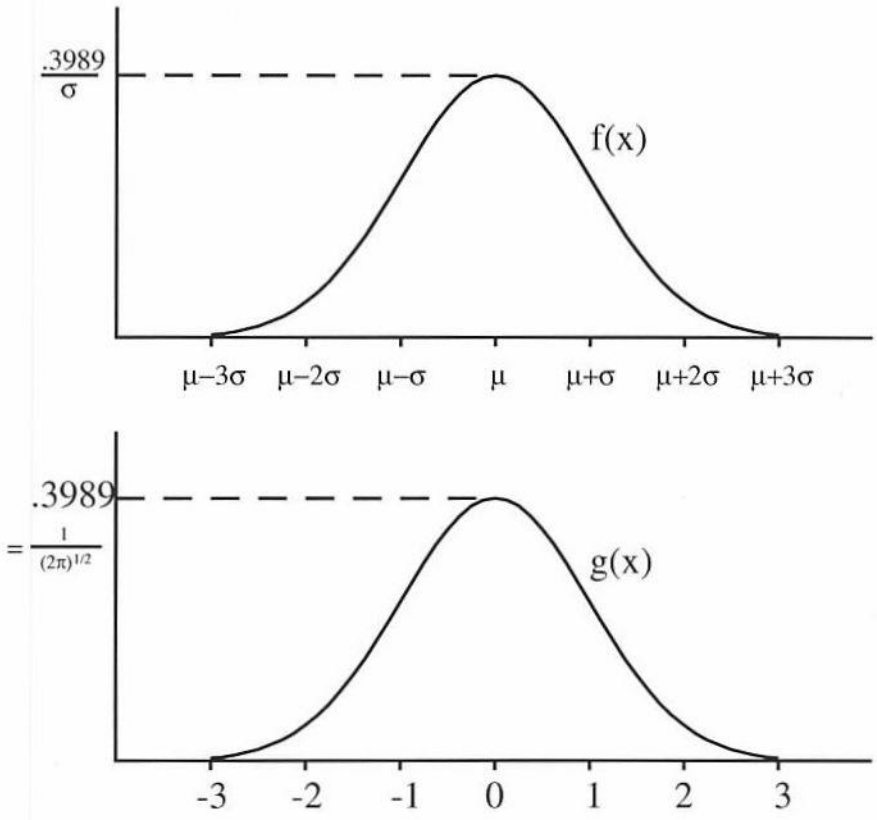


Abbildung 4.7: Dichte der Normalverteilung.

wobei

$$g(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} \quad .$$

G und g werden Verteilungsfunktion bzw. Dichtefunktion der *Standard-Normalverteilung* $N(0,1)$ bezeichnet. (Manchmal sind auch die griechischen Buchstaben Φ und ϕ (oder φ) üblich. Die Verteilung heißt auch Gauß'sch). Werte von G findet man in Tabellen (siehe Anhang, Tabelle A) und ein paar häufig verwendete sind

P(-1,1)	=	$\int_{-1}^{+1} g(t)dt$	=	2(G(1)-G(0))	=	68.26% [5]
P(-2,2)	=		=	2(G(2)-G(0))	=	95.45%
P(-3,3)	=				=	99.73%.

Beispiel 4.22: Ein Psychologe benützt ein Instrument, das ihm Werte liefert, die N(500, 10 000) verteilt sind. Gesucht ist die Wahrscheinlichkeit, dass ein Wert kleiner oder gleich 600 ist.

$$P(X \leq 600) = P(\frac{X - 500}{100} \leq \frac{600 - 500}{100}) = P(Z \leq 1) = G(1) = .8413 \quad .$$

[5] ☞: Z.B. `pnorm(1)-pnorm(-1)`

Wie groß ist die Schranke, unter der ein Wert mit 95% Wahrscheinlichkeit liegt?

$$.95 = P(X \leq x) = P(\frac{X - 500}{100} \leq \frac{x - 500}{100}) = G(\frac{x - 500}{100}) \quad .$$

Aus der Tabelle folgt

$$\frac{x - 500}{100} = 1.64 \quad ,$$

woraus sich die Schranke $x = 664$ ergibt.

Beispiel 4.23: Stichprobenwerte einer stetigen Verteilung. Gemessene Werte werden praktisch immer nur bis auf eine endliche Anzahl von Dezimalstellen registriert, sodass man eigentlich von einer diskreten Zufallsvariablen sprechen müsste. Es ist jedoch technisch sinnvoll, die Werte weiter als „stetig" zu betrachten.

Grumell und Dunningham prüften 250 Stichprobenwerte von Kohle eines bestimmten Ursprungs auf Asche. Die Werte sind in Form einer Häufigkeitstabelle festgehalten:

Intervall oder Zelle	Häufigkeit f	Anteil p	Geschätzte Wahrschein. (normal) $(\hat{p})$	Geschätzte Häufigkeit (normal) $e = 250(\hat{p})$	Standard. Residuen $\frac{f-e}{\sqrt{e}}$
(9.00-12.99)	(16)	0.064	0.064	16.01	-0.00
9.00-9.99	1	0.004			
10.00-10.99	3	0.012			
11.00-11.99	3	0.012			
12.00-12.99	9	0.036			
13.00-13.99	13	0.052	0.062	15.54	-0.67
14.00-14.99	27	0.108	0.096	24.02	0.61
15.00-15.99	28	0.112	0.128	31.88	-0.69
16.00-15.99	39	0.156	0.147	36.85	0.35
17.00-17.99	42	0.168	0.148	37.02	0.82
18.00-18.99	34	0.136	0.128	32.08	0.34
19.00-19.99	19	0.076	0.097	24.28	-1.07
20.00-20.99	14	0.056	0.064	15.88	-0.47
21.00-21.99	10	0.040			
22.00-22.99	4	0.016			
23.00-23.99	3	0.012			
24.00-24.99	0	0			
25.00-25.99	1	0.004			
(21.00-25.99)	(18)	0.072	0.065	16.34	0.41
	250	1.00		250.00	

Einige Zellen mit geringen Häufigkeiten sind zusammengefasst (in Klammern am Beginn und Ende der Tabelle). Die Anteile p sind auch in der folgenden Grafik (Histogramm) aufgetragen. Um zu sehen, wie gut die Dichte einer Normalverteilung $N(\mu, \sigma^2)$ dieses Histogramm approximiert, führen wir folgende informelle Rechnung

durch (Genaueres siehe später). Der Parameter μ stellt das Mittel dar und wird durch

$$\overline{x} = \frac{1}{n}\sum x_i = 17.015 \ ,$$

wobei x_i die Messwerte bezeichnen, geschätzt und σ^2 durch

$$s^2 = \frac{1}{n-1}\sum (x_i - \overline{x})^2 = 6.97$$

approximiert. Die damit definierte Wahrscheinlichkeitsdichte ist ebenfalls in der Grafik (Abbildung 4.8) über dem Histogramm dargestellt. Die durch dieses Modell geschätzten Wahrscheinlichkeiten p sowie die Häufigkeiten e für jedes Intervall sind in der Häufigkeitstabelle eingetragen.

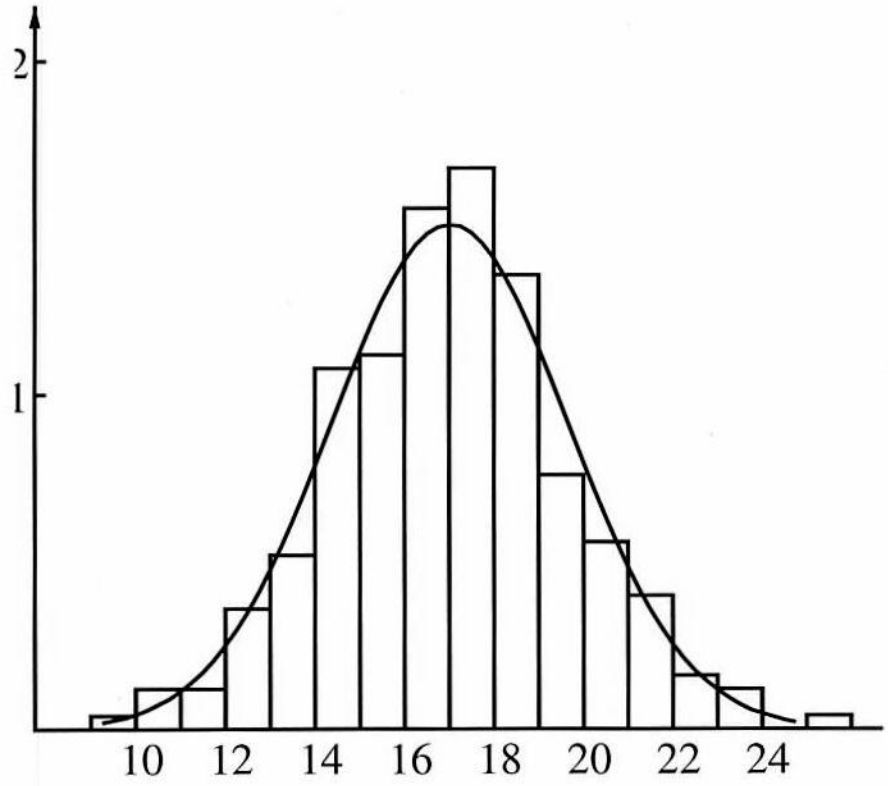


Abbildung 4.8: Histogramm und geschätzte Dichte von Aschenanteil.

Weiters wollen wir prüfen, ob das Modell der Normalverteilung richtig sein kann. Dazu sind in der letzten Spalte *standardisierte Residuen*

$$d = \frac{f - e}{\sqrt{e}}$$

angegeben, die unter der Annahme des Modells ungefähr standard-normalverteilt sind. Die Werte in der Tabelle widersprechen offensichtlich nicht dieser Annahme (alle absolut kleiner als 2). Ein weiterer einfacher „Test" besteht darin, die Größe

$$\chi^2 = \sum_{j=1}^{10} d_j^2 = 3.74 \quad ,$$

die ungefähr chi-quadrat mit 7 Freiheitsgraden verteilt ist, zu untersuchen. Der Vergleich von $\chi^2 = 3.74$ mit Werten aus Tabelle A des Anhangs gibt keine Indikation gegen die Annahme der Normalverteilung.

4.4.3 Wahrscheinlichkeitsnetz

Manchmal ist es für die Überprüfung einer Verteilung günstiger, die Verteilungsfunktion statt der Dichte zu betrachten. Wenn $x_1, \ldots, x_n$ n Datenpunkte bezeichnen, so heißt die Funktion

$$F_n(x) = \frac{1}{n} \sum_{i=1}^{n} I_{(-\infty,x]}(x_i)$$

empirische Verteilungsfunktion, wobei I die Indikatorfunktion ist. Diese Treppenfunktion F_n gibt im wesentlichen die relative Summenhäufigkeit an, weist n Sprünge der Größe $1/n$ auf und hat alle Eigenschaften einer Verteilungsfunktion.

Für den Zweck der Prüfung auf eine Normalverteilung ist es von Vorteil, die Skalierung der Ordinate zu ändern (zu verzerren). Im *Wahrscheinlichkeitspapier* oder *Wahrscheinlichkeitsnetz* wird die Ordinate zwischen 0 und 1 nicht in gleich große Intervalle geteilt, sondern es werden Abstände proportional G^{-1} aufgetragen, wobei G die Verteilungsfunktion der Standard-Normalverteilung ist. Bei Verwendung dieser Skalierung wird eigentlich nicht $F_n(x)$ über x grafisch dargestellt, sondern $G^{-1}(F_n(x))$ über x. Wenn nun die Daten ungefähr normalverteilt sind, so wird $F_n(x) \sim F(x)$ (also $N(\mu, \sigma^2)$) sein und

$$G^{-1}(F_n(x)) \sim G^{-1}(F(x)) = G^{-1}(G(\frac{x-\mu}{\sigma})) = \frac{x-\mu}{\sigma} \quad ,$$

sodass die Treppenfunktion ungefähr auf einer Geraden zu liegen kommt. Die Abbildung 4.9 soll das illustrieren. Man beachte, dass man auch sofort Schätzwerte für den Mittelwert μ und den Skalierungsparameter σ ablesen kann.

In der Praxis wird man nicht die Treppenfunktion, sondern nur die Punkte der Sprungstellen einzeichnen, die wiederum ungefähr auf einer Geraden liegen müssten. Die Abweichungen sind rein zufälliger Natur und werden stark von der Anzahl der Daten beeinflusst. Um ein gewisses Gefühl für „zufällige" Abweichungen zu vermitteln, haben wir in Abbildung 4.10 verschieden große Stichproben von

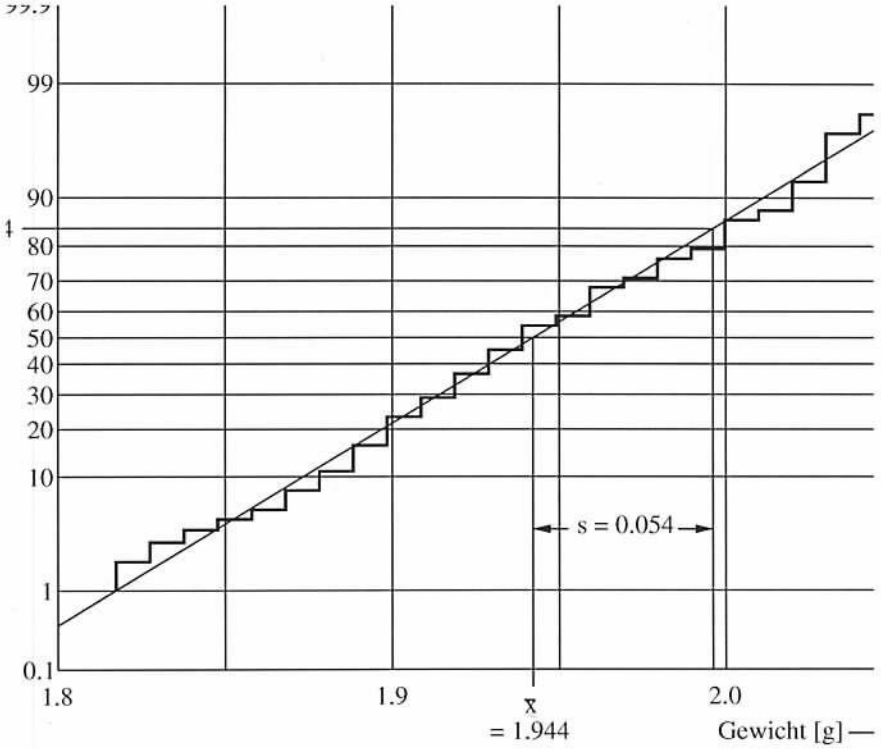


Abbildung 4.9: Wahrscheinlichkeitspapier.

normalverteilten Werten dargestellt. Ordinaten und Abszissen sind manchmal vertauscht, was aber natürlich keine Rolle spielt. Am einfachsten ist aber die Übung mit einem einfachen Befehl in ℝ.[6]

4.4.4 Funktionen einer Zufallsvariablen

Eine reelle Funktion h einer Zufallsvariablen X ist ebenfalls eine Zufallsvariable, etwa $Y = h(X)$, wenn gilt, dass die Urbilder der Borel-Mengen wieder Borel-Mengen sind, nämlich für alle reelle Zahlen y

$$h^{-1}(-\infty, y] \in \mathfrak{Z} \quad .$$

Das Wahrscheinlichkeitsmaß für Y ist dann definiert durch

$$P(Y \leq y) = P(X \in h^{-1}(-\infty, y]) \quad .$$

Beispiel 4.24: Eine diskrete Zufallsvariable X sei gegeben mit den Wahrscheinlichkeiten

[6] ℝ: `probplot(rnorm(100))` (Package: `e1071`)
ℝ: `qqnorm(x <- rnorm(100)); qqline(x)`

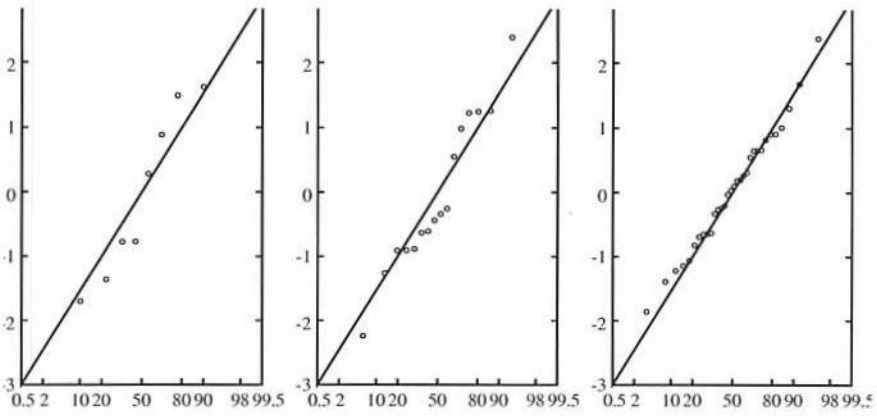


Abbildung 4.10: Simulierte normalverteilte Daten auf Wahrscheinlichkeitspapier.

i	−2	−1	0	1	2
P(X=i)	1/5	1/6	1/5	1/15	11/30

Wir betrachten die neue Zufallsvariable $Y = h(X) = X^2$ und erhalten

i	0	1	4
P(Y=i)	1/5	7/30	17/30

Beispiel 4.25: X sei $N(0,1)$ verteilt, d.h.

$$g(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}, \quad -\infty < x < \infty \quad .$$

Wir betrachten die Transformation $Y = h(X) = X^2$. Die Verteilungsfunktion von Y für $(y > 0)$ errechnet sich als

$$F_Y(y) = P(Y \leq y) = P(0 \leq Y \leq y) = P(X \in h^{-1}(0, y])$$

$$= P(-\sqrt{y} \leq X \leq \sqrt{y}) = G(\sqrt{y}) - G(-\sqrt{y}) \quad ,$$

woraus die Dichtefunktion sofort folgt:

$$f_Y(y) = \frac{dF_Y(y)}{dy} = \frac{1}{2\sqrt{y}} g(\sqrt{y}) + \frac{1}{2\sqrt{y}} g(-\sqrt{y}) \quad .$$

In unserem speziellen Fall der Normalverteilung bekommen wir

$$f_Y(y) = \frac{1}{\sqrt{2\pi y}} e^{-\frac{y}{2}} \qquad \text{für} \quad y > 0$$

und sonst 0. Diese Verteilung von Y wird auch *Chi-Quadrat-Verteilung* mit einem Freiheitsgrad genannt (χ_1^2).

Im Fall einer stetigen Zufallsvariablen X und einer Funktion h, die differenzierbar und streng monoton wachsend ist ($h'(x) > 0 \; \forall x$), können wir folgende Regel herleiten:

$$F_Y(y) = P(Y \leq y) = P(X \leq h^{-1}(y)) = F(h^{-1}(y)) \quad ,$$

woraus durch Ableiten folgt

$$\boxed{f_Y(y) = \frac{dF(h^{-1}(y))}{dy} \quad = f(h^{-1}(y))\frac{dh^{-1}(y)}{dy}} \quad .$$

Für streng monoton fallende h gilt Ähnliches.

Beispiel 4.26: Die Zufallsvariable X sei gleichverteilt im Intervall [0,1] : $X \sim$ R(0,1). Wir betrachten die Variable

$$Y = h(X) = -\ln X \quad ,$$

wobei offensichtlich die Funktion $h(x)$ für $x > 0$ streng monoton fallend ist. Die gewohnte Rechnung liefert für $y > 0$

$$F_Y(y) = P(Y \leq y) = P(-ln(X) \leq y) = P(X \geq e^{-y}) = 1 - F(e^{-y}) = 1 - e^{-y} \quad ,$$

wodurch

$$f_Y(y) = \frac{dF_Y(y)}{dy} = e^{-y} \quad .$$

Die Verteilung wird *Exponentialverteilung* genannt.

Beispiel 4.27: Die Zufallsvariable X sei normalverteilt: $X \sim N(\mu, \sigma^2)$. Betrachten wir $Y = e^X$. Die Funktion $h(x) = e^x$ ist differenzierbar und streng monoton steigend. Daher ist obige Regel anwendbar und wir erhalten

$$f_Y(y) = \frac{1}{\sqrt{2\pi}\sigma y} e^{-\frac{(\ln y - \mu)^2}{2\sigma^2}} \quad .$$

Die Verteilung von Y wird *Log-Normalverteilung* genannt, weil der Logarithmus von Y normalverteilt ist.

4.4.5 Erwartung

Bezeichne h eine reellwertige Funktion der Zufallsvariablen X. Dann ist der *Mittelwert* oder die *mathematische Erwartung* von $h(X)$ im Falle einer stetigen Zufallsvariablen durch

$$E[h(X)] = \int_{-\infty}^{\infty} h(x)f(x)dx \qquad (\text{vorausgesetzt} \int_{-\infty}^{\infty} \mid h(x) \mid f(x)dx < \infty)$$

und im Fall einer diskreten Zufallsvariablen durch

$$E[h(X)] = \sum_{i=1}^{\infty} h(x_i)p_i \qquad (\text{vorausgesetzt} \sum_{i=1}^{\infty} \mid h(x_i) \mid p_i < \infty)$$

definiert. Häufig sucht man die Erwartung einer Funktion der Form $ah_1(X) + bh_2(X)$ mit Konstanten a und b. Dann gilt (wie man sich leicht überzeugt)

$$E[ah_1(X) + bh_2(X)] = aE[h_1(X)] + bE[h_2(X)] \quad .$$

Ist h die Identität $h(x) = x$, dann erhalten wir die *Erwartung* oder den *Mittelwert* der Zufallsvariablen

$$\mu = E(X) = \int xf(x)dx$$

bzw.

$$\mu = E(X) = \sum x_i p_i \quad .$$

Die Größe

$$\sigma^2 = Var(X) = E(X - \mu)^2 = E[X - E(X)]^2$$

mit $h(x) = (x-\mu)^2$ bezeichnet man als *Varianz* oder *zentrales Moment 2. Ordnung.* σ heißt auch *Streuung* oder *Standardabweichung.* Es gilt

$$Var(X) = E(X)^2 - (EX)^2 \quad .$$

Beispiel 4.28:
(a) Poissonverteilung $P(\lambda)$:

$$\mu = EX = \sum_{i=0}^{\infty} i\frac{\lambda^i}{i!}e^{-\lambda} = \sum_{i=1}^{\infty} \frac{\lambda^{i-1}}{(i-1)!}\lambda e^{-\lambda} = \lambda \quad .$$

$$EX^2 = \sum_{i=0}^{\infty} i^2\frac{\lambda^i}{i!}e^{-\lambda} = \lambda + \sum_{i=0}^{\infty} i(i-1)\frac{\lambda^i}{i!}e^{-\lambda} = \lambda + \sum_{i=2}^{\infty} \lambda^2\frac{\lambda^{i-2}}{(i-2)!}e^{-\lambda} = \lambda + \lambda^2 \quad .$$

$$\sigma^2 = Var(X) = EX^2 - (EX)^2 = \lambda + \lambda^2 - \lambda^2 = \lambda \quad .$$

(b) Bernoulli-Verteilung $Bi(1,p)$:

$$\mu = EX = 1 \times p + 0 \times (1-p) = p \quad .$$

$$EX^2 = 1^2 p + 0^2(1-p) = p \quad .$$

$$\sigma^2 = p - p^2 = p(1-p) \quad .$$

(c) Binomialverteilung $Bi(n,p)$:

$$\mu = np$$

$$\sigma^2 = np(1-p) \qquad \text{(Herleitung später)} \quad .$$

(d) Rechtecksverteilung $R(a,b)$:

$$\mu = E(X) = \int_a^b x\frac{1}{b-a}dx = \frac{b^2-a^2}{2}\frac{1}{b-a} = \frac{a+b}{2} \quad .$$

$$EX^2 = \int_a^b x^2 \frac{1}{b-a} dx = \frac{b^3 - a^3}{3} \frac{1}{b-a} = \frac{b^2 + ab + a^2}{3} \quad .$$

$$\sigma^2 = VarX = \frac{b^2 + ab + a^2}{3} - (\frac{a+b}{2})^2 = \frac{(b-a)^2}{12} \quad .$$

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

$$EX = \int_{-\infty}^{\infty} \frac{x}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}} dx \quad .$$

Substitution von $z = \frac{x-u}{\sigma}$ liefert

$$EX = \int_{-\infty}^{\infty} \frac{\sigma z + \mu}{\sqrt{2\pi}} e^{-\frac{z^2}{2}} dz = 0 + \mu \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}} dz = \mu \quad .$$

$$Var(X) = \int_{-\infty}^{\infty} \frac{(x-\mu)^2}{\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} dx \quad .$$

Die gleiche Substitution und partielle Integration liefert

$$Var(X) = \int_{-\infty}^{\infty} \frac{\sigma^2 z^2}{\sqrt{2\pi}} e^{-\frac{z^2}{2}} dz$$

$$= \sigma^2 [-\frac{z}{\sqrt{2\pi}} e^{-\frac{z^2}{2}} |_{-\infty}^{\infty} + \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}} dz] = \sigma^2 \quad .$$

Übung 4.1: Man wirft einen Würfel, bis eine 6 erscheint. Wie groß ist die erwartete Anzahl von Würfen?

Zur weiteren Charakterisierung einer Verteilung definieren wir die *Schiefe* als

$$\gamma = \frac{1}{\sigma^3} E(X - \mu)^3$$

und die *Kurtosis* (4. Moment) als

$$\chi = \frac{1}{\sigma^4} E(X - \mu)^4 - 3 \quad .$$

4.4.6 Näherung der Binomialverteilung durch die Normalverteilung

Für große Werte von n lässt sich die Binomialverteilung durch die Normalverteilung mit dem Mittelwert

$$\mu = np$$

und der Varianz

$$\sigma^2 = np(1-p)$$

annähern. Dabei muss beachtet werden, dass die Binomialverteilung für Werte von p nahe 0 oder 1 sehr schief ist. Für p nahe .5 ist die Näherung allerdings auch für

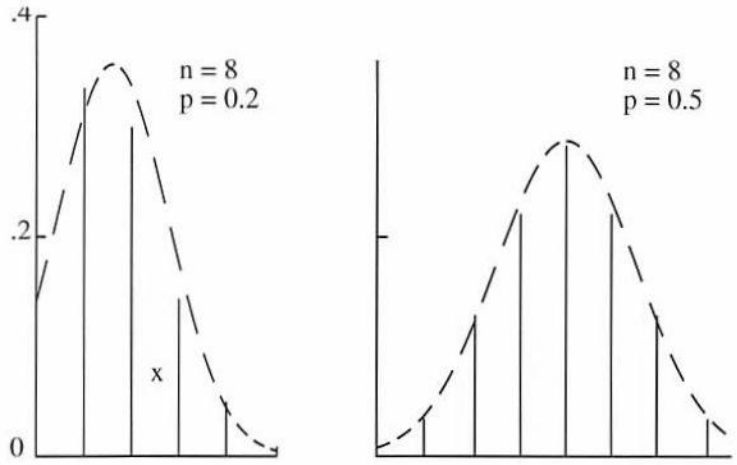


Abbildung 4.11: Binomialverteilung $Bi(8, p)$ und Normalverteilungsapproximation.

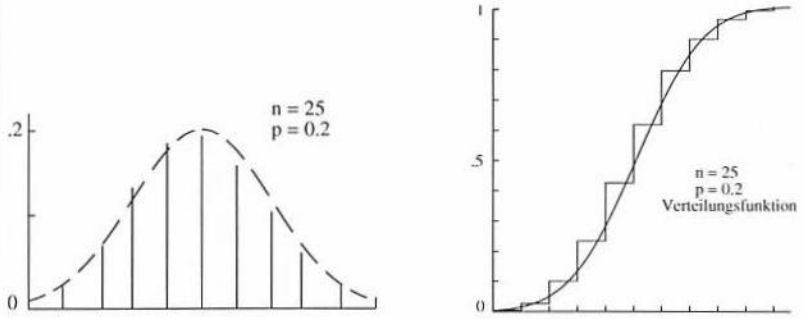


Abbildung 4.12: Binomialverteilung $Bi(25, 0.2)$ und Normalverteilungsapproximation.

kleinere Werte von n recht gut, was in den Abbildungen 4.11 und 4.12 illustriert wird.

Beispiel 4.29: Daten aus dem Beispiel auf Seite 44. Obwohl $n = 10\ 000$ sehr groß ist, stellt $p = .00005$ doch einen extremen Fall dar, wobei die Approximation noch sehr schlecht ist: Wir bekommen $\mu = np = .5$, $\sigma^2 = np(1-p) = .499975$ und

$$P(X \geq 3) = P(\frac{X - np}{\sqrt{np(1-p)}} \geq \frac{3 - .5}{\sqrt{.499975}}) \approx P(Z \geq 3.54) = .0002 \ ,$$

also .02% statt 1.4% von früher.

4.5 Mehrdimensionale Zufallsvariable

Die gleichzeitige Betrachtung mehrerer Zufallsvariablen ist aus verschiedenen Gründen wichtig. Erstens wird bei Zufallsexperimenten meistens mehr als eine Größe gemessen bzw. untersucht. Z.B. wird eine Bohrprobe nicht nur auf den Gehalt *eines* chemischen Elementes geprüft, sondern es werden viele Eigenschaften untersucht. Bei der Klassifizierung von Tieren oder Personen werden auch mehrere Merkmale wie Körpergröße, Gewicht, Augenfarbe, Schädelform, etc. betrachtet. Erforscht man jede Variable einzeln, so gehen die Zusammenhänge zwischen ihnen verloren.

Eine Zufallsvariable kann vom Ort oder von der Zeit abhängig sein, sodass sie zu verschiedenen Zeitpunkten zwar Eigenschaften des gleichen Merkmals angibt, aber verschiedene Verteilungen aufweist. Es kann also zu jedem Zeitpunkt eine andere Zufallsvariable angenommen werden.

Verteilungen mehrerer Variablen spielen auch bei der theoretischen Begründung statistischer Prüfverfahren eine Rolle. Eine Stichprobe des Umfangs n kann oft als „Realisation", als spezieller Wert einer n-dimensionalen Zufallsvariablen interpretiert werden. Man spricht dann von Zufallsvektoren.

Bezeichnen wir mit $X_1, \ldots, X_p$ p Zufallsvariable, d.h. $X_i : \Omega \to \mathbb{R}$ mit $X_i^{-1}(-\infty, x] \in \mathcal{A}$ für alle $x \in \mathbb{R}$ und $i = 1, \ldots, p$. Dann definieren wir einen p-dimensionalen *Zufallsvektor*

$$\boldsymbol{X} = (X_1, X_2, \ldots, X_p)^\top$$

als Abbildung von $(\Omega, \mathcal{A})$ in den p-dimensionalen Raum $(\mathbb{R}^p, \mathcal{B}^p)$, wobei das Urbild jedes Intervalls

$$I = \{(t_1, \ldots, t_p) \mid -\infty < t_i \leq x_i, \quad x_i \in \mathbb{R}, \quad i = 1, \ldots, p\}$$

in $\mathcal{A}$ liegen soll, d.h.

$$\boldsymbol{X}^{-1}(I) \in \mathcal{A} \ .$$

Die Wahrscheinlichkeit, dass $\boldsymbol{X}$ in das Intervall I fällt, ist offensichtlich definiert durch

$$P(\boldsymbol{X} \in I) = P(\omega \in \boldsymbol{X}^{-1}(I)) \ .$$

Beispiel 4.30: Ein Zufallsvektor $\boldsymbol{X}$ sei durch die Kodierung von p verschiedenen Merkmalen (z.B. Größe, Geschlecht, Anzahl der Zähne) einer Person definiert. Er stellt eine Abbildung der Elementarereignisse ω (z.B. eine bestimmte Person aus einer Bevölkerungsgruppe) in $\mathbb{R}^p$ dar. Die Wahrscheinlichkeit, dass $\boldsymbol{X}$ in ein spezifisches Intervall (aus $\mathbb{R}^p$) fällt, ist durch die Wahrscheinlichkeit des Urbildes dieses Intervalls gegeben, d.h. durch die Wahrscheinlichkeit, eine Person auszuwählen, deren kodierte Merkmale gerade in dieses Intervall fallen.

Der Einfachheit halber beschränken wir uns nun auf $p = 2$ Zufallsvariable.

Wir definieren wieder die *Verteilungsfunktion* F (jetzt auf $\mathbb{R}^2$) durch

$$F(x_1, x_2) = F(\boldsymbol{x}) = P(X_1 \leq x_1, X_2 \leq x_2)$$

$$= P((-\infty, x_1] \times (-\infty, x_2])$$
$$= P(\boldsymbol{X} \leq \boldsymbol{x}) \ .$$

F weist wie im eindimensionalen Fall folgende Eigenschaften auf: Sie ist in jeder Koordinate monoton wachsend und rechtsseitig stetig. Sie strebt gegen 1, wenn *alle* Argumente x_i gegen ∞ streben. Sie strebt gegen 0, wenn *ein* Argument x_i gegen $-\infty$ strebt.

Beispiel 4.31: 2 symmetrische Münzen werden auf einmal geworfen. $\boldsymbol{X}$ ist als Indikator der Köpfe definiert; d.h. X_1 ist 1, wenn die 1. Münze Kopf zeigt, und sonst 0; für X_2 gilt ähnliches bezüglich der 2. Münze. Die Verteilungsfunktion F ergibt sich dann folgendermaßen.

$$F(\boldsymbol{x}) = F(x_1, x_2) = \begin{cases} 0 & \text{wenn } -\infty < x_1 < 0 \text{ oder } -\infty < x_2 < 0 \\ 1/4 & 0 \leq x_1 < 1 \text{ und } 0 \leq x_2 < 1 \\ 1/2 & 0 \leq x_1 < 1 \quad 1 \leq x_2 < \infty \\ 1/2 & 0 \leq x_2 < 1 \quad 1 \leq x_1 < \infty \\ 1 & 1 \leq x_1 < \infty \quad 1 \leq x_2 < \infty. \end{cases}$$

Die Verteilung von $\boldsymbol{X}$ nennt man *diskret*, wenn $\boldsymbol{X}$ höchstens abzählbar viele verschiedene Werte annehmen kann, d.h.

$$\boldsymbol{X}(\omega) = (X_1(\omega), X_2(\omega)) = (x_1^{(i_1)}, x_2^{(i_2)}) \in \mathbb{R}^2 \ ,$$

$$i_j = 1, 2, \ldots, \quad j = 1, 2 \ .$$

Die Verteilungsfunktion F kann dann als

$$F(\boldsymbol{x}) = P(\boldsymbol{X} \leq \boldsymbol{x}) = \sum_{x_1^{(i_1)} \leq x_1, x_2^{(i_2)} \leq x_2} P(\boldsymbol{X} = (x_1^{(i_1)}, x_2^{(i_2)}))$$

beschrieben werden, wobei die Summation über alle Punkte $(x_1^{(i_1)}, x_2^{(i_2)})$ durchgeführt wird, deren Komponenten kleiner oder gleich den Komponenten von $\boldsymbol{x}$ sind (siehe auch Abbildung 4.13).

$$p_{i_1 i_2} = P(X_1 = x_1^{(i_1)}, X_2 = x_2^{(i_2)})$$

wird wieder als *Wahrscheinlichkeitsfunktion* bezeichnet.

Beispiel 4.32: 2 symmetrische Münzen werden 2 mal geworfen. $\boldsymbol{X} = (X_1, X_2)$ gibt die Orientierung der Münzenpaare auf dem Tisch an, d.h. X_i ist der Winkel, den die Verbindungsgerade der beiden Münzen nach dem i-ten Wurf $(i = 1, 2)$ mit einer bestimmten Tischkante einschließt. Der Wertebereich von X_i ist also 0 bis 180^o. Aus Symmetriegründen finden wir Wahrscheinlichkeiten von Intervallen z.B. als

$$P((0, 10] \times (0, 10]) = P((0, 10] \times (10, 20]) = \ldots$$

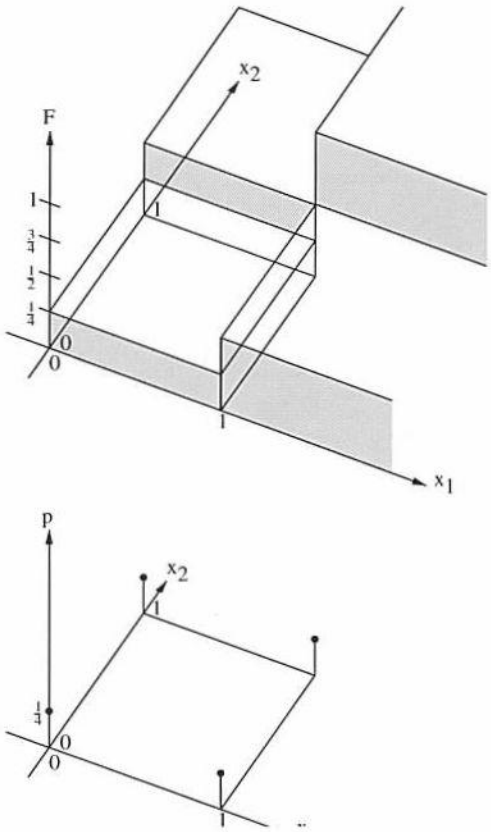


Abbildung 4.13: Verteilungsfunktion für das Werfen von zwei Münzen.

$$= P((170, 180] \times (170, 180]) = \frac{10^2}{180^2} \quad .$$

Aus immer feiner werdenden Unterteilungen schließen wir, dass

$$F(\boldsymbol{x}) = F(x_1, x_2) = \begin{cases} 0 & \text{wenn} \quad -\infty < x_1 < 0 \quad \text{oder} \quad -\infty < x_2 < 0, \\ \frac{x_1 x_2}{180^2} & \quad 0 \le x_1 < 180 \quad \text{und} \quad 0 \le x_2 < 180 \\ \frac{x_1}{180} & \quad 0 \le x_1 < 180 \quad\quad\; 180 \le x_2 < \infty \\ \frac{x_2}{180} & \quad 180 \le x_1 < \infty \quad\quad\; 0 \le x_2 < 180 \\ 1 & \quad 180 \le x_1 < \infty \quad\quad\; 180 \le x_2 < \infty. \end{cases}$$

Die Wahrscheinlichkeit ist also gleichmäßig verteilt über dem Quadrat $(0, 180] \times (0, 180]$, und zwar mit der Dichte $1/180^2$ innerhalb des Quadrates.

Die Verteilung von $\boldsymbol{X}$ heißt *absolut stetig*, wenn die Verteilungsfunktion F als p-faches Integral (sagen wir, $p = 2$)

$$F(\boldsymbol{x}) = F(x_1, x_2) = \int_{-\infty}^{x_2} \int_{-\infty}^{x_1} f(t_1, t_2) dt_1 dt_2$$

über eine nichtnegative Funktion f dargestellt werden kann. f heißt wieder Wahrscheinlichkeitsdichte, und die Wahrscheinlichkeit eines Ereignisses $B \in \mathfrak{B}^2$ kann

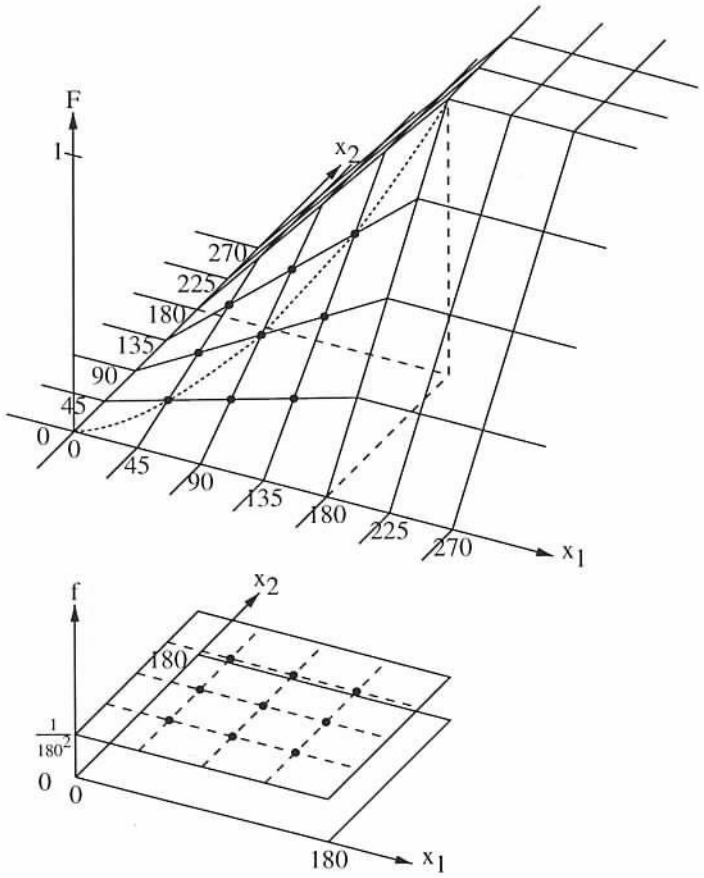


Abbildung 4.14: 2-dimensionale stetige Verteilungsfunktion.

als

$$P(B) = P(\boldsymbol{X} \in B) = \underbrace{\int\int}_{B} f(t_1, t_2) dt_1 dt_2$$

berechnet werden. Für das Beispiel auf Seite 58 folgt z.B.

$$P((0, 10] \times (0, 10]) = \int_0^{10} \int_0^{10} f(t_1, t_2) dt_1 dt_2 =$$

$$= \int_0^{10} \int_0^{10} \frac{1}{180^2} dt_1 dt_2 = \frac{10^2}{180^2} \quad .$$

4.5.1 Randverteilung und Unabhängigkeit

Es sei der Zufallsvektor $\boldsymbol{X} = (X_1, \ldots, X_p)$ mit der entsprechenden Verteilungsfunktion $F(\boldsymbol{x})$ gegeben. Oft ist die Verteilung einer Komponente X_j $(j \in \{1, \ldots, p\})$ ohne Berücksichtigung der anderen von Interesse. Man spricht von der *Randvertei-*

lung der Zufallsvariablen X_j, die durch Summation über alle möglichen Zustände der anderen gefunden wird.

Wenn die Dimension des Zufallsvektors $\boldsymbol{X}$ gleich 2 ist, z.B. $\boldsymbol{X} = (X, Y)^\top$, vereinfachen sich die komplizierten Formeln wesentlich. Im diskreten Fall ergibt sich mit

$$p_{ij} = P(X = x_i, Y = y_j)$$

die Randverteilung von X als

$$p_{i.} = p_{X,i} = P(X = x_i) = \sum_{j=1}^{\infty} p_{ij}$$

und von Y als

$$p_{.j} = p_{Y,j} = P(Y = y_j) = \sum_{i=1}^{\infty} p_{ij} \quad .$$

Im stetigen Fall erhalten wir

$$f_X(x) = \int_{-\infty}^{\infty} f(x, y)dy$$

bzw.

$$f_Y(y) = \int_{-\infty}^{\infty} f(x, y)dx \quad .$$

Für die Verteilungsfunktionen der Randverteilungen gilt

$$F_X(x) = F(x, \infty)$$

bzw.

$$F_Y(y) = F(\infty, y) \quad .$$

Beispiel 4.33: 3 symmetrische Münzen werden geworfen. X gibt die Anzahl von „Kopf" an, Y den Absolutbetrag der Differenz zwischen der Anzahl von „Kopf" und der Anzahl von „Wappen". Die Wahrscheinlichkeitsverteilung des Zufallsvektors (X, Y) wird mit $p_{ij} = P(X = i, Y = j)$, $i = 0, 1, 2, 3$, $j = 1, 3$ festgelegt. In der folgenden Tabelle werden p_{ij} sowie die Randverteilungen angegeben.

Y \ X	0	1	2	3	$P(Y = j) = p_{.j}$
1	0	$\frac{3}{8}$	$\frac{3}{8}$	0	$\frac{6}{8}$
3	$\frac{1}{8}$	0	0	$\frac{1}{8}$	$\frac{2}{8}$
$p_{i.} = P(X = i)$	$\frac{1}{8}$	$\frac{3}{8}$	$\frac{3}{8}$	$\frac{1}{8}$	1

Beispiel 4.34: Die gemeinsame Dichte der Zufallsvariablen (X, Y) sei

$$f(x, y) = \begin{cases} 2 & \text{für} \quad 0 < x < y < 1 \\ 0 & \text{sonst.} \end{cases}$$

Die Randverteilungen ergeben sich als

$$f_X(x) = \int_x^1 f(x,y)dy = \begin{cases} 2-2x & \text{für} \quad 0 < x < 1 \\ 0 & \text{sonst,} \end{cases}$$

$$f_Y(y) = \int_0^y f(x,y)dx = \begin{cases} 2y & \text{für } 0 < x < 1 \\ 0 & \text{sonst.} \end{cases}$$

Man sagt, dass zwei Zufallsvariable X und Y voneinander *unabhängig* sind, wenn die Verteilungsfunktion des Zufallsvektors (X,Y) gleich dem Produkt der Randverteilungsfunktionen ist, d.h.

$$F(x,y) = F_X(x)F_Y(y) \text{ für alle } x,y \in \mathbb{R} \ .$$

Man kann sich überlegen, dass die Unabhängigkeit zweier Zufallsvariablen gleichbedeutend der Unabhängigkeit der Ereignisse bezüglich der einen von allen Ereignissen bezüglich der anderen Zufallsvariablen ist.

Es lässt sich auch zeigen, dass für Wahrscheinlichkeitsfunktionen bzw. Dichtefunktionen im Falle der Unabhängigkeit äquivalent gilt

$$p_{ij} = p_{i\cdot}p_{\cdot j} \quad \text{für alle } i,j,$$

bzw.

$$f(x,y) = f_X(x)f_Y(y) \quad \text{für alle } x,y \in \mathbb{R} \ .$$

Beispiel 4.35:

(i) X und Y im Beispiel auf Seite 61 sind nicht unabhängig, weil

$$P(X=0,Y=1) = 0 \neq P(X=0)P(Y=1) = \frac{1}{8} \times \frac{6}{8} \ .$$

(ii) X und Y im nächsten Beispiel auf Seite 61 sind ebenfalls nicht unabhängig. (Übung!)

(iii) X_1 und X_2 in den Beispielen 4.31 und 4.32 auf Seite 58 sind unabhängig. (Übung!)

4.5.2 Funktionen eines Zufallsvektors

Sei $\boldsymbol{X}$ ein p-dimensionaler Zufallsvektor. Dann definiert eine Funktion, die R^p in R^q abbildet, einen q-dimensionalen Zufallsvektor $\boldsymbol{Y} = h(\boldsymbol{X})$, wenn für jedes $B \in \mathfrak{B}^q$ gilt

$$h^{-1}(B) \in \mathfrak{B}^p \ .$$

Die Wahrscheinlichkeit, dass $\boldsymbol{Y} \in B$, folgt aus

$$P(\boldsymbol{Y} \in B) = P(\boldsymbol{X} \in h^{-1}(B)) \ .$$

Beispiel 4.36: X und Y geben die Augenzahlen von jeweils einem von 2 geworfenen Würfeln an. Es interessiert die Verteilung von $Z = X + Y$, der Summe der Augenzahlen. Es gilt offensichtlich

$$P(X = i) = p_{i.} = 1/6, \quad i = 1, \ldots, 6 \ ,$$

$$P(Y = j) = p_{.j} = 1/6, \quad j = 1, \ldots, 6 \ ,$$

und wegen der Unabhängigkeit

$$P(X = i, Y = j) = p_{ij} = p_{i.}p_{.j} \ .$$

Für $P(Z = k)$, $k = 2, 3, \ldots, 12$ bekommen wir

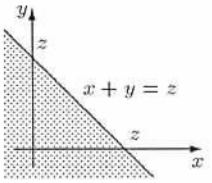
$$p_k = P(Z = k) = \sum_{i+j=k} p_{ij} \ ,$$

wobei die Summation über alle Kombinationen (i, j) mit $i + j = k$ geht. Dies lässt sich auch in die folgende Form bringen,

$$p_k = \sum_{i=1}^{k-1} p_{i,k-i} = \sum_{i=1}^{k-1} p_{i.}p_{.k-i} = \sum_{j=1}^{k-1} p_{k-j.}p_{.j} \ .$$

Ähnlich können wir den wichtigen allgemeinen Fall der Summe zweier unabhängiger Zufallsvariablen behandeln. Wir nehmen an, dass X und Y unabhängig seien und untersuchen die Verteilung von $Z = X + Y$. Im Fall von stetigen Verteilungen erhalten wir

$$\begin{aligned} \boxed{F_Z(z)} &= P(Z \leq z) = P(X + Y \leq z) \\ &= \int_{-\infty}^{\infty} \int_{-\infty}^{z-y} f(x, y)dxdy \\ &= \int_{-\infty}^{\infty} f_Y(y) \int_{-\infty}^{z-y} f_X(x)dxdy \\ &= \underline{\boxed{\int_{-\infty}^{\infty} f_Y(y)F_X(z - y)dy}} \ . \end{aligned}$$



Substitution von $t = y + x$ ergibt

$$F_Z(z) = \int_{-\infty}^{\infty} f_Y(y) \int_{-\infty}^{z} f_X(t - y)dtdy$$

und Vertauschen der Integrationsreihenfolge sowie Ableitung liefert

$$\boxed{f_Z(z)} = \frac{dF_Z(z)}{dz} = \frac{d}{dz} \int_{-\infty}^{z} \int_{-\infty}^{\infty} f_Y(y)f_X(t - y)dydt$$

$$\underline{= \boxed{\int_{-\infty}^{\infty} f_Y(y)f_X(z - y)dy}} \ .$$

Für den diskreten Fall können wir wie im obigen Beispiel angeben,

$$\boxed{p_k = \sum_{i=1}^{k-1} p_i . p_{.k-i}} \quad .$$

X und Y können natürlich in den Formeln beliebig vertauscht werden. Die eingerahmten Rechenvorschriften bezeichnet man auch als *Faltungen.*

Beispiel 4.37: Die Verteilung der Summe von n unabhängigen Zufallsvariablen erhält man durch vollständige Induktion:

(i) $X_i \sim Bi(1,p)$. Dann besitzt $Z = X_1 + X_2 + \ldots + X_n$ eine $Bi(n,p)$-Verteilung.

(ii) $X_i \sim P(\lambda)$. Dann ist $Z \sim P(n\lambda)$.

(iii) $X_i \sim N(\mu_i, \sigma_i^2)$. Dann ist $Z \sim N(\sum \mu_i, \sum \sigma_i^2)$.

(iv) $X_i \sim N(0,1)$. Dann ist $X_i^2 \sim \chi_1^2$, d.h. chiquadrat-verteilt mit einem Freiheitsgrad (siehe das Beispiel auf Seite 52). $Z = \sum X_i^2$ ist dann χ_n^2-verteilt, d.h. chiquadrat mit n Freiheitsgraden. (Diese Verteilung spielt eine wichtige Rolle als Testverteilung).

4.5.3 Erwartung

Sei h eine reellwertige Funktion des Zufallsvektors $\boldsymbol{X}$. Dann ist die *mathematische Erwartung* oder der *Mittelwert* von $h(\boldsymbol{X}) = h(X_1, \ldots, X_p)$ ähnlich wie im eindimensionalen Fall definiert durch

$$E[h(\boldsymbol{X})] = \int_{-\infty}^{\infty} \ldots \int_{-\infty}^{\infty} h(x_1, \ldots, x_p) f(x_1, \ldots, x_p) dx_1 \ldots dx_p$$

im stetigen Verteilungsfall, und im diskreten Fall durch

$$E[h(\boldsymbol{X})] = \sum_{i_1, \ldots, i_p} h(x_1^{(i_1)}, \ldots, x_p^{(i_p)}) p_{i_1 \ldots i_p} \quad ,$$

vorausgesetzt, dass das p-fache Integral über den Absolutbeträgen der Funktionswerte existiert, und bei der Reihe gilt Analoges.

Wie im eindimensionalen Fall kann man leicht zeigen, dass für 2 Funktionen h_1 und h_2 und Konstante a und b gilt

$$E[ah_1(\boldsymbol{X}) + bh_2(\boldsymbol{X})] = aE[h_1(\boldsymbol{X})] + bE[h_2(\boldsymbol{X})] \quad .$$

Wenn $p = 2$, d.h. $\boldsymbol{X} = (X_1, X_2)$ 2-dimensional ist, und h_1 nur von der ersten Komponente X_1 (d.h. $h_1(\boldsymbol{X}) = h_1(X_1)$) und h_2 nur von X_2 (d.h. $h_2(\boldsymbol{X}) = h_2(X_2)$) abhängt, dann kann man im stetigen Fall schließen (im diskreten ist es ähnlich), dass z.B. für h_1 gilt

$$E[h_1(\boldsymbol{X})] = \int h_1(x_1) \int f(x_1, x_2) dx_2 dx_1 =$$

$$= \int h_1(x_1) f_{X_1}(x_1) dx_1 = E[h_1(X_1)] \quad .$$

Da für h_2 das gleiche gilt, folgt für $a = b = 1$

$$E[h_1(X_1) + h_2(X_2)] = E[h_1(X_1)] + E[h_2(X_2)]$$

und insbesondere für Identitäten $h_1(x) = h_2(x) = x$

$$E(X_1 + X_2) = E(X_1) + E(X_2) \quad .$$

Diese Schlussweise lässt sich leicht auf n Zufallsvariable erweitern, und wir erhalten den *Additionssatz für Mittelwerte.* Der Mittelwert einer Summe von Zufallsvariablen, deren Mittelwerte existieren, ist gleich der Summe dieser Mittelwerte,

$$E(X_1 + X_2 + \cdots + X_n) = E(X_1) + E(X_2) + \cdots + E(X_n) \quad .$$

Beispiel 4.38: Für $i = 1, \ldots, n$ sei $X_i \sim Bi(1, p)$. Die Erwartung von X_i ist $EX_i = p$. Die Summe der X_i, $Z = X_1 + \cdots + X_n$, ist $Bi(n, p)$ verteilt und besitzt die Erwartung

$$EZ = EX_1 + \cdots + EX_n = np \quad .$$

Beim Produkt zweier Zufallsvariablen ist es nicht so einfach. Wenn z.B. X die Augenzahl eines geworfenen Würfels darstellt, so gilt

$$E(X \times X) = EX^2 = 1^2 \frac{1}{6} + 2^2 \frac{1}{6} + \cdots + 6^2 \frac{1}{6} = \frac{91}{6} \neq (\frac{7}{2})^2 = (EX)^2 \quad ,$$

wobei natürlich $EX = \frac{7}{2}$.

Wenn allerdings zwei Zufallsvariable X und Y unabhängig sind, so sieht man im stetigen Fall (und analog im diskreten Fall), dass

$$E(XY) = \int \int xy f(x, y) dx dy = \int x f_X(x) \int y f_Y(y) dy dx =$$

$$= \int x f_X(x) dx \int y f_Y(y) dy = (EX)(EY) \quad .$$

Dies lässt sich zum *Multiplikationssatz für Mittelwerte* erweitern: Für n unabhängige Zufallsvariable $X_1, \ldots, X_n$, deren Mittelwerte existieren, gilt

$$E(X_1 X_2 \ldots X_n) = (EX_1)(EX_2) \ldots (EX_n) \quad .$$

Betrachten wir nun die *Varianz* der Summe zweier Zufallsvariablen $Z = X + Y$, wobei $\sigma_X^2 = Var(X)$, $\sigma_Y^2 = Var(Y)$ und $\sigma_Z^2 = Var(Z)$ bezeichnen soll. Durch einfaches Ausmultiplizieren zeigt man schnell, dass

$$\begin{aligned} \sigma_Z^2 &= E(Z - EZ)^2 = E(X + Y - E(X + Y))^2 \\ &= E(X - EX)^2 + E(Y - EY)^2 + 2E[(X - EX)(Y - EY)] \\ &= \sigma_X^2 + \sigma_Y^2 + 2\sigma_{XY}, \end{aligned}$$

wobei

$$\sigma_{XY} = E[(X - EX)(Y - EY)] = E(XY) - (EX)(EY)$$

(die Erwartung der Kreuzprodukte) als *Kovarianz* der Zufallsvariablen X und Y bezeichnet wird.

Für den Fall der Unabhängigkeit von X und Y gilt E(XY)=(EX)(EY), sodass folgt

$$\sigma_{XY} = 0 \ ,$$

die Kovarianz verschwindet. (**Der umgekehrte Schluss ist nicht zulässig!**) Die Varianz von Z reduziert sich dann auf

$$\sigma_Z^2 = \sigma_X^2 + \sigma_Y^2 \ .$$

Dieses Ergebnis lässt sich auch auf die Summe von mehr als zwei Zufallsvariablen erweitern, was zum *Additionssatz für Varianzen* führt: Die Varianz einer Summe *unabhängiger* Zufallsvariablen, deren Varianzen existieren, ist gleich der Summe der Varianzen. Bemerkung: Konstante Faktoren gehen quadratisch in die Varianz ein, d.h. $Var(aX) = a^2 Var(X)$!

Beispiel 4.39: Für $i = 1, \ldots, n$ sei $X_i \sim Bi(1, p)$. Die Varianz von X_i ist $VarX_i = p(1-p)$, sodass für die Varianz der Summe $Z = X_1 + \ldots + X_n$, die $Bi(n, p)$ verteilt ist, folgt,

$$Var(Z) = Var(X_1) + \cdots + Var(X_n) = np(1-p) \ .$$

Die Kovarianz σ_{XY} von zwei Zufallsvariablen X und Y stellt ein Maß für die Abhängigkeit der beiden dar. Es ist allerdings günstiger, dieses Maß zu standardisieren, indem man es durch die Streuungen von X und Y dividiert. Wir erhalten so die *Korrelation*

$$\rho_{XY} = \frac{\sigma_{XY}}{\sigma_X \sigma_Y}$$

als dimensionslose Größe, die – wie man zeigen kann – Werte zwischen -1 und +1 annimmt.

Beispiel 4.40: Die theoretische Erwartung EX einer Zufallsvariablen X wird bei Vorliegen einer Stichprobe $x_1, \ldots, x_n$ durch den Mittelwert $\bar{x}$ approximiert. Die Varianz σ^2 wird durch

$$s_X^2 = \frac{1}{n-1} \sum (x_i - \bar{x})^2$$

angenähert. Bei Paaren von Messungen (x_i, y_i), die Werte einer zweidimensionalen Zufallsvariablen (X, Y) darstellen, wird die Kovarianz σ_{XY} durch

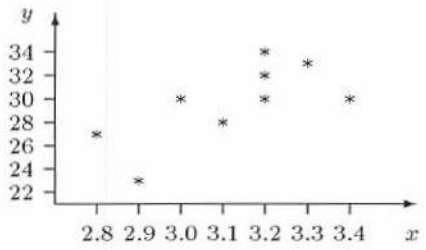
$$s_{XY} = \frac{1}{n-1} \sum (x_i - \bar{x})(y_i - \bar{y})$$

approximiert. Der (Stichproben-)*Korrelationskoeffizient* ergibt sich dann als

$$r_{XY} = \frac{s_{XY}}{s_X s_Y} \ .$$

Beispiel 4.41: Betrachten wir die Abhängigkeit des Eisengehaltes Y (in %) kieseliger Hämatiterze von der Dichte X (g/cm^3) und nehmen an, dass folgende Werte gemessen wurden (Quelle: H. Bottke, Bergbauwiss. 10, 1963, 377):

x_i	2.8	2.9	3.0	3.1	3.2	3.2	3.2	3.3	3.4
y_i	27	23	30	28	30	32	34	33	30



Wir berechnen die Größen:

$\bar{x} = \frac{1}{n}\sum x_i = 3.12$

$\bar{y} = 29.67$

$s_X^2 = \frac{1}{n-1}\sum(x_i - \bar{x})^2 = .0369$

$s_Y^2 = \frac{1}{n-1}\sum(y_i - \bar{y})^2 = 11.25$

$s_{XY} = \frac{1}{n-1}\sum(x_i - \bar{x})(y_i - \bar{y}) = .4458$

$r_{XY} = s_{XY}/(s_X s_Y) = .69$

Als Illustration der Korrelation ρ und seiner empirischen Approximation r stellen wir in der Abbildung 4.15 jeweils 100 künstlich erzeugte Werte mit verschiedenen Werten von ρ dar. (Die Stichproben wurden von einer bivariaten Normalverteilung mit $\sigma_X = 1$, $\sigma_Y = 1$ und $\rho = 0, .3, .5, .7, .9$ gewonnen).

4.6 Ein Beispiel: Zentraler Grenzwertsatz

Eine interessante Anwendung der Normalverteilung findet man bei der Betrachtung der Verteilung des arithmetischen Mittels. Dies soll an einem einfachen Versuch illustriert werden.

Wir nehmen an, dass die gesamte Population aus Werten 60, 61, 62, ..., 69 besteht, die gleichmäßig verteilt sind, d.h. gleiche Häufigkeiten aufweisen. Für den Versuch werfen wir 300 Kärtchen, von denen 30 die Aufschrift '60', 30 die Aufschrift '61' usw. erhalten, in eine Urne. Nach Durchmischung wird eine Stichprobe mit 2 Kärtchen (n=2) gezogen und das arithmetische Mittel $\bar{x}$ berechnet. Diese Prozedur (mit Zurücklegen) wird 200 mal wiederholt und mit den Werten von $\bar{x}$ ein Histogramm gezeichnet (siehe Abbildung 4.16). Das Ganze führen wir noch mit Stichprobengrößen $n = 4$, 8 und 16 durch. Die erhaltenen Häufigkeiten mit den Klassenmittelpunkten und einigen Kenngrößen sind in der Tabelle 4.1 ersichtlich.

Bezeichnet X die Zufallsvariable, die die Anschrift eines gezogenen Kärtchens repräsentiert, so bezeichnet $\bar{X}$ konsequenterweise das entsprechende arithmetische Mittel, das ebenfalls als Zufallsvariable zu interpretieren ist. Der mittlere Wert von X errechnet sich aus der Verteilung der Grundgesamtheit als $\mu = 64.5$ und die Varianz als $\sigma^2 = 8.25$. Man sieht aus der Tabelle, dass die empirischen Mittel von $\bar{X}$ für verschiedene n natürlich auch alle ungefähr gleich sind und sich nur um einen kleinen zufälligen Fehler unterscheiden. Die berechneten Varianzen $s_{\bar{X}}^2$

Tabelle 4.1: Häufigkeiten aus einem einfachen Versuch.

Gesamtheit n=1		n=2		n=4		n=8		n=16	
X	h	$\bar{X}$	h	$\bar{X}$	h	$\bar{X}$	h	$\bar{X}$	h
69	30	68.75	9	68.50	3	67.250	2	66.1875	3
68	30	67.75	12	67.75	5	66.625	4	65.8750	5
67	30	66.75	23	67.	4	66.	26	65.5625	21
66	30	65.75	37	66.25	21	67.375	36	65.2500	27
65	30	64.75	33	65.50	38	64.750	40	64.9375	24
64	30	63.75	38	64.75	35	64.125	43	64.6250	25
63	30	62.75	22	64.	53	63.500	30	64.3125	35
62	30	61.75	13	63.25	25	62.875	15	64.	26
61	30	60.75	13	62.50	8	62.250	3	63.6875	18
60	30			61.75	8	61.625	1	63.3750	15
								63.0625	1
Summe	300		200		200		200		200
Mittel von $\bar{X}$		64.66		64.23		64.57		64.58	
Varianz von $\bar{X}$		4.173		1.868		1.105		.509	
$s_{\bar{X}}$		2.043		1.367		1.051		.714	
$ns^2_{\bar{X}}$		8.35		7.47		8.84		8.14	

von $\bar{X}$ werden dagegen bei steigendem Stichprobenumfang n kleiner. Multipliziert man sie allerdings mit n, so erhält man ungefähr den gleichen Wert von $\sigma^2 = 8.25$. Betrachtet man die Histogramme der Mittel, so sieht man deutlich die Annäherung an die Glockenkurve. Dies wird noch durch den folgenden theoretischen Satz unterstrichen:

Satz: Besitzt die Verteilung der Grundgesamtheit eine endliche Varianz, so ist die Verteilung der arithmetischen Mittel von Zufallsstichproben approximativ normal, sofern der Stichprobenumfang genügend groß ist.

In den meisten praktischen Anwendungen besitzt die Verteilung eine endliche Varianz. Die Frage der Stichprobengröße ist schwieriger. Falls jedoch die zugrundeliegende Verteilung normal ist, dann ist auch die Verteilung der arithmetischen Mittel normal.

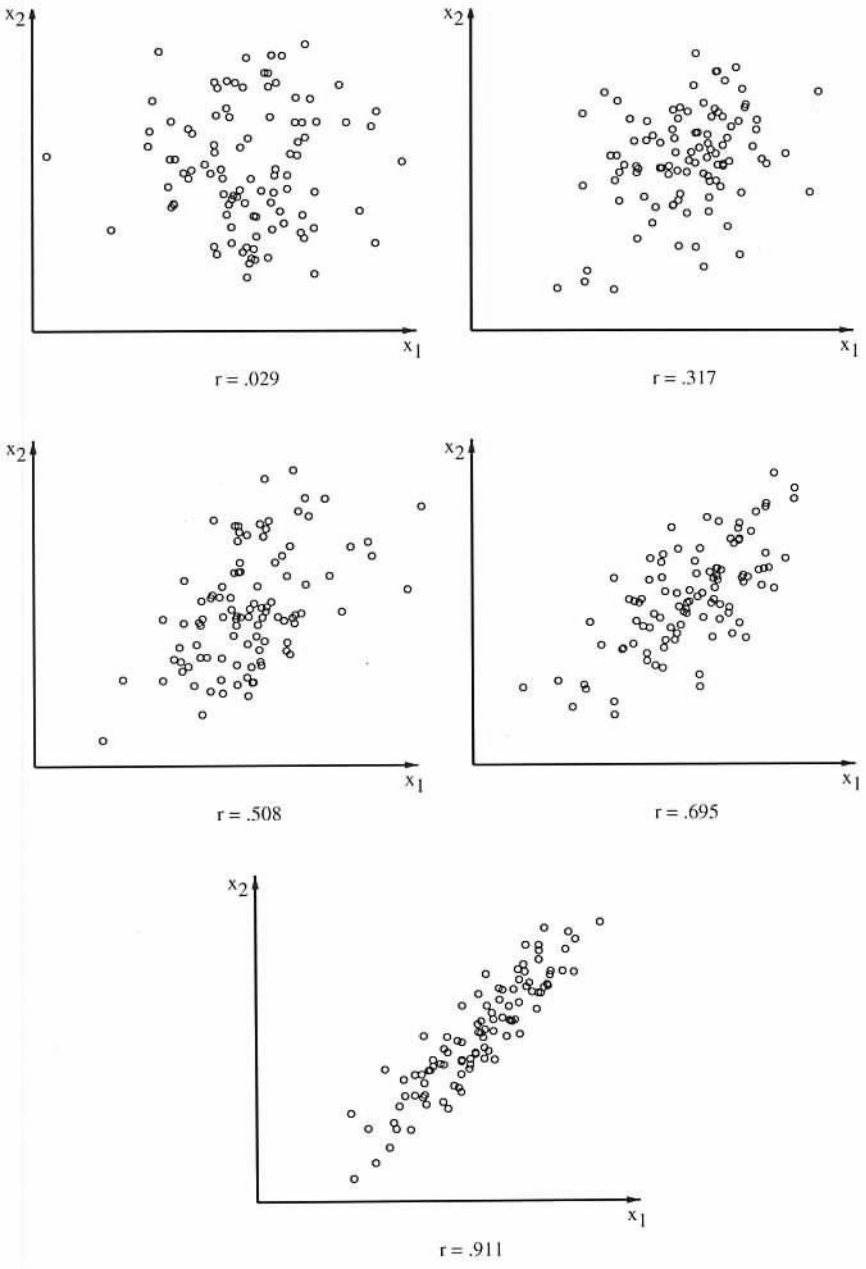


Abbildung 4.15: Streuungsdiagramme mit $\rho = 0,.3,.5,.7,.9$.

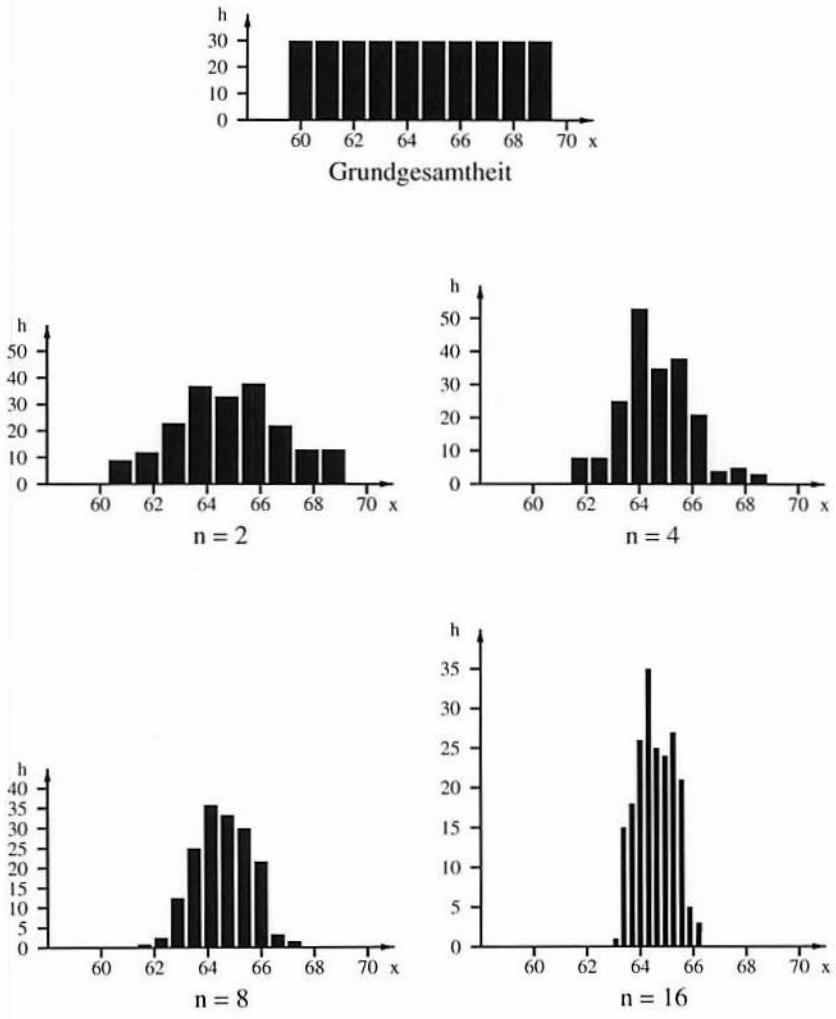


Abbildung 4.16: Histogramm für $\overline{x}_n$ und Ausgangsverteilung.

Kapitel 5

Analytische Statistik: Schätzungen und Tests

In der beschreibenden Statistik wird versucht, statistische Daten in eine möglichst übersichtliche Form zu bringen. Aus diesen Darstellungen können oft schon gewisse Schlüsse gezogen und Hypothesen aufgestellt werden.

In der analytischen Statistik soll die Verbindung zwischen der Theorie und der Wirklichkeit hergestellt werden. Dabei interessieren Fragen, inwieweit und wie Schlüsse von einer Stichprobe (allgemeiner von gesammelten Daten) auf die Grundgesamtheit gezogen werden können. Wenn die hypothetische Verteilung der Grundgesamtheit unbekannte Parameter enthält, möchte man wissen, wie und mit welcher Genauigkeit man auf diese Parameter schließen kann. Ist es bei einer bestimmten, geforderten Sicherheit überhaupt möglich, dass die Stichprobe von dieser Grundgesamtheit kommt? Kommen zwei Stichproben von zwei verschiedenen Grundgesamtheiten? Weisen Paare von Messungen Abhängigkeiten auf, etc.?

5.1 Stichproben

Wie im Abschnitt 3.4 ausgeführt wurde, stellt eine Stichprobe eine Untermenge einer Population dar. Nach der Einführung von Zufallsvariablen X kann ein Stichprobenwert x_i auch als *Realisation*, als konkret angenommener Wert von X aufgefasst werden. Die verschiedenen Stichprobenwerte $(x_1, \ldots, x_n)$ sind dann wiederholte Realisationen der Zufallsvariablen X, die normalerweise als unabhängig voneinander betrachtet werden. Der n-dimensionale Vektor $\boldsymbol{x} = (x_1, \ldots, x_n)$ der Stichprobenwerte kann allerdings auch als einzige Realisierung eines Zufallsvektors $\boldsymbol{X} = (X_1, \ldots, X_n)$, bei dem jedes Element X_i die *identische* Verteilung wie X aufweist, interpretiert werden.

Wenn z.B. die Besitzer eines Fernsehapparates in einem Land die Grundgesamtheit darstellen, dann könnte X_i die kodierte Antwort einer Person auf die Frage der Konsumierung eines bestimmten Programms sein. Um auf das Verhalten der

Grundgesamtheit schließen zu können, muss die Person aus allen Fernseherbesitzern zufällig ausgewählt werden, d.h. jede Person soll die gleiche Chance (Wahrscheinlichkeit) haben, ausgewählt zu werden. Nur dann liefern (zumindest einfache) statistische Verfahren vernünftige Ergebnisse. Normalerweise wird meistens auch gefordert, dass eine Person von den anderen Personen unabhängig ausgewählt wird. Dies bedeutet natürlich auch, dass es in diesem Beispiel theoretisch möglich wäre, dass zweimal die gleiche Person (zufällig) ausgewählt wird.

Die praktische, zufällige Auswahl von n Elementen aus einer endlichen Gesamtheit ist i.a. gar nicht so einfach. Häufig werden dazu Tafeln von Zufallszahlen (z.B. Rand Corporation: A Million Random Digits with 100,000 Normal Deviates. The Free Press, Clencoe, Ill. 1955) oder Pseudozufallszahlen eines Computers verwendet (linearer Kongruenzgenerator $z_{i+1} = az_i mod(m)$).

Wir werden zunächst einfache statistische Verfahren besprechen und dabei unter *Stichprobe* den n-dimensionalen Zufallsvektor $(X_1, \ldots, X_n)$ mit unabhängig und identisch verteilten Elementen X_i verstehen. Die *Stichprobenwerte* $(x_1, \ldots, x_n)$ stellen eine Realisation von $(X_1, \ldots, X_n)$ dar.

5.2 Punktschätzungen

Nehmen wir an, dass die Verteilung der Stichprobenelemente X_i einen unbekannten Parameter θ enthält ($X_i \sim F_\theta$) und dass es eine Funktion t gibt, die aus den Stichprobenwerten den Wert von θ näherungsweise berechnet. Die Approximation wäre also

$$\hat{\theta} = t(x_1, \ldots, x_n) \quad .$$

Allgemein bezeichnet man eine Funktion der Stichprobe $T = t(X_1, \ldots, X_n)$ als *Statistik*, die natürlich wieder eine Zufallsvariable ist. Im Falle der Verwendung zur näherungsweisen Bestimmung (Schätzung) gewisser Kenngrößen spricht man von einer *Schätzfunktion* t oder kurz von einem *Schätzer*. Eine Realisation des Schätzers $T = t(X_1, \ldots, X_n)$, etwa $t = t(x_1, \ldots, x_n)$, heißt *Schätzwert* oder *Schätzung*.

Beispiel 5.1: Gegeben sei eine Zufallsvariable X, deren Verteilung einen Parameter μ enthält, der den Erwartungswert $\mu = EX$ darstellt (z.B. die Normalverteilung $N(\mu, \sigma^2)$). Der Wert von μ als Mittel von X könnte durch das arithmetische Mittel aus den Stichprobenwerten approximiert werden, nämlich

$$\hat{\mu} = \hat{\mu}(x_1, \ldots, x_n) = \frac{1}{n} \sum_{i=1}^{n} x_i = \bar{x} \quad .$$

Der Schätzer wird als

$$\bar{X} = \bar{X}(X_1, \ldots, X_n) = \frac{1}{n} \sum_{i=1}^{n} X_i$$

definiert und eine Realisation mit $\bar{x}$ bezeichnet.

Da ein Schätzer eine Zufallsvariable ist, können wir seine mathematische Erwartung untersuchen. Man wird bei wiederholtem Schätzen eines Parameters als Resultat auch gerade den Parameter erwarten (im populären Sinn). Wenn also t den Parameter θ schätzt, so soll gelten

$$E(T) = E_\theta t(X_1, \ldots, X_n) = \theta \quad .$$

In diesem Fall spricht man von einem *erwartungstreuen* oder *unverzerrten* Schätzer.

Beispiel 5.2: Sei $\mu = EX_i$, $i = 1, \ldots, n$. Dann gilt

$$E\bar{X} = E(\frac{1}{n}\sum X_i) = \frac{1}{n}\sum EX_i = \mu \quad .$$

Das arithmetische Mittel $\bar{X}$ der Stichprobe ist also ein erwartungstreuer Schätzer des Mittels der Verteilung oder des Populationsmittels. Wenn die Varianz $\sigma^2 = Var(X_i)$ existiert, kann man wegen

$$Var(\bar{X}) = Var(\frac{1}{n}\sum X_i) = \frac{1}{n^2}Var(\sum X_i) = \frac{1}{n^2}\sum VarX_i = \frac{\sigma^2}{n}$$

zeigen, dass $\bar{X}$ auch einen *konsistenten* Schätzer darstellt. Ähnliches gilt für den Schätzer der Varianz

$$S^2 = \frac{1}{n-1}\sum(X_i - \bar{X})^2 \quad ,$$

der konsistent ist. Wie man sich leicht überzeugt, gilt auch die Erwartungstreue

$$ES^2 = \sigma^2 \quad .$$

Beispiel 5.3: Wenn die Verteilung symmetrisch ist, stellt der Median $\tilde{X}$ ebenfalls einen konsistenten und erwartungstreuen Schätzer der Erwartung dar.

Die Güte eines Schätzers hängt von seiner Variabilität ab, d.h. je kleiner seine Varianz, desto besser. Zum Beispiel gilt für $\bar{X}$ immer $Var\bar{X} = \sigma^2/n$, sofern die Varianz existiert. Dagegen kann man für den Median $\tilde{X}$ im Fall einer zugrundeliegenden Normalverteilung zeigen, dass die Varianz für große n

$$Var_N(\tilde{X}) = \frac{\sigma^2}{.64n} \quad .$$

Dies bedeutet, dass man bei Verwendung des Medians (und zugrundeliegender Normalverteilung) ca. 1/3 mehr Beobachtungen braucht, um die gleiche „Genauigkeit" wie bei Verwendung von $\bar{X}$ zu erhalten. Es ist natürlich klar, dass die Situation bei anderen Verteilungen völlig verschieden sein kann. Bei einer bestimmten Verteilung wird man allerdings bestrebt sein, einen Schätzer mit möglichst geringer Varianz zu verwenden.

Man sagt, ein erwartungstreuer Schätzer ist *wirksam* oder *effizient*, wenn er die kleinstmögliche Varianz aufweist.

Beispiel 5.4: Bei zugrundeliegender Normalverteilung $N(\mu, \sigma^2)$ stellt $\bar{X}$ einen effizienten Schätzer für μ dar. S^2 ist für σ^2 nur asymptotisch effizient (d.h. für $n \to \infty$).

Es gibt verschiedene Verfahren, um brauchbare Schätzer für Parameter einer Verteilung zu finden. Die *Maximum-Likelihood-Methode* ist die wichtigste. Sie wählt im wesentlichen jenen Wert des Parameters, der die Stichprobe als „wahrscheinlichstes" Resultat erscheinen lässt. Zur genauen Beschreibung der Methode brauchen wir den Begriff der *Likelihood-Funktion*, die mit der Stichprobe $(x_1, \ldots, x_n)$ im Falle einer stetigen Verteilung als

$$\ell(\theta; x_1, \ldots, x_n) = f_\theta(x_1) f_\theta(x_2) \ldots f_\theta(x_n)$$

definiert wird. $f_\theta(x_i)$ bezeichnet dabei die Dichte der zugrundeliegenden Verteilung, die von θ abhängen soll. θ wird nun so gewählt, dass dieses Produkt der Dichtefunktion ein Maximum annimmt. Bei Vorliegen einer diskreten Verteilung wird analog die Wahrscheinlichkeitsfunktion verwendet. Die Methode illustrieren wir hier an einem Beispiel.

Beispiel 5.5: Die zugrundeliegende Verteilung sei die Normalverteilung $N(\mu, \sigma^2)$; μ und σ sind zu schätzen. Die Likelihood-Funktion ergibt sich als

$$\ell = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x_1-\mu)^2}{2\sigma^2}} \ldots \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x_n-\mu)^2}{2\sigma^2}} = (\frac{1}{\sqrt{2\pi}})^n \frac{1}{\sigma^n} e^{-\frac{q}{2\sigma^2}} \quad ,$$

wobei q die Summe der quadratischen Abweichungen

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

bezeichnet. Statt ℓ zu maximieren (hier bezüglich μ *und* σ) können wir auch den Logarithmus $\ln \ell$ maximieren. Durch Logarithmieren ergibt sich

$$\ln \ell = -n \ln(\sqrt{2\pi}) - n \ln \sigma - \frac{q}{2\sigma^2} \quad .$$

Eine notwendige Bedingung für das Maximum ist das Verschwinden der 1. Ableitung. Die Differentiation bezüglich μ liefert

$$\frac{\partial \ln \ell}{\partial \mu} = -\frac{1}{2\sigma^2} \frac{\partial q}{\partial \mu} = \frac{1}{\sigma^2} \sum_{i=1}^{n} (x_i - \mu) = 0$$

und bezüglich σ

$$\frac{\partial \ln \ell}{\partial \sigma} = -\frac{n}{\sigma} + q/\sigma^3 = 0 \quad .$$

Aus der ersten Gleichung folgt die Lösung für μ,

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

was wiederum in die zweite Gleichung eingesetzt

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

ergibt. Interessanterweise ist dieser Maximum-Likelihood-Schätzer für σ^2 nicht erwartungstreu (n statt $n-1$), aber für große n spielt das natürlich keine Rolle.

5.3 Intervallschätzungen

Eine praktisch durchgeführte Punktschätzung ist nur bei Angabe der Genauigkeit des Resultates sinnvoll. Deshalb werden in der sogenannten „Fehlerrechnung" häufig der geschätzte Mittelwert und die Standardabweichung s in der Form

$$\bar{x} \pm s/\sqrt{n}$$

(oder manchmal $\bar{x} \pm s$) angegeben, ohne Berücksichtigung der wirklich zugrundeliegenden Verteilung. $s/\sqrt{n}$ wird auch als *mittlerer Fehler* oder *mittlerer quadratischer Fehler* bezeichnet.

Diese Angabe ist aber manchmal, insbesondere bei schiefen Verteilungen, irreführend und es ist günstiger, *Konfidenz-* oder *Vertrauensintervalle* zu verwenden. Sie geben einen Bereich an, in dem der unbekannte Parameter mit einer bestimmten Wahrscheinlichkeit liegt.

Bei einer Stichprobe aus einer Population mit Mittel μ und Varianz σ^2 besitzt die Verteilung des arithmetischen Mittels $\bar{X}$ das Mittel μ und die Varianz σ^2/n. Die Form der Verteilung ist - zumindest bei entsprechend großen Stichproben - ungefähr normal $N(\mu, \sigma^2/n)$. Deshalb kann aus Tabellen (z.B. Tabelle A des Anhangs) die Wahrscheinlichkeit gefunden werden, dass die standardisierte Variable

$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}}$$

in einen bestimmten Bereich fällt. Bezeichnen wir mit z_α das α-Quantil der N(0,1)-Verteilung, d.h.

$$P(Z \leq z_\alpha) = \alpha \quad ,$$

wobei folgende Werte häufig Verwendung finden:

α	.95	.975	.995	.9995
z_α	1.645	1.960	2.576	3.291

,

dann ist die Wahrscheinlichkeit, dass Z in den mittleren Bereich fällt,

$$P(z_{\frac{\alpha}{2}} \leq \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \leq z_{1-\frac{\alpha}{2}}) = P(-z_{1-\frac{\alpha}{2}} \leq \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \leq z_{1-\frac{\alpha}{2}}) = 1 - \alpha \quad .$$

Die erste Ungleichung umgeformt ergibt

$$\mu \leq \bar{X} + z_{1-\frac{\alpha}{2}}\sigma/\sqrt{n}$$

und die zweite

$$\bar{X} - z_{1-\frac{\alpha}{2}}\sigma/\sqrt{n} \leq \mu \quad ,$$

sodass

$$P(\bar{X} - z_{1-\frac{\alpha}{2}}\sigma/\sqrt{n} \leq \mu \leq \bar{X} + z_{1-\frac{\alpha}{2}}\sigma/\sqrt{n}) = 1 - \alpha \quad .$$

Dies bedeutet, dass die Wahrscheinlichkeit, dass das Intervall

$$(\bar{X} - z_{1-\frac{\alpha}{2}}\sigma/\sqrt{n}, \bar{X} + z_{1-\frac{\alpha}{2}}\sigma/\sqrt{n})$$

das wahre Mittel überdeckt, gleich 1-α ist. Dieses Intervall ist das 100(1-α)% Konfidenzintervall für μ. 1-α heißt auch *Überdeckungswahrscheinlichkeit* und α *Konfidenzzahl.*

Beispiel 5.6: Angenommen, σ sei bekannt als $\sigma = 2.80$ und 16 Beobachtungen wurden gemacht, die $\bar{x} = 15.70$ liefern. Das 95%-Konfidenzintervall ($\alpha = .05$) errechnet sich dann als

$$(15.70 - 1.96\frac{2.80}{\sqrt{16}}, 15.70 + 1.96\frac{2.80}{\sqrt{16}}) = (15.70 - 1.37, 15.70 + 1.37)$$

$$= (14.33, 17.07) \quad .$$

Manchmal schreibt man das Resultat auch als

$$\mu = 15.70 \pm 1.37 \quad ,$$

was aber eindeutig als Konfidenzintervall identifiziert und nicht mit dem obigen mittleren Fehler verwechselt werden sollte.

Übung 5.1: Man bestimme das 95%-Konfidenzintervall für das Mittel der Daten aus der Übung auf Seite 21, wobei angenommen wird, dass die Standardabweichung $\sigma = 1.86$ bekannt ist.

Beispiel 5.7: (Bestimmung des Stichprobenumfangs): Die Genauigkeit einer Schätzung nimmt natürlich mit dem Stichprobenumfang n zu. Die obige Fragestellung des Konfidenzintervalls bei gegebener Stichprobe lässt sich auch umdrehen: Wie groß muss diese bei gegebener maximaler Länge d des Konfidenzintervalls sein? Aus $d = 2z_{1-\frac{\alpha}{2}}\sigma/\sqrt{n}$ erhalten wir sofort

$$n = (2z_{1-\frac{\alpha}{2}}\sigma/d)^2 \quad .$$

Im vorigen Beispiel würde man z.B. für $d = 1$

$$n = (2 \times 1.96 \times 2.80/1)^2 = 120.5 \sim 121$$

errechnen.

Konfidenzintervalle können für beliebige Parameter θ gefunden werden, indem man 2 Größen (Schätzstatistiken) U und O sucht, die die Grenzen angeben und für die bei gegebener Überdeckungswahrscheinlichkeit 1-α gilt

$$P(U \leq \theta \leq O) = 1 - \alpha \ .$$

Die Grenzen U und O des Konfidenzintervalls (U, O) sind Funktionen der Stichprobe.

Wenn in unserem obigen Beispiel der Normalverteilung die Standardabweichung unbekannt ist und aus der Stichprobe geschätzt werden muss, wird das Konfidenzintervall länger, weil die Information über σ der Stichprobe entnommen wird. Da wir die Zufallsvariablen X_i nicht mehr mit dem bekannten σ standardisieren können, benötigen wir folgende, wesentliche, theoretische Hilfsmittel:

$X_1, \ldots, X_n$ seien unabhängige, identisch normalverteilte Zufallsvariable. Dann sind die beiden Variablen

$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}}$$

und

$$Y = \frac{1}{\sigma^2} \sum_{i=1}^{n} (X_i - \bar{X})^2$$

unabhängig verteilt, und Y besitzt eine Chi-Quadrat-Verteilung χ^2_{n-1} mit $n-1$ Freiheitsgraden. Die Variable

$$T = \frac{Z}{\sqrt{Y/(n-1)}}$$

ist *(Student-)t*-verteilt mit $n-1$ Freiheitsgraden (t_{n-1}).

Setzen wir den üblichen Schätzer

$$S^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X})^2$$

für σ^2 in die Definition der Variablen Y ein und substituieren wir noch Z und Y in T, so sehen wir, dass

$$T = \frac{\bar{X} - \mu}{S/\sqrt{n}}$$

t_{n-1} verteilt ist. Dieses Resultat können wir ähnlich wie oben zur Konstruktion eines Konfidenzintervalles für μ bei unbekanntem σ verwenden. Bezeichnen wir mit $t_{n-1;\alpha}$ das α–Quantil der t_{n-1}-Verteilung, das wieder Tabellen entnommen werden kann (Tabelle A des Anhangs), dann erhalten wir

$$P(-t_{n-1;1-\frac{\alpha}{2}} \leq \frac{\bar{X} - \mu}{S/\sqrt{n}} \leq t_{n-1;1-\frac{\alpha}{2}}) = 1 - \alpha$$

oder umgeformt,

$$P(\bar{X} - t_{n-1;1-\frac{\alpha}{2}} S/\sqrt{n} \leq \mu \leq \bar{X} + t_{n-1;1-\frac{\alpha}{2}} S/\sqrt{n}) = 1 - \alpha \ ,$$

und das Konfidenzintervall[1] mit Überdeckungswahrscheinlichkeit 1-α ist

$$(\bar{X} - t_{n-1;1-\frac{\alpha}{2}} S/\sqrt{n}, \bar{X} + t_{n-1;1-\frac{\alpha}{2}} S/\sqrt{n}) \ .$$

Beispiel 5.8: Es seien Daten wie im Beispiel auf Seite 76 gegeben, nur σ sei unbekannt, jedoch aus den Daten als $\hat{\sigma} = s = 2.80$ geschätzt. Da $n = 16$ und ein 95%-Konfidenzintervall gesucht ist, benötigen wir $t_{15;.975}$, was aus Tabelle A als $t_{15;.975} = 2.131$ folgt. Das entsprechende Konfidenzintervall wird daher

$$(15.70 - 2.131\frac{2.80}{\sqrt{16}}, 15.70 + 2.131\frac{2.80}{\sqrt{16}}) = (15.70 - 1.49, 15.70 + 1.49)$$

$$= (14.21, 17.19) \ .$$

5.4 Tests von Hypothesen

Schätzungen, wie wir sie in den letzten Abschnitten betrachtet haben, sind nur unter gewissen Annahmen sinnvoll. Viele können gut mit statistischen Tests geprüft werden. In der Statistik verstehen wir unter *Hypothese* eine Annahme über die Verteilung einer Zufallsvariablen. Ein Beispiel einer Hypothese wäre die Annahme, dass die Verteilung einen gewissen Mittelwert (= 23.2) aufweist. Ein statistischer *Test* einer Hypothese ist dann ein Prüfverfahren, nach dem die Hypothese *angenommen* oder *verworfen* werden kann; man kann ihr Vertrauen schenken und sie für richtig halten oder nicht. Diese Tests dienen als Entscheidungshilfe.

Typische Beispiele von Hypothesen, die mittels statistischer Tests untersucht werden, sind:

(a) Eine Reihe von Beobachtungen stellt eine Stichprobe einer Population mit einem spezifizierten Mittel μ dar (eine Auswahl von Glühlampen hat Standard-Qualität, die Durchschnittsintelligenz der Studenten dieser Klasse ist gleich der aller Studenten).

(b) Eine Reihe von Beobachtungen stellt eine Stichprobe einer Population mit einer spezifizierten Varianz σ^2 dar (diese Klasse ist bezüglich der Intelligenz genauso variabel wie andere Klassen).

(c) Zwei Gruppen von Beobachtungen stellen Stichproben von Populationen mit gleichem Mittel dar (Methode A ist gleich oder besser als Methode B).

[1] R: `confint(lm(Daten ~ 1))`

5.4.1 Mittel einer Population

Tests bezüglich des Mittels μ der Population stützen sich auf das Stichprobenmittel $\bar{X}$ und dessen Verteilung. Die Hypothese soll lauten $\mu = \mu_0$, wobei μ_0 ein spezieller Wert ist. Wenn sie richtig ist, werden die Werte von $\bar{X}$ zufällig um μ_0 herum streuen, nämlich mit der Standardabweichung $\sigma/\sqrt{n}$, wobei σ^2 wie üblich die Varianz der Population bezeichnet. Liegt der Wert von $\bar{X}$ „zu weit" von μ_0 entfernt, wird man vielleicht nicht mehr bereit sein, die Hypothese aufrecht zu erhalten und wird sie *verwerfen*. Jene Werte, die zur Verwerfung der Hypothese führen, definieren den *kritischen Bereich* des Tests. Die Wahrscheinlichkeit, dass (unter der Annahme der Hypothese) $\bar{X}$ in diesen Bereich fällt, wird *Signifikanzzahl* oder *Signifikanzniveau* genannt. Die Statistik, auf die sich der Test stützt (hier $\bar{X}$), heißt *Teststatistik*.

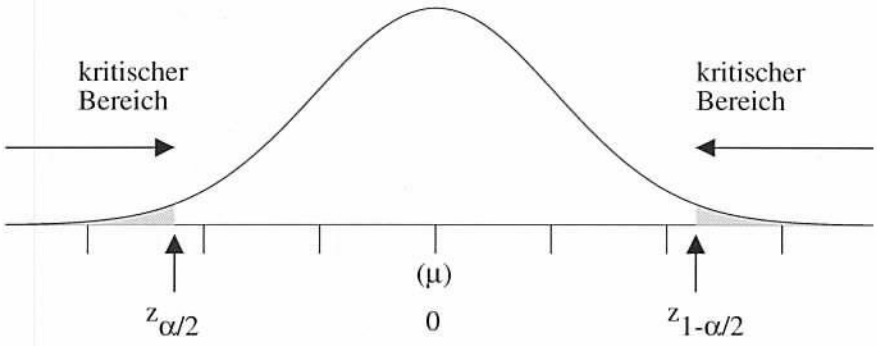


Abbildung 5.1: Verteilung einer Teststatistik.

Wir setzen nun voraus, dass X normalverteilt ist, sodass unsere Voraussetzung besagt: $X \sim N(\mu_0, \sigma^2)$. Dann gilt bekanntlich $\bar{X} \sim N(\mu_0, \sigma^2/n)$ und es ist besser mit der standardisierten Größe $Z = (\bar{X} - \mu_0)/(\sigma/\sqrt{n})$ zu arbeiten, die $N(0,1)$ verteilt ist. Wenn nun ein konkreter Wert von Z absolut größer als eine bestimmte Schranke (etwa $z_{1-\frac{\alpha}{2}}$ für ein vorher festgelegtes α) ist, also in den kritischen Bereich fällt, wird die Hypothese verworfen. α wird im allgemeinen klein gewählt werden (z.B. .1, .05, .01), und z_α bezeichnet wieder das α-Quantil der $N(0,1)$-Verteilung. Bei Zutreffen der Hypothese gilt

$$P(\mid Z \mid > z_{1-\frac{\alpha}{2}}) = \alpha \quad ,$$

d.h. Z fällt mit Wahrscheinlichkeit α in den kritischen Bereich. Man überlege sich, dass im Falle einer unbekannten Varianz σ^2, wie bei der Berechnung von Konfidenzintervallen, σ^2 aus der Stichprobe geschätzt und z durch t_{n-1} ersetzt werden kann (Einstichproben-*t-Test*[2]).

Als erste Zusammenfassung können wir auf folgende Schritte bei der Durchführung eines Tests einer Hypothese hinweisen:

[2] ℝ: `t.test(Daten, mu=70)`

- Formulierung der Voraussetzungen und der Hypothese (z.B. $H_o : \mu = \mu_o$)
- Wahl des Signifikanzniveaus α
- Wahl der Teststatistik und des kritischen Bereiches
- Präsentierung der Rechnungen
- Vollständige Angabe von Schlussfolgerungen.

Beispiel 5.9: Die Daten aus dem Beispiel auf Seite 76 seien gegeben (Normalverteilung, $\sigma = 2.80$ bekannt, $\bar{x} = 15.70$ aus 16 Datenpunkten gemessen). Es sei ein Anlass zur Aufstellung der Hypothese bezüglich des Mittelwertes $\mu_0 = 14$ gegeben. Muss diese Hypothese auf Grund der gemessenen Werte verworfen werden? Wir wählen als Signifikanzniveau $\alpha = .05$. Die Teststatistik ist $Z = (\bar{X} - \mu_0)/(\sigma/\sqrt{n})$ und der kritische Bereich wird durch $\mid Z \mid > z_{1-\frac{\alpha}{2}}$ definiert. In unserem Fall ist der Wert der Teststatistik

$$z = (15.70 - 14)/(2.80/4) = 2.43 \quad ,$$

der absolut größer ist als $z_{1-\frac{\alpha}{2}}$=1.96. Man sagt, dass auf dem Signifikanzniveau α = .05 der Wert der Teststatistik *signifikant* ist (oder der errechnete Wert $\bar{x} = 15.70$ *signifikant* von 14 verschieden ist). Als Schlussfolgerung gilt also, dass bei diesem α die Hypothese $\mu = 14$ verworfen werden muss.

Hätten wir $\alpha = .01$ gewählt, sodass

$$z_{1-\frac{\alpha}{2}} = 2.576 \quad ,$$

dann wäre das Resultat des Tests nicht signifikant. Die Hypothese könnte also nicht verworfen werden.

5.4.2 Verschiedene Arten von Fehlern

Das Signifikanzniveau α gibt die Wahrscheinlichkeit an, mit der die Hypothese verworfen wird, obwohl sie richtig ist. Die Schlussfolgerung ist ein Fehler, und man bezeichnet daher auch α als Fehlerwahrscheinlichkeit 1. Art oder kurz (aber fälschlich) als Fehler *1. Art*. Es kann natürlich auch sein, dass die Hypothese falsch ist. Wenn wir sie verwerfen, ist die Entscheidung richtig, ansonsten falsch. Im letzten Fall spricht man von einem Fehler *2. Art*. Bezeichnen wir die Wahrscheinlichkeit des Auftretens eines Fehlers 2. Art mit β, so können α und β entsprechend der zugrundeliegenden Hypothese interpretiert werden, wie aus der folgenden Entscheidungstabelle mit den Wahrscheinlichkeiten ersichtlich ist.

		Hypothese	
		richtig	falsch
Entscheidung:	annehmen	$1 - \alpha$	β
	ablehnen	α	$1 - \beta$
		1	1

In der Praxis wird man versuchen, die Fehlerwahrscheinlichkeiten 1. und 2. Art möglichst klein zu halten. Ein Verkleinern von beiden kann aber i.a. nur durch Vergrößerung des Stichprobenumfangs n erreicht werden.

Um β definieren zu können, müssen wir über die zugrundeliegende „alternative" Wahrscheinlichkeitsstruktur Bescheid wissen. Man stellt daher *Alternativ-* oder *Gegenhypothesen* auf (im Gegensatz dazu heißt die ursprüngliche Hypothese *Null-Hypothese*). Betrachten wir der Einfachheit halber wieder das vorige Beispiel des Tests bezüglich des Mittels einer Normalverteilung. Halten wir einmal die Fehlerwahrscheinlichkeit 1. Art und den kritischen Bereich fest, dann wird β aber noch vom wahren Wert des Mittels μ abhängen. Liegt μ nahe μ_0, so wird die Wahrscheinlichkeit der Annahme der Null-Hypothese, d.h. der Wert von β, groß sein, dagegen wenn μ sehr weit von μ_0 entfernt ist, wird β relativ klein ausfallen. Der Wert von β wird auch als *Operationscharakteristik* des Tests bezeichnet.

Beispiel 5.10: Angenommen, es sei bekannt, dass die Körpergröße von Männern eine Standardabweichung von $\sigma = 3$ *Zoll* $= 3''$ aufweist, und es werde die Hypothese aufgestellt, dass das Mittel der Gesamtheit $\mu_0 = 67''$ beträgt, d.h. $H_0 : \mu = 67''$. Es werden $n = 25$ Männer gemessen und die Hypothese soll auf dem Niveau $\alpha = .05$ getestet werden. Nachdem die Streuung bekannt ist, kann die hypothetische Verteilung von $\bar{X}$ $(\sigma/\sqrt{n} = 3/\sqrt{25} = .6)$ dargestellt werden. In der Abbildung 5.2 wird diese mit dem kritischen Bereich (i) mit $\mu = 67''$ und (ii) mit $\mu = 68''$ illustriert. Bei $\mu = 68''$ berechnet man $\beta = .62$ und $1 - \beta$ als 38%.

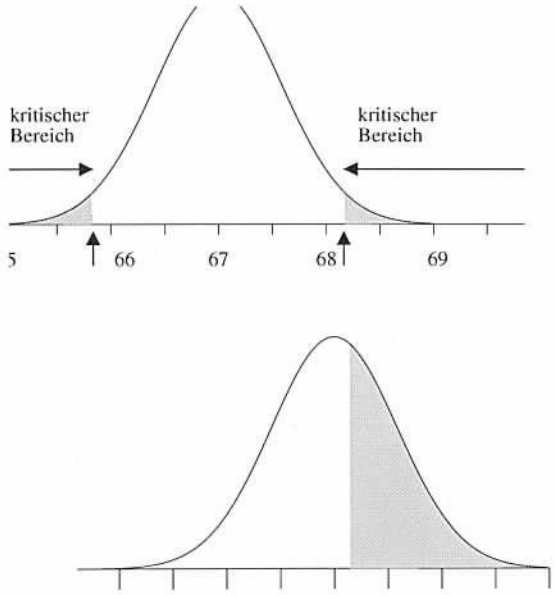


Abbildung 5.2: Kritischer Bereich bei verschiedenen zugrundeliegenden Verteilungen.

1-β als Funktion von μ aufgetragen heißt *Macht (Schärfe)* des Tests. Diese hängt natürlich stark vom Stichprobenumfang n ab, was in der Abbildung 5.3

illustriert wird.

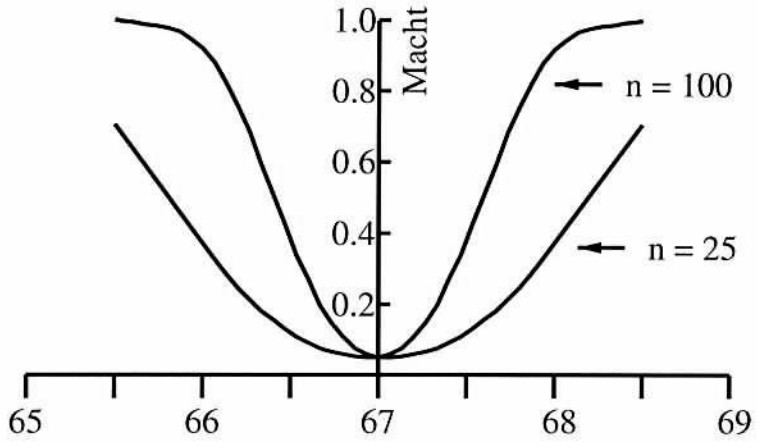


Abbildung 5.3: Macht eines Tests bei verschiedenen Stichprobengrößen.

In der Praxis ist es meist schwierig, beide Fehlerwahrscheinlichkeiten α und β klein zu machen. α kann man i.a. klein wählen, aber solange man über β, das dabei sehr groß sein kann, nichts weiß, sollte man vorsichtig sein. Man kann die Schwierigkeit sprachlich hervorheben, indem man nicht von „Hypothese annehmen", sondern nur von „Hypothese *nicht* verwerfen" spricht.

5.4.3 Typen von Alternativen

Bei den Tests bezüglich des Mittels einer normalverteilten Population wurde die Nullhypothese als

$$H_0 : \mu = \mu_0$$

und die Alternativ-Hypothese allgemein als

$$H_1 : \mu \neq \mu_0$$

formuliert. Diese „zweiseitige Alternative" tritt z.B. bei Messungen auf, die weder zu klein noch zu groß sein dürfen, wie etwa die Durchmesser von Wellen. Den kritischen Bereich haben wir dann für $Z = (\bar{X} - \mu_0)/(\sigma/\sqrt{n})$ mit

$$\mid Z \mid > z_{1-\frac{\alpha}{2}}$$

angegeben. Der kritische Bereich für $\bar{X}$ folgt daraus als

$$\bar{X} < \mu_0 - z_{1-\frac{\alpha}{2}}\sigma/\sqrt{n} \quad \text{und} \quad \bar{X} > \mu_0 + z_{1-\frac{\alpha}{2}}\sigma/\sqrt{n} \ .$$

Dies weist eine überraschende Vergleichbarkeit mit dem Konfidenzintervall für μ auf:

$$\bar{X} - z_{1-\frac{\alpha}{2}}\sigma/\sqrt{n} \leq \mu \leq \bar{X} + z_{1-\frac{\alpha}{2}}\sigma/\sqrt{n} \ .$$

Die Grenzen des Konfidenzintervalls und des kritischen Bereichs fallen zusammen. Wenn also μ_0 in das Konfidenzintervall fällt, wird auch dieser Test die Hypothese $H_0 : \mu = \mu_0$ nicht verwerfen und umgekehrt. Dies gilt allerdings i.a. nicht für anders formulierte Tests.

Andere Möglichkeiten der Alternativ-Hypothese sind

$$H_1 : \mu > \mu_0$$

oder

$$H_1 : \mu < \mu_0 \ .$$

Man spricht von *einseitigen* Alternativen bzw. von einseitigen Tests. Diese finden z.B. bei Festigkeitsuntersuchungen Anwendung, wo nur eine bestimmte Mindestfestigkeit gefordert wird. Hier wird der kritische Bereich auch nur auf einer Seite liegen, und statt $z_{1-\frac{\alpha}{2}}$ verwendet man $z_{1-\alpha}$.

5.5 Anteile: Schätzungen und Tests

Die Methode der Schätzung des Mittel einer Verteilung kann auch zur Schätzung des Anteils (Verhältnisses) von Individuen einer Population mit bestimmten Charakteristiken verwendet werden. Ein Beispiel wäre der Anteil der Wähler, die für eine bestimmte Partei sind.

Der relative Anteil wird als spezieller Fall des arithmetischen Mittels $\bar{X}$ interpretiert, indem jedem Individuum die Zahl 1 zugeordnet wird, wenn es die betrachtete Charakteristik aufweist, und sonst die Zahl 0 („scores"). Der Anteil $\hat{p}$ berechnet sich dann als Mittel dieser Zahlen. Die Anzahl der „Einser" in einer Stichprobe vom Umfang n ist $Bi(n,p)$-verteilt. Der Mittelwert ist daher np und die Varianz $\sigma^2 = np(1-p)$. Die Statistik $\bar{X}$ (Stichprobe) besitzt das Mittel p (den wahren Anteil) und die Standardabweichung (den Standardfehler)

$$\sigma/n = \sqrt{p(1-p)/n} \ .$$

Für exakte Tests und Schätzungen bräuchte man die Werte der Verteilungsfunktion für die betrachtete Binomialverteilung, die aber besonders für große n aufwendig zu berechnen sind. Daher behilft man sich oft mit Kurvenblättern (Nomogrammen). Aus der Abbildung 5.4 können z.B. 95%-Konfidenzintervalle bzw. 2-seitige Tests zum Signifikanzniveau 5% abgelesen werden. Meistens wird jedoch bei genügend großem n die Normalverteilungsapproximation für $\bar{X}$ ausgenützt. Leider enthält die Standardabweichung den unbekannten Parameter p. Man bemerkt aber, dass der Faktor $p(1-p)$ nie größer als $1/4$ werden kann, sodass sich als Abschätzung nach oben für σ/n die Größe

$$.5/\sqrt{n}$$

anbietet. Für mittlere und große Werte von n nimmt man meist als Näherung

$$\sigma/n \sim \sqrt{\bar{x}(1-\bar{x})/n} \ .$$

Damit können Konfidenzintervalle wie früher gefunden werden. Betrachten wir zum Beispiel 500 Personen, die befragt wurden und vor denen 200 angaben, für eine bestimmte Sache zu sein. Ein ungefähres 95%-Konfidenzintervall für den wahren Anteil in der Bevölkerung errechnet sich als

$$\frac{200}{500} - 1.96\sqrt{\frac{.4(1-.4)}{500}} < p < \frac{200}{500} + 1.96\sqrt{\frac{.4(1-.4)}{500}}$$

oder

$$(.357, .443) \ .$$

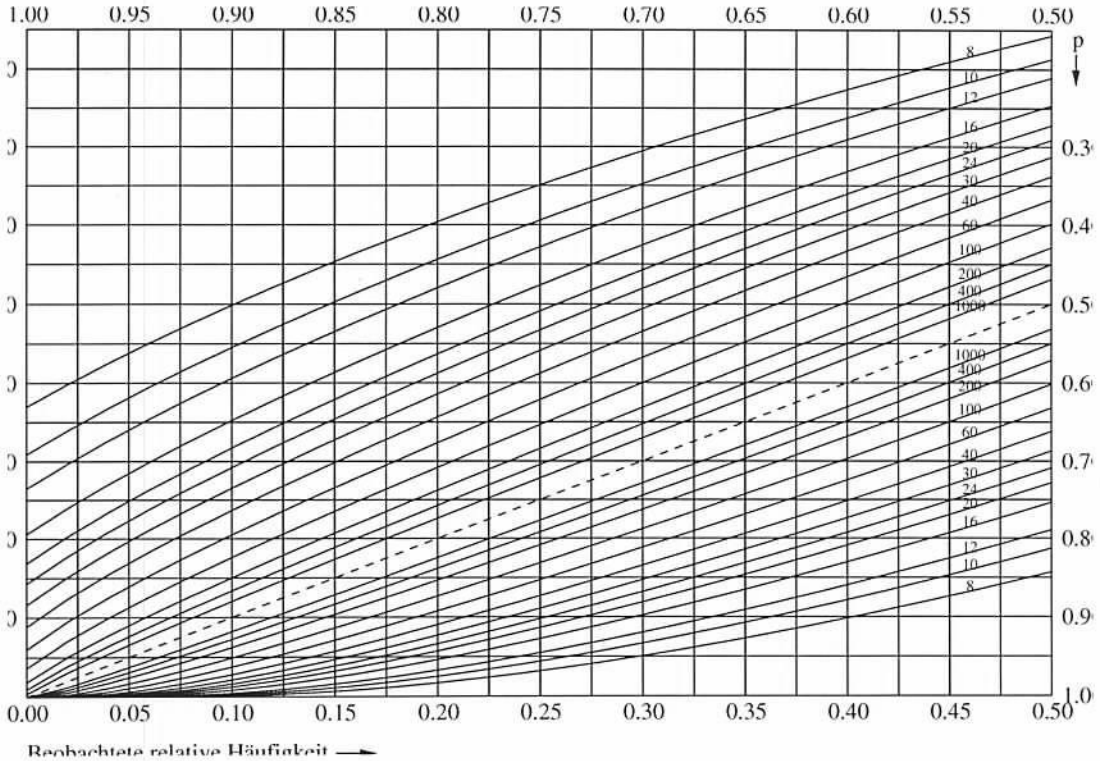


Abbildung 5.4: Kurvenblatt der Quantile der Binomialverteilung zur Bestimmung von 95%-Konfidenzintervallen für den Anteil p.

Ein Hypothesentest bezüglich des wahren Anteils sieht algorithmisch folgendermaßen aus:

1. Beschreibung der Zielvorstellung.
2. Formulierung der Hypothese $H_0 : p = p_0$ mit der Alternative $H_1 : p \neq p_0$ oder $p > (<) p_0$.
3. Wahl von α.

4. Teststatistik
$$Z = \frac{\bar{X} - p_0}{\sqrt{p_0(1-p_0)/n}} \ .$$
5. Annahme der approximativen Normalverteilung für Z mit Mittelwert 0 und Varianz 1.
6. Bestimmung von $z_{1-\alpha/2}$ für den 2-seitigen Test oder von $z_{1-\alpha}$ für den 1-seitigen.
7. Berechnung der Werte von $\bar{X}$ und Z.
8. Statistische Schlussfolgerung.
9. Experimentelle Schlussfolgerung.

Beispiel 5.11: (1) Es soll auf dem 5%-Niveau getestet werden, ob in einer bestimmten Insektenart 50% männliche und 50% weibliche Individuen vorhanden sind. Dazu wird bei 100 Insekten das Geschlecht überprüft.
(2) $H_0 : p = .5$, Alternative $p \neq .5$.
(3) $\alpha = .05$.
(4) $Z = (\bar{X} - .5)/\sqrt{.5(1-.5)/100}$.
(5) Annahme der Normalverteilung für Z.
(6) Verwerfung von H_0, wenn z ≤ -1.96 oder $z \geq 1.96$.
(7) Die Stichprobe der 100 Insekten ergibt 40 männliche und 60 weibliche Tiere. Daher ist $\bar{x} = .4$ und $z = (.4 - .5)/\sqrt{.5(1-.5)/100} = -2$.
(8) Aus $-2 < -1.96$ wird geschlossen, dass die Hypothese $H_0 : p = .5$ auf dem 5%-Niveau nicht gehalten werden kann.
(9) Ein Ungleichgewicht zwischen den Geschlechtern dieser Insektenart wird auf Grund dieser Stichprobe gefolgert. (Der geschätzte Anteil ist vom postulierten signifikant verschieden!)

5.6 Standardabweichung und Varianz

Die Standardabweichung dient als Maß für die Variabilität einer Messgröße, das ebenso wichtig ist wie das Mittel. Bei der Erzeugung von Kettengliedern ist nicht nur die mittlere Belastbarkeit der Glieder von Interesse, sondern es wird auch eine geringe Variabilität gefordert.

5.6.1 Konfidenzintervall

Als Basis zur Schätzung eines Konfidenzintervalls für die Varianz σ^2 dient die Verteilung des Schätzers
$$S^2 = \frac{1}{n-1} \sum (X_i - \bar{X})^2 \ .$$

Wir nehmen wieder an, dass die Stichprobe aus einer normalverteilten Größe $N(\mu, \sigma^2)$ erzeugt wurde. Die modifizierte Zufallsvariable

$$Y = \frac{1}{\sigma^2}\sum_{i=1}^{n}(X_i - \bar{X})^2 = (n-1)S^2/\sigma^2$$

besitzt dann eine χ^2_{n-1}-Verteilung mit $n-1$ Freiheitsgraden. Diese χ^2-Verteilung ist nicht symmetrisch, sodass wir 2 Quantile auswählen müssen, etwa $\chi^2_{n-1;\frac{\alpha}{2}}$ und $\chi^2_{n-1;1-\frac{\alpha}{2}}$ für festes α. Dann gilt

$$P(\chi^2_{n-1;\frac{\alpha}{2}} \leq (n-1)S^2/\sigma^2 \leq \chi^2_{n-1;1-\frac{\alpha}{2}}) = 1-\alpha \quad .$$

Die Ungleichungen werden wie für das Konfidenzintervall für μ umgeformt, sodass die Formel

$$P((n-1)S^2/\chi^2_{n-1;1-\frac{\alpha}{2}} \leq \sigma^2 \leq (n-1)S^2/\chi^2_{n-1;\frac{\alpha}{2}}) = 1-\alpha$$

ein Konfidenzintervall

$$((n-1)S^2/\chi^2_{n-1;1-\frac{\alpha}{2}}, (n-1)S^2/\chi^2_{n-1;\frac{\alpha}{2}})$$

für σ^2 liefert. Man bemerkt, dass μ nicht bekannt sein muss.

Beispiel 5.12: Es seien die Daten wie im Beispiel auf Seite 78 gegeben ($n = 16$, $s = 2.80$) und ein 95%-Konfidenzintervall wird gesucht ($\alpha = .05$). Aus Tabelle A des Anhangs erhalten wir die Quantile

$$\chi^2_{15;.975} = 27.49, \quad \chi^2_{15;.025} = 6.26 \quad ,$$

woraus das Konfidenzintervall sofort folgt:

$$(15 * 2.80^2/27.49, 15 * 2.80^2/6.26) = (4.28, 18.8) \quad .$$

Konfidenzintervalle für σ können aus dem obigen sofort durch Wurzelziehen der Grenzen gefunden werden.

5.6.2 Hypothesentest

Betrachten wir nun einen Test bezüglich der Varianz, der meist mit einseitiger Alternative $\sigma^2 > \sigma_o^2$ konstruiert wird. Die Null-Hypothese heißt natürlich

$$H_0 : \sigma^2 = \sigma_0^2 \quad .$$

Als Teststatistik verwenden wir eine Funktion der Stichprobe, die σ^2 verwendet, und deren Verteilung wir kennen. Die Variable

$$Y = \frac{1}{\sigma^2}\sum_{i=1}^{n}(X_i - \bar{X})^2 = (n-1)\frac{S^2}{\sigma^2}$$

besitzt eine ${\chi_{n-1}}^2$-Verteilung, wenn die $X_i \sim N(\mu, \sigma^2)$ verteilt sind. Es ist naheliegend, die Hypothese zu verwerfen, wenn der Wert von Y zu groß ausfällt. Wir definieren daher den kritischen Bereich von Y mit

$$Y > \chi^2_{n-1;1-\alpha} \quad ,$$

wobei

$$P(Y > \chi^2_{n-1;1-\alpha}) = \alpha \quad .$$

Beispiel 5.13: Ähnlich wie im vorigen Beispiel seien normalverteilte Daten ($n = 16$) gegeben. Allerdings wird $s^2 = \frac{1}{n-1}\sum(x_i - \bar{x})^2$ mit $\bar{x} = 15.70$ als $s^2 = 11.29$ berechnet. Es soll getestet werden, ob $\sigma^2 = \sigma_o^2 = 7.84$ mit der Alternativ-Hypothese $\sigma^2 > \sigma_o^2$. Wir wählen als Signifikanzniveau $\alpha = .05$. Dann finden wir aus Tabelle A des Anhangs die Grenze des kritischen Bereichs $\chi^2_{n-1;1-\alpha} = 24.996$. Der Wert unserer Teststatistik

$$y = (n-1)\frac{s^2}{\sigma_0^2} = 15\frac{11.29}{7.84} = 21.6$$

ist aber kleiner als 24.996, also nicht signifikant. Die Hypothese $\sigma^2 = 7.84$ kann also nicht verworfen werden.

5.7 Zwei Populationen

In vielen Forschungsstudien liegt das Hauptinteresse im Vergleich zweier Gruppen statt im Vergleich einer Gruppe mit irgendwelchen bekannten Werten. Zum Beispiel können zwei Lehrmethoden verglichen werden, die Wirkungen zweier Medikamente, Produktionsmethoden, etc.

5.7.1 Vergleich der Mittel

Beim Vergleich zweier Gruppen von Beobachtungen betrachtet man i.a. ihre Mittelwerte und untersucht sie auf signifikante Unterschiede. Dabei nehmen wir an, dass beide Populationen, von denen die Beobachtungen kommen, normalverteilt sind und gleiche Varianzen aufweisen (ein Test auf gleiche Varianzen wird im nächsten Unterabschnitt besprochen). Bezüglich der beiden Stichproben können 2 wesentliche Fälle auftreten:

(i) Jeder Wert der einen Stichprobe hängt mit genau einem der anderen zusammen (z.B. Paare von Messungen an verschiedenen Objekten). Dann bildet man am besten Differenzen zwischen den zusammengehörigen Werten und testet auf Mittelwert gleich 0.

(ii) Die beiden Stichproben sind voneinander unabhängig (und nicht notwendigerweise gleich groß). Bezeichnen wir die beiden Stichproben mit $X_1, \ldots, X_{n_1}$

und $Y_1, \ldots, Y_{n_2}$ mit Mittel μ_X bzw. μ_Y und die entsprechenden Schätzer für Mittel und Varianzen mit $\bar{X}$, $\bar{Y}$, S_X^2 und S_Y^2. Dann ist (ohne Beweis) die Größe

$$T = \sqrt{\frac{n_1 n_2 (n_1 + n_2 - 2)}{n_1 + n_2}} \cdot \frac{\bar{X} - \bar{Y}}{\sqrt{(n_1 - 1)S_X^2 + (n_2 - 1)S_Y^2}}$$

t-verteilt mit n_1+n_2-2 Freiheitsgraden. Für den Test $\mu_X = \mu_Y$ wird also T als Teststatistik verwendet und der kritische Bereich ergibt sich beim einseitigen Test (Alternative $\mu_X > \mu_Y$) als

$$T > t_{n_1+n_2-2;1-\alpha} \quad .$$

(Man bezeichnet diesen Test auch als *2-Stichproben-t-Test.*[3])

Beispiel 5.14: 2 Farbstoffe sollen getestet werden. Der 1. ist billiger, sodass der 2. nur interessant erscheint, wenn er signifikant besser dem Wetter standhält. Es werden 5 unabhängige Versuche durchgeführt, die zu folgenden Kennzahlen führen:

Farbstoff I	85	87	92	80	84
Farbstoff II	89	89	90	84	88

Bezeichnen wir die Werte für den Farbstoff I mit X und für den anderen mit Y. Dann heißt die Null-Hypothese $\mu_X = \mu_Y$ (oder besser $\mu_X \geq \mu_Y$) und die Alternative $\mu_X < \mu_Y$. Wir wählen $\alpha = .05$. Die Werte der einzelnen Statistiken berechnen sich als

$$\bar{x} = 85.6, \quad \bar{y} = 88.0, \quad s_X^2 = 19.3, \quad s_Y^2 = 5.5$$

$$t = \sqrt{\frac{5 * 5 * 8}{10}} \frac{85.6 - 88.0}{\sqrt{4 * 19.3 + 4 * 5.5}} = -1.08 \quad .$$

Der kritische Bereich ist hier allerdings durch

$$T < -t_{n_1+n_2-2;1-\alpha} = t_{n_1+n_2-2;\alpha}$$

gegeben. Aus der Tabelle A (Anhang) finden wir

$$t_{n_1+n_2-2;\alpha} = t_{8;.05} = -t_{8;.95} = -1.860 \quad .$$

t fällt also nicht in den kritischen Bereich. Daraus folgt, dass auf Grund dieser Versuchsserie der 2. Farbstoff nicht signifikant besser ist.

[3] R: `t.test(Daten1,Daten2)`

5.7.2 Vergleich der Varianzen

Oft ist es interessant zu wissen, ob die Varianzen zweier Normalverteilungen, deren Mittel nicht bekannt sein müssen, als gleich angesehen werden können. Dazu brauchen wir folgendes theoretische Hilfsmittel über das Verhältnis zweier Quadratsummen, das auch in der Varianzanalyse Verwendung findet.

Es seien V_1 und V_2 zwei unabhängige Zufallsvariable, die χ^2-Verteilungen mit m bzw. n Freiheitsgraden besitzen. Dann besitzt die Zufallsvariable

$$V = \frac{V_1/m}{V_2/n}$$

eine F-Verteilung mit m und n Freiheitsgraden ($F_{m,n}$-Verteilung). Quantile dieser Verteilung sind in der Tabelle A des Anhanges aufgeführt.

Wenn wir nun die empirische Varianz

$$S_X^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$$

einer $\mathrm{N}(\mu, \sigma^2)$-verteilten Zufallsvariablen X betrachten, so wissen wir, dass $(n-1)S_X^2/\sigma^2$ eine χ^2_{n-1}-Verteilung besitzt. Das Verhältnis der empirischen Varianzen zweier unabhängiger Normalverteilungen mit vorausgesetzter gleicher Varianz ist im wesentlichen F-verteilt. Genauer, wenn von X mit der Verteilung $\mathrm{N}(\mu_X, \sigma^2)$ n_1 Stichprobenwerte und von Y mit der von X unabhängigen Verteilung $\mathrm{N}(\mu_Y, \sigma^2)$ n_2 Stichprobenwerte zur Verfügung stehen, so ist

$$F = \frac{S_X^2/\sigma^2}{S_Y^2/\sigma^2} = \frac{S_X^2}{S_Y^2}$$

F_{n_1-1,n_2-1} verteilt. Diese Variable F können wir direkt als Teststatistik für den Vergleich der Varianzen zweier normalverteilter, unabhängiger Zufallsvariablen verwenden. Betrachten wir z.B. die Hypothese $\sigma_X^2 = \sigma_Y^2$ gegen die Alternative $\sigma_X^2 > \sigma_Y^2$ für Normalverteilungen unter Benutzung unabhängiger Stichproben und stellen dafür einen statistischen Test auf. Ist die Alternativ-Hypothese $\sigma_X^2 > \sigma_Y^2$ wahr, dann wird die Teststatistik F wahrscheinlich große Werte annehmen, sodass wir den kritischen Bereich mit

$$F > F_{n_1-1,n_2-1;1-\alpha}$$

festlegen, wobei die Grenze wieder das 1-α-Quantil der F-Verteilung darstellt, d.h.

$$P(F \leq F_{n_1-1,n_2-1;1-\alpha}) = 1 - \alpha \ .$$

Wir werden also die empirischen Varianzen s_X^2 und s_Y^2 berechnen und die Null-Hypothese verwerfen, wenn für gewähltes α der Wert von $F = s_X^2/s_Y^2$ die Schranke $F_{n_1-1,n_2-1;1-\alpha}$ übersteigt[4].

[4] R: `var.test(Daten1,Daten2)`

Beispiel 5.15: Bei je 16 Schrauben mit gewalztem bzw. gefrästem Gewinde wurde der Flankendurchmesser bestimmt. Es ergaben sich folgende Werte: $s_X^2 = .001382\ mm^2, s_Y^2 = .000433\ mm^2$. Unter Annahme der Normalverteilung sollte die Hypothese $\sigma_X^2 = \sigma_Y^2$ gegenüber $\sigma_X^2 > \sigma_Y^2$ getestet werden.
Wir wählen $\alpha = .05$. Die Freiheitsgrade der F-Verteilung betragen $n_1 - 1 = 15$ bzw. $n_2 - 1 = 15$, sodass wir aus Tabelle A des Anhangs den entsprechenden F-Wert $F_{15,15;.95} = 2.40$ finden. Der Wert der Teststatistik F errechnet sich als $F = .001382/.000433 = 3.19$, ist also größer als 2.40. Die Hypothese wird verworfen: Die Varianz s_X^2 ist signifikant größer als s_Y^2.

5.8 Anpassungstests

5.8.1 Chi-Quadrat-Test

Die bisher besprochenen Tests beziehen sich immer nur auf einzelne Parameter von Verteilungen. Die Form der Verteilung (z.B. normal) wird dabei als bekannt vorausgesetzt. Manchmal ist aber interessant, eine Hypothese über die Form der Verteilung zu prüfen. Der wohl am weitesten verbreitete Test ist der *Chi-Quadrat-Test*.

Es wird wie bei einem Histogramm eine Klasseneinteilung getroffen, und die (empirischen) Häufigkeiten werden mit den theoretischen (hypothetischen) verglichen. Weichen sie zu stark voneinander ab, wird man die Hypothese verwerfen, sonst annehmen. Nehmen wir an, es seien k Klassen (Intervalle) gewählt worden. Dann bezeichnen wir mit h_i die absolute Häufigkeit, also die Anzahl von Datenpunkten in der i-ten Klasse. Die aus der Hypothese über die Verteilung entsprechende, theoretische Wahrscheinlichkeit, dass ein Wert in die i-te Klasse fällt, bezeichnen wir mit p_i und weiters die unter der Hypothese erwartete absolute Häufigkeit $e_i = np_i$, wobei wie üblich n die Anzahl der Daten angibt. Die Teststatistik T verwendet die quadratischen Abweichungen, genauer

$$T = \sum_{i=1}^{k} \frac{(h_i - e_i)^2}{e_i} \quad .$$

Die ungefähre Verteilung von T ergibt sich aus dem folgenden theoretischen Hilfsmittel: Wenn die Hypothese über die Wahrscheinlichkeitsverteilung zutrifft, strebt die Verteilung von T gegen eine χ_{k-1}^2-Verteilung.

Da die Hypothese verworfen wird, wenn die Abweichungen und damit der Wert von T zu groß ausfällt, wird man für eine gegebene Signifikanzzahl α den kritischen Bereich mit

$$T > \chi_{k-1;1-\alpha}^2$$

definieren. Die Verteilung von T ist nur „asymptotisch" bekannt, stimmt also annähernd nur für große n. Man sollte daher die Faustregel beachten: Großes n und die Klasseneinteilung so wählen, dass in jede Klasse mindestens 5 Punkte fallen.

Wenn die hypothetische Verteilung noch unbekannte Parameter enthält, die mit den gleichen Daten geschätzt werden müssen, so wirkt sich das natürlich in der Verteilung von T aus: Werden r Parameter geschätzt, so besitzt T die asymptotische Verteilung χ^2_{k-r-1}. Die Freiheitsgrade werden also um r verringert.

Beispiel 5.16: Betrachten wir nun nochmals die Rechnung im Beispiel auf Seite 48. 250 Stichprobenwerte über Asche von Kohle waren gegeben. Die Hypothese der Normalverteilung der Daten soll geprüft werden. Der Wert der Teststatistik

$$T = \sum_{j=1}^{10} d_j^2 = 3.74$$

entspricht einer Realisation einer Zufallsvariablen, die χ^2 mit 7 Freiheitsgraden verteilt ist. Diese Freiheitsgrade errechnen sich aus der Klassenzahl (10) minus 2 geschätzten Parameter für μ und σ^2 minus 1 (wegen $\sum e_i = \sum h_i$). Der Vergleich von $T = 3.74$ mit Werten aus Tabelle A des Anhangs gibt keine Indikation gegen die Annahme der Normalverteilung.

5.8.2 Kolmogorov-Smirnov-Test

Hier muss im Prinzip angenommen werden, dass die Stichprobenvariablen $X_1, \ldots, X_n$ eine stetige Verteilungsfunktion F haben. Um die zugrundeliegende Verteilung F auf eine hypothetische F_0 zu testen, d.h. $H_0 : F(x) = F_0(x)\ \forall x$ ist es naheliegend, die absolute Differenz

$$\mid F_n(x) - F_0(x) \mid$$

bezüglich der empirischen Verteilungsfunktion F_n zu betrachten.

Beispiel 5.17: Es sei zu testen, ob für einen bestimmten PKW-Typ der Benzinverbrauch in Litern pro 100 km bei einer Geschwindigkeit von 100 km/h normalverteilt mit $\mu = 12$ und $\sigma = 1$ ist. Die Stichprobenwerte von 10 Fahrzeugen dieses Typs ergaben

12.4 11.8 12.9 12.6 13.0 12.5 12.0 11.5 13.2 12.8

die hypothetische Verteilung N(12,1) und die empirische sind in der Abbildung 5.5 dargestellt.

Die Teststatistik (Kolmogorov-Smirnov) lautet nun[5]:

$$K_n = \sup_x \mid F_0(x) - F_n(x) \mid \ ,$$

$$K_n^+ = \sup_x (F_0(x) - F_n(x))$$

oder

$$K_n^- = \sup_x (F_n(x) - F_0(x))$$

[5] R: `ks.test(Daten,'pnorm')`

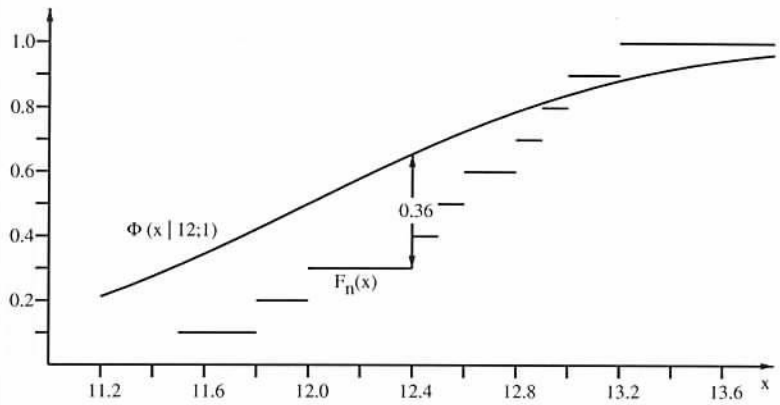


Abbildung 5.5: Hypothetische und empirische Verteilungsfunktion.

entsprechend den drei Gegenhypothesen

$$H_1: \quad F(x) \neq F_0(x)$$
$$H_1: \quad F(x) < F_0(x)$$
$$H_1: \quad F(x) > F_0(x) \text{ für mindestens ein } x.$$

Man kann leicht feststellen, dass die Verteilung der Teststatistiken nicht von der tatsächlichen Verteilung F abhängt, also *verteilungsfrei* ist (wie die χ^2-Statistik im vorigen Unterabschnitt). Quantile der Verteilung sind tabelliert (Tabelle A.6 des Anhangs). Man bemerke den Zusammenhang zwischen der grafischen Analyse mit Wahrscheinlichkeitspapier und diesem formalen Test.

Interessanterweise kann im obigen Beispiel die Nullhypothese auf dem 5%-Niveau nicht abgelehnt werden (kleine Stichprobe!).

Kapitel 6

Varianzanalyse

Die monatliche Gewichtszunahme einer Anzahl von Tieren variiert von Tier zu Tier, auch wenn alle anderen Bedingungen wie Futterart und -menge gleich sind. Die Variation erscheint rein zufälliger Art. Werden die Tiere allerdings unterschiedlich gefüttert, so kommt noch die Variation bezüglich des Futters dazu. Zur grundsätzlichen Fragestellung, ob die Futterart auf die Gewichtszunahme einen Einfluss hat, d.h. ob die durchschnittlichen Gewichtszunahmen gleich sind, müssen wir versuchen, die beiden Variationen zu trennen. Dies wird typisch in der *Varianzanalyse* durchgeführt, und wir sprechen von der Varianzzerlegung. Dieses Beispiel zählt zur *einfachen Varianzanalyse.*

Wenn wir zwei Einflüsse gleichzeitig untersuchen wollen, wie z.B. Futtermenge und Futterart, müssen diese voneinander und von der zufälligen Variation getrennt werden. Man spricht von der *doppelten Varianzanalyse.*

Die Varianzanalyse verwendet als Werkzeug im wesentlichen die Zerlegung der *Quadratsumme*, d.h. der Summe der Quadrate der Abweichungen der Stichprobenwerte vom Mittelwert.

6.1 Vergleich der Mittelwerte mehrerer Normalverteilungen

Angenommen, n unabhängige Stichprobenwerte

$$x_{ij},\ i = 1, \ldots, n_j,\ j = 1, \ldots, k\ ,$$

von normalverteilten Zufallsvariablen mit *gleicher* Varianz seien gegeben, d.h. für die Zufallsvariablen soll gelten

$$X_{ij} \sim N(\mu_j, \sigma^2)$$

mit $j = 1, \ldots, k$ und

$$n = \sum_{j=1}^{k} n_j\ .$$

Wir möchten auf Gleichheit aller Mittelwerte

$$H_o : \mu_1 = \mu_2 = \ldots = \mu_k$$

gegen

$$H_1 : \mu_r \neq \mu_s \text{ für mindestens ein } r \neq s, \quad r, s = 1, \ldots, k \quad ,$$

testen.

Betrachten wir die Quadratsumme der Abweichungen

$$q = \sum_{j=1}^{k} \sum_{i=1}^{n_j} (x_{ij} - \bar{x})^2$$

und erweitern die Abweichungen

$$(x_{ij} - \bar{x}) = (x_{ij} - \bar{x}_j) + (\bar{x}_j - \bar{x}) \quad ,$$

wobei natürlich gilt

$$\bar{x}_j = \frac{1}{n_j} \sum_{i=1}^{n_j} x_{ij} \quad ,$$

quadrieren und summieren, so erhalten wir

$$q = \sum_{j=1}^{k} \sum_{i=1}^{n_j} (x_{ij} - \bar{x})^2 = \sum_{j=1}^{k} \sum_{i=1}^{n_j} (x_{ij} - \bar{x}_j)^2 + \sum_{j=1}^{k} \sum_{i=1}^{n_j} (\bar{x}_j - \bar{x})^2 \quad ,$$

weil das gemischte Glied

$$\sum_{j=1}^{k} \sum_{i=1}^{n_j} (x_{ij} - \bar{x}_j)(\bar{x}_j - \bar{x}) = \sum_{j=1}^{k} (\bar{x}_j - \bar{x}) \sum_{i=1}^{n_j} (x_{ij} - \bar{x}_j) = 0$$

verschwindet. Wir haben also q in 2 Quadratsummen zerlegt, nämlich

$$q = q_I + q_Z \quad ,$$

wobei

$$q_I = \sum_{j=1}^{k} \sum_{i=1}^{n_j} (x_{ij} - \bar{x}_j)^2$$

die Summe der quadratischen Abweichungen *innerhalb jeder Stichprobe* und

$$q_Z = \sum_{j=1}^{k} \sum_{i=1}^{n_j} (\bar{x}_j - \bar{x})^2 = \sum_{j=1}^{k} n_j (\bar{x}_j - \bar{x})^2$$

die Quadratsumme zwischen den Stichproben darstellt.

Nun kann man zeigen, dass unter der Null-Hypothese, d.h. alle x_{ij} sind Realisationen der Verteilung $N(\mu_o, \sigma^2)$ mit $\mu_1 = \mu_2 = \ldots = \mu_k = \mu_o$, die Verteilungen

der entsprechenden Zufallsvariablen von q_I/σ^2 bzw. q_Z/σ^2 unabhängig und gleich χ^2_{n-k} bzw. χ^2_{k-1} sind. Daraus folgt, dass das Verhältnis

$$F = \frac{q_Z/(k-1)}{q_I/(n-k)}$$

einer $F_{k-1,n-k}$-Verteilung genügt. Diese Variable F wird als Teststatistik verwendet. Fällt ihr Wert zu groß aus, muss die Null-Hypothese verworfen werden. Für eine bestimmte Signifikanzzahl α ergibt sich der kritische Bereich als

$$F > F_{k-1,n-k;1-\alpha} \quad .$$

Bei der praktischen Berechnung des Tests werden die Zahlenwerte gerne in einer *Varianzanalyse-Tafel*[1] angeordnet:

Variation	Freiheitsgrade FG	Quadratsumme q	mittlere Quadrats. s^2	F
Zwischen den Gr.	$k-1$	q_z	$s_Z^2 = q_Z/(k-1)$	$\frac{q_Z/(k-1)}{q_I/(n-k)}$
Innerhalb der Gr.	$n-k$	q_I	$s_I^2 = q_I/(n-k)$	
Total	$n-1$	q		

Beispiel 6.1: Es wurden Staubuntersuchungen in Abgasen durchgeführt. Die nachfolgenden Werte dienten dazu, festzustellen, ob die Geschwindigkeiten (in Fuß/sec.) zeitlich einigermaßen konstant sind. Zwischen den Messwerten jeder Gruppe liegt ein größerer Zeitabstand.

Gruppe	Geschwindigkeit (Fuß/sec)									
A	20	21	20	20	23	21	26			
B	24	25	27	23	22	22	24	27	26	25
C	25	28	22	24	26	26				

Die Null-Hypothese heißt also: Die Mittelwerte der Gruppen sind gleich. Als Signifikanzzahl wählen wir $\alpha = .05$. Für die Mittelwerte der Gruppen erhalten wir

$$\bar{x}_1 = 21.57, \quad \bar{x}_2 = 24.5, \quad \bar{x}_3 = 25.17 \quad .$$

Die Abweichungsquadrate innerhalb der Gruppen ergeben

$$q_I = \sum_{j=1}^{3} \sum_{i=1}^{n_j} (x_{ij} - \bar{x}_j)^2 = 81.05$$

mit $n - k = \sum n_j - 3 = 20$ Freiheitsgraden. Die Abweichungen zwischen den Gruppen ergeben mit $\bar{x} = 23.78$

$$q_Z = \sum_{j=1}^{3} n_j (\bar{x}_j - \bar{x})^2 = 50.87$$

[1] R: `summary(aov(Y ~ FakA, data = Daten))`

mit $k-1=2$ Freiheitsgraden, sodass wir für den F-Wert

$$F = 6.27 \ ,$$

erhalten, und die Varianzanalyse-Tafel bekommt folgende Gestalt.

Variation	FG	q	s^2	F
Zwischen den Gruppen	2	$q_Z = 50.87$	$s_Z^2 = 25.44$	6.27
Innerhalb der Gruppen	20	$q_I = 81.05$	$s_I^2 = 4.05$	
Total	22	$q = 131.92$		

Der F-Wert ist größer als das maßgebende Quantil der F-Verteilung

$$F_{2,20;.95} = 3.49 \ .$$

Wir schließen daraus, dass die Werte zwischen den Gruppen signifikant verschieden sind, d.h. die durchschnittlichen Gasgeschwindigkeiten können nicht als gleich angesehen werden.

6.2 Doppelte Varianzanalyse *****

Wenn man Daten nach 2 Gesichtspunkten einteilt und sie nach diesen analysieren will, kann man die doppelte Varianzanalyse verwenden. Die beiden Gesichtspunkte könnten z.B. Düngemittel und Bewässerung beim Ernteertrag sein.

Gegeben sei eine Stichprobe von n Werten, die sich in k Gruppen und jede Gruppe in genau p Klassen einteilen lässt. Wir bezeichnen die Daten x_{ij} mit 2 Indizes, sodass wir sie in Matrixform anordnen können:

$$\begin{array}{c} \\ k \text{ Zeilen} \\ (\text{Gruppen}) \end{array} \overbrace{\begin{array}{c} \\ \left\{ \begin{array}{cccc} x_{11} & x_{12} & \dots & x_{1p} \\ x_{21} & x_{22} & \dots & x_{xp} \\ \vdots & & \vdots & \\ x_{k1} & x_{k2} & \dots & x_{kp} \end{array} \right. \end{array}}^{p \text{ Spalten (Klassen)}}$$

Wir nehmen an, dass x_{ij} unabhängige Realisationen von normalverteilten Zufallsvariablen mit *gleicher* Varianz sind. Es interessiert uns, ob die theoretischen Mittelwerte der Spalten gleich sind, d.h.

$$H_o : \bar{\mu}_{.1} = \bar{\mu}_{.2} = \ldots = \bar{\mu}_{.p} \ ,$$

und/oder ob die Mittelwerte der Zeilen gleich sind, d.h.

$$H_o : \bar{\mu}_{1.} = \bar{\mu}_{2.} = \ldots = \bar{\mu}_{k.} \ ,$$

wobei der durch einen Punkt ersetzte Index Mittelbildung bedeutet. Gegenhypothese ist immer „mindestens“ ein $\neq$.

Wir bezeichnen den Mittelwert in der i-ten Gruppe (Zeile) mit

$$\bar{x}_{i.} = \frac{1}{p}\sum_{j=1}^{p} x_{ij}$$

und den Mittelwert der j-ten Klasse mit

$$\bar{x}_{.j} = \frac{1}{k}\sum_{i=1}^{k} x_{ij} \quad ,$$

sowie den Mittelwert der gesamten Stichprobe mit

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{k}\sum_{j=1}^{p} x_{ij}$$

mit $n = kp$. Die „totale“ Quadratsumme

$$q = \sum_{i=1}^{k}\sum_{j=1}^{p}(x_{ij} - \bar{x})^2$$

spalten wir jetzt in 3 Teile auf. Zuerst trennen wir die Abweichungen in

$$x_{ij} - \bar{x} = (\bar{x}_{i.} - \bar{x}) + (\bar{x}_{.j} - \bar{x}) + (x_{ij} - \bar{x}_{i.} - \bar{x}_{.j} + \bar{x}) \quad ,$$

quadrieren und summieren, sodass wir erhalten

$$q = p\sum_{i=1}^{k}(\bar{x}_{i.} - \bar{x})^2 + k\sum_{j=1}^{p}(\bar{x}_{.j} - \bar{x})^2 + \sum_{i}\sum_{j}(x_{ij} - \bar{x}_{i.} - \bar{x}_{.j} + \bar{x})^2 \quad ,$$

weil die gemischten Glieder wieder wegfallen. q haben wir also aufgespalten in

$$q_Z = p\sum_{i=1}^{k}(\bar{x}_{i.} - \bar{x})^2 \quad ,$$

der Quadratsumme der Mittelwerte zwischen den Zeilen, in

$$q_S = k\sum_{j=1}^{p}(\bar{x}_{.j} - \bar{x})^2 \quad ,$$

der Quadratsumme der Mittelwerte zwischen den Spalten, und in die quadratische Restsumme

$$q_R = \sum_{i}\sum_{j}(x_{ij} - \bar{x}_{i.} - \bar{x}_{.j} + \bar{x})^2 \quad .$$

Ist die Null-Hypothese richtig, so kann man zeigen, dass die den Quadratsummen q_Z, q_S, q_R entsprechenden Zufallsvariablen Q_Z, Q_S, Q_R voneinander unabhängig sind und Q_Z/σ^2, Q_S/σ^2 und Q_R/σ^2 χ^2-verteilt sind mit $k-1, p-1$, bzw. $(k-1)(p-1)$ Freiheitsgraden. Daraus ergibt sich, dass die Verhältnisse der mittleren Quadratsummen

$$S_Z^2/S_R^2 \text{ und } S_S^2/S_R^2$$

mit $S_Z^2 = Q_Z/(k-1)$, $S_S^2 = Q_S/(p-1)$ und $S_R^2 = Q_R/[(k-1)(p-1)]$ F-Verteilungen mit $[(k-1),(k-1)(p-1)]$ bzw. $[(p-1),(k-1)(p-1)]$ Freiheitsgraden besitzen. Diese Verhältnisse verwenden wir als entsprechende Teststatistiken, um wie in der einfachen Varianzanalyse zu prüfen, ob es zwischen den Zeilen bzw. Spalten signifikante Unterschiede gibt. Fällt ein Verhältnis zu groß aus, etwa

$$F = S_Z^2/S_R^2 > F_{k-1,(k-1)(p-1);1-\alpha} \quad ,$$

dann werden wir die Hypothese der Gleichheit der Zeilenmittel verwerfen. Analog prüft man mit S_S^2/S_R^2 auf einen signifikanten Unterschied zwischen den Spalten. Die Werte der Statistiken trägt man häufig in eine Varianzanalyse-Tafel[2] ein:

Variation	Freiheitsgrade FG	Quadratsumme q	mittlere Quadratsumme s^2	F
Zwischen den Zeilen	$k-1$	q_Z	$s_Z^2 = q_Z/(k-1)$	s_Z^2/s_R^2
Zwischen den Spalten	$p-1$	q_S	$s_S^2 = q_S/(p-1)$	s_S^2/s_R^2
Rest	$(k-1)(p-1)$	q_R	$s_R^2 = q_R/(k-1)(p-1)$	
Total	$n-1$	q		

Bemerkung: Nachdem $q = q_Z + q_S + q_R$ gilt, wird man die restliche Quadratsumme der Einfachheit halber immer als $q_R = q - q_Z - q_S$ berechnen.

Beispiel 6.2: Die folgenden Daten stammen aus einem Versuch über Befestigungsarten von Werkzeugen, um Unfälle und Zeitverluste bei Arbeiten an einem Raketensilo zu vermeiden. Angegeben ist die zum Entfernen eines durch 6 Schrauben gehaltenen Teils benötigte Zeit (sec). Der Schraubenzieher war an einem Riemen befestigt, dessen anderes Ende am Gürtel hing (Typ A) oder das Handgelenk fest (Typ B) bzw. lose (Typ C) umschloss. Weiterhin interessierten die Unterschiede des Arbeitens der nach Alter und Ausbildung differierenden Teilnehmer. Die Hypothesen, dass (1) zwischen den 4 Befestigungsarten und (2) zwischen den 12 Teilnehmern kein Unterschied besteht, sollen überprüft werden.

[2] R: `summary(aov(Y ~ FakA + FakB, data = Daten))`

	Arbeiter Nr.											
Halter	1	2	3	4	5	6	7	8	9	10	11	12
Typ A	93	98	91	65	74	80	84	81	55	94	49	64
Typ B	97	62	100	70	76	68	73	73	61	85	61	61
Typ C	133	64	71	76	66	76	74	94	64	82	49	67
ohne	108	62	62	62	68	78	74	67	53	71	47	63
$\bar{x}_{.j}$	107.8	71.5	81	68.3	71	75.5	76.3	78.8	58.3	83	51.5	63.8
$\sum_i (x_{ij} - \bar{x}_{.j})^2$	970.8	939.0	922.0	112.8	68.0	83.0	80.8	408.8	78.8	270.0	123.0	18.8

Die berechneten Mittelwerte und Quadratsummen der Spalten und Zeilen sind an den Rändern der Tafel eingetragen.

	$\bar{x}_{i.}$	$\sum_j (x_{ij} - \bar{x}_{i.})^2$
Typ A	77.33	2844.7
Typ B	73.92	2034.9
Typ C	76.33	4834.7
ohne	67.92	2544.9
$\bar{x}$	73.883	

Die Varianzanalyse-Tafel bekommt nun folgende Gestalt:

Variation	FG	q	s^2	F
Zwischen den Zeilen	3	$q_Z = 642.08$	$s_Z^2 = 214.03$	2.06
Zwischen den Spalten	11	$q_S = 8825.75$	$s_S^2 = 802.34$	7.71
Rest	33	$q_R = 3433.42$	$s_R^2 = 104.04$	
Total	47	$q = 12901.25$		

Wählen wir $\alpha = .05$. Dann ist die Variation zwischen den Zeilen nicht signifikant groß ($F = 2.06 < F_{3,33;.95} = 2.92$), d.h., die Hypothese, dass die Befestigungsart keinen Einfluss hat, kann nicht verworfen werden; die Variation zwischen den Spalten ist allerdings signifikant ($7.71 > F_{11,33;.95} = 2.16$). Daraus schließen wir, dass die Null-Hypothese, dass die mittlere Arbeitszeit der Arbeiter gleich ist, verworfen werden muss, also der Einfluss der Person auf die Arbeitszeit ist *statistisch gesichert.*

Bemerkung: Die Zufallsvariablen X_{ij} im Modell der obigen, doppelten Varianzanalyse können auch als

$$X_{ij} = \mu + a_i + b_j + \epsilon_{ij}, \qquad i = 1, \ldots, k, \; j = 1, \ldots, p$$

dargestellt werden. Dabei bezeichnen $\epsilon_{ij} \sim N(0, \sigma^2)$-verteilte, unabhängige Zufallsvariable, und die a_i bzw. b_j werden *Zeilen-* bzw. *Spalteneinflüsse* (Einflüsse des Faktors A bzw. B) genannt, die hier der Einfachheit halber „additiv" wirken und deren Summen a_i und b_j gleich Null sind. Die Mittel der Zeilen bzw. Spalten werden dann als

$$\bar{\mu}_{i.} = \mu + a_i$$

bzw.

$$\bar{\mu}_{.j} = \mu + b_j$$

geschrieben, und die Nullhypothesen lauten

$$a_i = 0,\ i = 1, \ldots, k \quad ,$$

bzw.

$$b_j = 0,\ j = 1, \ldots, p \quad .$$

Kapitel 7

Regression und Korrelation

Ein *Regressionsproblem* behandelt die Verteilung einer Variablen, wenn mindestens eine andere gewisse Werte in nicht zufälliger Art annimmt. Ein *Korrelationsproblem* dagegen betrachtet die gemeinsame Verteilung von zwei Variablen, von denen keine durch den Experimentator fixiert wird, beide sind also zufällig. Typische *Regressionsprobleme* sind z.B. beim Studium des Ernteertrages mit verschiedenen Mengen von Dünger, bei der Lebensdauer von Tieren bei verschiedenen Strahlungsdosen etc., zu finden. Dabei werden immer die Werte einer Variablen festgehalten, und diese unterliegen keiner zufälligen Variation. Ein typisches *Korrelationsproblem* wäre das Studium des Zusammenhangs zwischen Intelligenzquotienten und Schulleistung von Kindern.

7.1 Das Regressionsproblem

Als einfaches Beispiel wollen wir den Zusammenhang der Verteilung des Gewichts von Männern mit ihrer Größe studieren. Dann wählen wir zu vorgegebenen Körpergrößen Männer zufällig aus und erhalten z.B. folgende Daten (siehe Abbildung 7.1).

Für jede gewählte Größe x bekommen wir eine gewisse Verteilung der Gewichte Y der Männer mit dieser Größe. Von dieser können eventuell Mittel $\mu_{y.x}$ und Varianz $\sigma^2_{y.x}$ angegeben werden. Weil die Verteilung von Y von den Werten von x abhängt, wird Y auch als *abhängige* und x als *unabhängige* Variable bezeichnet. Es muss aber festgehalten werden, dass x hier *keine* Zufallsvariable darstellt. Normalerweise wird die Varianz $\sigma^2_{y.x}$ als konstant über x angenommen.

In vielen Anwendungsbeispielen der Regressionsanalyse kann die Abhängigkeit der Mittelwerte von Y ($\mu_{y.x}$) von x im Bereich der x-Werte durch eine gerade Linie angegeben werden. Man spricht von einfacher, *linearer* Regression und schreibt z.B.

$$\mu_{y.x} = a + b(x - \bar{x}) \quad ,$$

wobei a und b feste Parameter darstellen.

x [cm]	$y = Y(\omega)$ [kg]
150	55
150	67.5
150	60
155	60
155	70
155	65
155	67.5
160	75
160	72.5
175	85
175	92.5
175	80

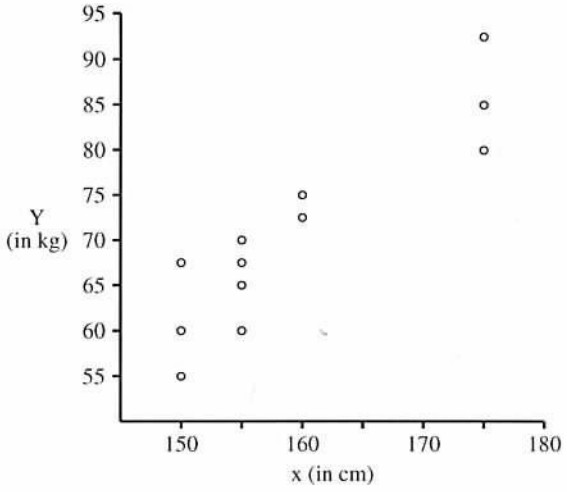


Abbildung 7.1: Körpergewichte über den Größen.

7.2 Schätzung der Parameter

Die Parameter der Regressionsgeraden müssen aus den Daten geschätzt werden. Dies geschieht zumeist mit der Methode der kleinsten Quadrate. Eine lineare, erwartungstreue Schätzung[1] für a ist dann das arithmetische Mittel der Y-Werte,

$$\hat{a} = \bar{y} \quad ,$$

und für b

$$\hat{b} = \frac{s_{xy}}{s_x^2} = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x_i - \bar{x})^2} \quad ,$$

wobei $s_x^2 = \frac{1}{n-1}\sum(x_i - \bar{x})^2$ die empirische Varianz der x-Werte und

$$s_{xy} = \frac{1}{n-1}\sum(x_i - \bar{x})(y_i - \bar{y})$$

die empirische Kovarianz (siehe später) zwischen x und Y bezeichnet. Sei $\hat{y}_x$ der geschätzte mittlere Wert von Y an der Stelle x (also von $\mu_{y.x}$). Dann gilt

$$\hat{\mu}_{y.x} = \hat{y}_x = \hat{a} + \hat{b}(x - \bar{x}) \quad .$$

Eine erwartungstreue Schätzung für $\sigma^2 = \sigma_{y.x}^2$ ist

$$s^2 = \frac{1}{n-2}\sum(y_i - \hat{y}_i)^2 = \frac{1}{n-2}\sum[y_i - \hat{a} - \hat{b}(x_i - \bar{x})]^2 \quad .$$

[1] R: `lm(y ~ x)` , aber Vorsicht: In R wird $\bar{x}$ nicht abgezogen!

s heißt auch mittlerer Fehler oder *Standardabweichung der Beobachtungen.*

Bei der Berechnung der Werte für das obige Beispiel der Körpergewichte von Männern ergibt sich folgendes:

$$\begin{array}{lllll}
\sum x_i & = & 1\,915 & \bar{x} & = & 159.58 \\
\sum y_i & = & 850 & \bar{y} & = & 70.83 \\
\sum x_i y_i & = & 136\,725 & & & \\
\sum x_i^2 & = & 306\,675 & & & \\
\sum y_i^2 & = & 61\,525 & & &
\end{array}$$

$$\begin{array}{lllll}
s_x^2 & = & \frac{1}{11}(306675 - 12 * 159.58^2) & = & 97.54 \\
s_y^2 & = & \frac{1}{11}(61525 - 12 * 70.83^2) & = & 119.70 \\
s_{xy} & = & \frac{1}{11}(136725 - 12 * 159.38 * 70.83) & = & 98.11 \\
\hat{b} & = & 98.11/97.54 & = & 1.01 \\
s^2 & = & \frac{11}{10}(119.70 - 1.01^2 * 97.54) & = & 23.12 \\
\hat{y}_x & = & 70.83 + 1.01(x - 159.58) & &
\end{array}$$

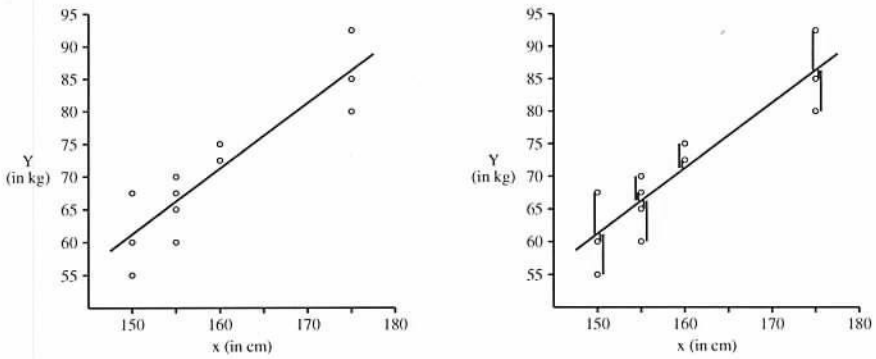


Abbildung 7.2: Regression der Körpergewichte über den Größen.

In der rechten Skizze der Abbildung 7.2 sind auch die Residuen $y_i - \hat{y}_i$, also die Differenzen zwischen den gemessenen und geschätzten Werten, angedeutet. Die Art der obigen Berechnung der Parameter $\hat{a}$ und $\hat{b}$ ergibt sich auch aus dem Prinzip der *kleinsten Quadrate*, das heißt, die Gerade wird so gewählt, dass die Summe der quadrierten Residuen minimal wird. Die Motivierung kommt auch aus der Ausgleichsrechnung.[2]

7.3 Schätzungen und Tests bei Normalverteilung

7.3.1 Konfidenzintervalle der Parameter

Bis jetzt wurde nur angenommen, dass die Varianz $\sigma_{y.x}^2 = \sigma^2$ für alle Werte von x gleich und dass die Regression linear ist. Wenn wir nun zusätzlich die Verteilung

[2] ℝ: `summary(lm(y ~ x))`

von Y bei jedem Wert x als normal annehmen, können wir Konfidenzintervalle für die Parameter a, b, σ^2 und $\mu_{y.x}$ angeben. Es gilt dann, dass die Statistiken

$$T_a = \frac{(\bar{Y} - a)\sqrt{n}}{S}$$

und

$$T_b = \frac{(\hat{b} - b)s_x\sqrt{n-1}}{S}$$

eine t-Verteilung mit $n-2$ Freiheitsgraden besitzen, die Verteilung von

$$(n-2)\frac{S^2}{\sigma^2}$$

ist χ^2_{n-2} mit $n-2$ Freiheitsgraden. Konfidenzintervalle[3] mit der Konfidenzzahl α erhält man folglich sofort als

$$\bar{Y} - t_{n-2;1-\frac{\alpha}{2}}\frac{S}{\sqrt{n}} < a < \bar{Y} + t_{n-2;1-\frac{\alpha}{2}}\frac{S}{\sqrt{n}} \quad ,$$

$$\hat{b} - t_{n-2;1-\frac{\alpha}{2}}\frac{S}{s_x\sqrt{n-1}} < b < \hat{b} + t_{n-2;1-\frac{\alpha}{2}}\frac{S}{s_x\sqrt{n-1}}$$

und

$$(n-2)\frac{S^2}{\chi^2_{n-2;1-\frac{\alpha}{2}}} < \sigma^2 < (n-2)\frac{S^2}{\chi^2_{n-2;\frac{\alpha}{2}}} \quad .$$

Für unser obiges Beispiel ergeben sich 90%-Konfidenzintervalle als

$$70.83 - 1.81\sqrt{\frac{23.12}{12}} < a < 70.83 + 1.81\sqrt{\frac{23.12}{12}}$$

oder

$$68.32 < a < 73.34 \quad ,$$

für b

$$1.01 - 1.81\sqrt{\frac{23.12}{97.54 * 11}} < b < 1.01 + 1.81\sqrt{\frac{23.12}{97.54 * 11}}$$

oder

$$.74 < b < 1.28 \quad ,$$

und für σ^2

$$10\frac{23.12}{18.31} < \sigma^2 < 10\frac{23.12}{3.94}$$

oder

$$12.63 < \sigma^2 < 58.68 \quad .$$

[3] R: `confint(lm(y ~ x))`

7.3.2 Schätzung der Mittelwerte und zukünftiger Beobachtungen

Ein Konfidenzintervall[4] für den Mittelwert $\mu_{y.x}$ an der Stelle x erhält man mit der Formel

$$\hat{y}_x - t_{n-2;1-\frac{\alpha}{2}} S \sqrt{\frac{1}{n} + \frac{(x-\bar{x})^2}{(n-1)s_x^2}} < \mu_{y.x} < \hat{y}_x + t_{n-2;1-\frac{\alpha}{2}} S \sqrt{\frac{1}{n} + \frac{(x-\bar{x})^2}{(n-1)s_x^2}}.$$

In unserem Beispiel erhalten wir für Männer mit $x = 162.5\ cm$ Körpergröße einen geschätzten mittleren Wert für das Körpergewicht

$$\hat{y}_{162.5} = 70.83 + 1.01(162.5 - 159.58) = 73.78$$

und ein 95%-Konfidenzintervall

$$73.78 - 2.23\sqrt{23.12[\frac{1}{12} + \frac{(162.5-159.58)^2}{11 * 97.54}]} < \mu_{y.162.5}$$

$$< 73.78 + 2.23\sqrt{23.12[\frac{1}{12} + \frac{(162.5-159.58)^2}{11 * 97.54}]}$$

oder

$$70.54 < \mu_{y.162.5} < 77.02 \ .$$

Wollen wir eine Aussage über eine zukünftige Beobachtung y an der Stelle x machen[5], so kommt zur Varianz von $\hat{y}_x$ noch ein σ^2 dazu und wir erhalten

$$\hat{y}_x - t_{n-2;1-\frac{\alpha}{2}} S \sqrt{1 + \frac{1}{n} + \frac{(x-\bar{x})^2}{(n-1)s_x^2}} < y$$

$$< \hat{y}_x + t_{n-2;1-\frac{\alpha}{2}} S \sqrt{1 + \frac{1}{n} + \frac{(x-\bar{x})^2}{(n-1)s_x^2}} \ .$$

Dies ist ein Toleranzintervall für *einen* an der Stelle x zu beobachtenden Wert, das auf Grund der Information aus der Stichprobe gefunden wurde. Für unser Beispiel erhalten wir an der Stelle $x = 162.5$ ($\alpha = .05$)

$$73.78 - 2.23\sqrt{23.12[1 + \frac{1}{12} + \frac{(162.5-159.58)^2}{11 * 97.54}]} < y_x < \ldots$$

oder

$$62.58 < y_x < 84.98 \ .$$

[4] R: `predict(lm(y ~ x), interval='confidence')`
[5] R: `predict(lm(y ~ x), interval='prediction')`

7.3.3 Test auf Abhängigkeit

Eine häufig aufgestellte Hypothese ist die der Abhängigkeit der Variablen Y von x. Eine Methode, diese zu testen, ist auf Gleichheit der Mittelwerte von Y bei allen Werten von x zu testen. Dieser Fall bedeutet aber in der betrachteten linearen Regression

$$H_o : b = 0 \quad .$$

Algorithmisch würde ein Test so aussehen:

1. Die Hypothese $b = 0$ wird getestet. Wird sie verworfen, so gibt dies auf Grund der Stichprobe genügend Grund zur Annahme, dass Y von x abhängt.
2. $H_o : b = 0$ mit der Alternative $b \neq 0$ (oder > 0 oder < 0).
3. Man wähle ein α.
4. Die Teststatistik sei
$$T = \frac{(\hat{b} - 0)s_x\sqrt{n-1}}{S} \quad .$$
5. Wenn die Verteilung von Y normal mit gleichem Mittel und Varianz für jedes x ist, so besitzt T eine t-Verteilung mit $n - 2$ Freiheitsgraden.
6. Der kritische Bereich wird dann als $(-\infty, -t_{n-2;1-\frac{\alpha}{2}}) \cup (t_{n-2;1-\frac{\alpha}{2}}, \infty)$ berechnet.
7. Man berechne den Wert für T und sehe nach, ob er in den kritischen Bereich fällt.
8. Man verwerfe oder akzeptiere entsprechend die Nullhypothese.
9. Man ziehe die Schlussfolgerung über die Abhängigkeit oder Unabhängigkeit zwischen Y und x.

In unserem numerischen Beispiel ergibt sich ein Wert für T als

$$1.01\sqrt{\frac{97.54 \times 11}{23.12}} = 6.88 \quad ,$$

wobei der kritische Bereich (bei $\alpha = .05$) mit $T < -2.23$ und $T > 2.23$ gegeben ist, sodass wir auf Abhängigkeit des Körpergewichts von der Körpergröße schließen müssen.

7.4 Das Korrelationsproblem

Im Gegensatz zur Abhängigkeit einer Zufallsvariablen von einer deterministischen Größe betrachten wir jetzt den Zusammenhang zwischen zwei zufälligen Größen. In einer Stichprobe müssen hier immer paarweise Messungen vorliegen. Meistens werden Analysen unter der Annahme, dass das Paar der betrachteten Zufallsvariablen (X, Y) eine *bivariate Normalverteilung* aufweist, durchgeführt. Diese ist in Abbildung 7.3 dargestellt. Es ist keine der Variablen ausgezeichnet. Bei jedem fixen Wert von X besitzt Y eine Normalverteilung und umgekehrt. Neben den Mittelwerten μ_X, μ_Y und den Varianzen $\sigma_X^2 = E(X - \mu_X)^2$, $\sigma_Y^2 = E(Y - \mu_Y)^2$ dient zur Charakterisierung dieser bivariaten Verteilung als Maß der Abhängigkeit zwischen X und Y noch die *Kovarianz*

$$\sigma_{XY} = E[(X - \mu_X)(Y - \mu_Y)] \ .$$

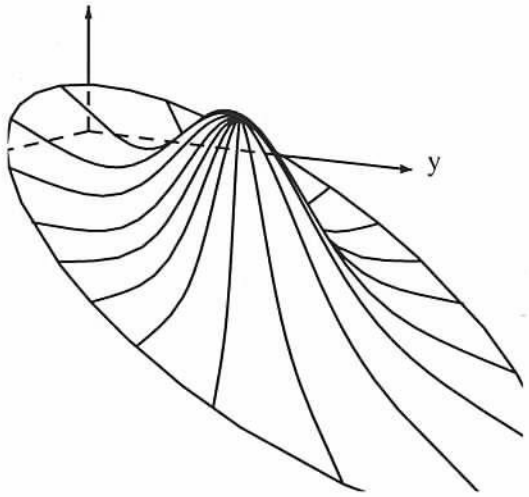


Abbildung 7.3: Dichte der bivariaten Normalverteilung.

Als relative (dimensionslose) Größe ist die *Korrelation* zwischen X und Y als

$$\rho_{XY} = \frac{\sigma_{XY}}{\sigma_X \sigma_Y}$$

definiert. Ihr Wert liegt zwischen -1 und +1. Unabhängigkeit der beiden Variablen bedeutet $\sigma_{XY} = 0$ und damit $\rho_{XY} = 0$. Als Schätzung für ρ dient meistens der empirische Korrelationskoeffizient

$$r_{XY} = \frac{1}{s_X s_Y} \frac{1}{n-1} \sum (x_i - \bar{x})(y_i - \bar{y}) \ .$$

Das am Anfang des Kapitels angeführte Beispiel der Körpergrößen und Gewichte kann natürlich auch als Korrelationsproblem interpretiert werden. Als em-

pirischen Korrelationskoeffizient errechnen wir

$$r_{XY} = \frac{98.11}{\sqrt{97.54 * 119.70}} = .91 \quad .$$

Test auf Unkorreliertheit

Sind die beiden Zufallsvariablen X und Y voneinander unabhängig und normalverteilt, so besitzt die Statistik

$$T = R\sqrt{\frac{n-2}{1-R^2}}$$

eine t_{n-2}-Verteilung, wobei R die Zufallsvariable bezeichnet, die die Werte des empirischen Korrelationskoeffizienten r_{XY} annimmt. T kann sofort als Teststatistik zum Testen der Nullhypothese $H_o : \rho = 0$ verwendet werden. Bei Spezifizierung der Gegenhypothese $H_1 : \rho \neq 0$ ergibt sich als kritischer Bereich

$$| T | > t_{n-2;1-\frac{\alpha}{2}} \text{ [6]} .$$

Beispiel 7.1: Betrachten wir die Abhängigkeit des Eisengehaltes Y (in %) kieseliger Hämatiterze von der Dichte X (g/cm^3), wie im Beispiel auf Seite 67. Nun testen wir $H_o : \rho = 0$ gegen $H_1 : \rho \neq 0$ mit $\alpha = .05$. Der Wert des empirischen Korrelationskoeffizienten R beträgt $r = .69$. Mit $n = 9$ ergibt sich der Wert der Teststatistik T als

$$t = r\sqrt{\frac{n-2}{1-r^2}} = .69\sqrt{\frac{7}{1-.69^2}} = 2.52 \quad ,$$

was absolut größer als $t_{n-2;1-\frac{\alpha}{2}} = t_{7;.975} = 2.365$ ausfällt. Die Hypothese der Unkorreliertheit muss daher verworfen werden.

[6] R: `cor.test(Daten1, Daten2)`

Kapitel 8

Zählstatistiken

In vielen Problemen ist man nicht an der Messung von gewissen Variablen interessiert, sondern an der Zählung von Dingen, die in eine gewisse Kategorie fallen. Zum Beispiel möchten wir wissen, wie viele Stimmen auf eine Partei entfallen, wie viele unbrauchbare Stücke in der Produktion von Maschinenelementen anfallen, etc. Die drei wichtigsten Verteilungen, die bei der Analyse auftreten, sind die χ^2-, die Binomial- und die Poissonverteilung.

Nehmen wir an, es gäbe k Kategorien und eine Stichprobe der Größe n, deren Werte in jeweils genau eine Kategorie fallen. Die beobachteten Häufigkeiten in jeder Kategorie seien $h_1, \ldots, h_k$, wobei natürlich gilt $\sum h_j = n$. Nehmen wir weiters an, dass es theoretische Häufigkeiten $e_j = np_j$, $j = 1, \ldots, k$, für die Kategorien gäbe. Wenn wir die Frage untersuchen, ob die beobachteten Häufigkeiten von den theoretischen nur zufällig abweichen, dann verwenden wir die Teststatistik

$$T = \sum_{j=1}^{k} \frac{(h_j - e_j)^2}{e_j} \quad ,$$

die eine ungefähre χ^2-Verteilung mit $k - 1$ Freiheitsgraden aufweist. (Häufig bezeichnet man diese Statistik auch gleich mit χ^2.) Diese Vorgangsweise vollzieht sich analog dem Anpassungstest, den wir im Abschnitt 5.8 behandelt haben. Allerdings kann die Kategorien- (Klassen-)einteilung auch natürlich gegeben sein. Man sollte nochmals bemerken, dass die Anzahl der Beobachtungen n in die Teststatistik gar nicht eingeht (nur k); die Verteilung ist allerdings nur approximativ gegeben, und der Test sollte für kleine n's mit Vorsicht durchgeführt werden.

8.1 Einfache Klassifizierung

In einem einfachen Klassifizierungsproblem sind die theoretischen Anteile der Klassen von vornherein festgelegt.

Klassen	beobachtet	theoretisch
1	h_1	e_1
2	h_2	e_2
.	.	.
.	.	.
.	.	.
k	h_k	e_k
Total	n	n

Beispiel 8.1: Mendel erhielt bei einem seiner Kreuzungsversuche an Erbsenpflanzen folgende Werte.

Klassen	beobachtet	theoretisch	$h_j - e_j$	$\frac{(h_j-e_j)^2}{e_j}$
rund und gelb	315	312.75	2.25	.016
kantig und gelb	101	104.25	-3.25	.101
rund und grün	108	104.25	3.75	.135
kantig und grün	32	34.75	-2.75	.218
Total	556	556.00		.470

Die Theorie spricht für 9:3:3:1 als Verhältnis zwischen den Klassen. Gibt diese Stichprobe (der Kreuzungsversuch) diesem Statement recht? Die χ^2-Statistik ergibt einen Wert von .47 bei 3 Freiheitsgraden. Nimmt man $\alpha = .05$, so erhält man einen kritischen Wert von $\chi^2_{3;.95} = 7.81$, der weit höher liegt. Dies gibt *nicht* genügend Grund, um die Hypothese zu verwerfen.

8.2 Zweifache Klassifizierung

Wir nehmen an, dass die Daten nach zwei Gesichtspunkten (Merkmalen) eingeteilt werden können. Wir werden das Problem der Unabhängigkeit dieser beiden Merkmale untersuchen. Unabhängigkeit heißt dabei (wie früher), dass die Verteilung des einen Merkmals die gleiche ist für jeden beliebigen Wert des anderen.

Wenn zum Beispiel die Augenfarbe von der Haarfarbe der Menschen unabhängig ist, so müsste der Anteil der Blauäugigen mit hellem Haar gleich dem der Blauäugigen mit dunklem sein. Bei einer Stichprobe treten natürlich Abweichungen auf, und wir untersuchen, ob diese „signifikant" sind.

	Haarfarbe		
Augenfarbe	hell	dunkel	Total
blau	32(24.1)	12(19.9)	44
braun	14(19.7)	22(16.3)	36
anders	6(8.2)	9(6.8)	15
Total	52	43	95

Die Abweichungen werden wieder mit der χ^2-Statistik getestet. Die „theoretischen" Anteile schätzen wir aus den Randwerten. Es werden 95 Personen untersucht, von denen 52 helle Haarfarbe aufweisen. Dieser Anteil sollte bei Annahme der Unabhängigkeit auch bei den Blauäugigen auftreten, nämlich bei $\frac{52}{95} \times 44 = 24.1$. Analog errechnet sich der Anteil der Braunäugigen mit $\frac{52}{95} \times 36 = 19.7$. (Die Werte sind in der Tafel zwischen Klammern gegeben). Die χ^2-Statistik errechnet sich wie früher, nämlich

$$\chi^2 = \sum \frac{(h_{ij} - \hat{e}_{ij})^2}{\hat{e}_{ij}} = \frac{(32 - 24.1)^2}{24.1} + \cdots = 10.67 \quad ,$$

wobei $\hat{e}_{ij}$ die geschätzte, erwartete Anzahl der Objekte der Kategorie (i, j) bezeichnet. Die Teststatistik besitzt wieder eine ungefähre χ^2- Verteilung. Bei der Berechnung der Freiheitsgrade muss allerdings noch die Anzahl der geschätzten Anteile, hier der Randwerte, abgezogen werden. Bei r Zeilen und c Spalten $(k = rc)$ erhalten wir also $rc - 1 - (r - 1) - (c - 1) = (r - 1)(c - 1)$ Freiheitsgrade. In unserem Beispiel finden wir auf dem Niveau $\alpha = .05$ das Quantil $\chi^2_{2;.95} = 5.99$, wobei der Wert der Teststatistik viel größer ist, und die Hypothese der Unabhängigkeit muss verworfen werden.

So eine Tafel der zweifachen Klassifizierung wird auch häufig *Kontingenztafel* genannt. Allgemein stellt sie sich folgendermaßen dar, wobei

$$\hat{e}_{ij} = h_{i.} h_{.j} / n$$

und für die Randsumme

$$h_{i.} = \sum_{j=1}^{c} h_{ij}, \; h_{.j} = \sum_{i=1}^{r} h_{ij}$$

gilt

	Merkmal B				
Merkmal A	B_1	B_2	$\ldots$	B_c	$\sum$
A_1	$h_{11}(\hat{e}_{11})$	$h_{12}(\hat{e}_{12})$	$\ldots$	$h_{1c}(\hat{e}_{1c})$	$h_{1.}$
A_2	$h_{21}(\hat{e}_{21})$	$h_{22}(\hat{e}_{22})$	$\ldots$	$h_{2c}(\hat{e}_{2c})$	$h_{2.}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
A_r	$h_{r1}(\hat{e}_{r1})$	$h_{r2}(\hat{e}_{r2})$	$\ldots$	$h_{rc}(\hat{e}_{rc})$	$h_{r.}$
$\sum$	$h_{.1}$	$h_{.2}$	$\ldots$	$h_{.c}$	$h_{..} = n$

R:

```
> print(farb[1:10, ])

   Augenfarbe Haarfarbe
1        blau      hell
2      anders    dunkel
3       braun    dunkel
4        blau    dunkel
5        blau      hell
6       braun    dunkel
7        blau      hell
8        blau      hell
9      anders      hell
10     anders    dunkel

> print(table(farb))

          Haarfarbe
Augenfarbe dunkel hell
    anders      9    6
    blau       12   32
    braun      22   14

> print(chisq.test(table(farb)))

        Pearson's Chi-squared test

data:  table(farb)
X-squared = 10.7122, df = 2, p-value = 0.004719
```

Literaturverzeichnis

[1] ISO 2859. *Sampling Procedures and Tables for Inspection by Attributes.* International Organization for Standardization, Genf, 1974.

[2] ISO 3951. *Sampling Procedures and Charts for Inspection by Variables for Percent Defective.* International Organization for Standardization, Genf, 1976.

[3] A. Behr. *Einführung in die Statistik mit R.* Vahlen, München, 2005.

[4] R.M. Betha and R.R. Rhinehart. *Applied Engineering Statistics.* Statistics: Textbooks and Monographs Series, 121. Marcel Dekker, Inc., New York, 1991.

[5] J.P. Bläsing. *Statistische Qualitätskontrolle.* gfmt – Gesellschaft für Management und Technologie AG, St. Gallen, 1989.

[6] H. Büning und G. Trenkler. *Nichtparametrische statistische Methoden.* Walter de Gruyter, Berlin, 1994.

[7] I.W. Burr. *Statistical Quality Control Methods.* Marcel Dekker Inc., New York, 1976.

[8] W.J. Dixon and F.J. Massey. *Introduction to Statistical Analysis.* McGraw Hill, New York, 1969.

[9] John Fox. *Applied Regression Analysis, Linear Models, and Related Methods.* Sage Publications, Thousand Oaks, CA, USA, 1997. ISBN 0-8039-4540-X.

[10] John Fox. *An R and S-Plus Companion to Applied Regression.* Sage Publications, Thousand Oaks, CA, USA, 2002. ISBN 0-761-92279-2.

[11] I. Fraser. *Probability and Statistics: Theory Applications.* Duxbury Press, North Scituate, Mass., 1976.

[12] I. Guttmann. *Statistical Tolerance Regions: Classical and Bayesian.* Griffin, London, 1970.

[13] I. Guttmann, S.S. Wilks and J.S. Hunter. *Introductory Engineering Statistics.* Wiley & Sons, New York, 1982.

[14] J. Hartung, B. Elpelt und H.-K. Klösener. *Statistik. Lehr- und Handbuch der angewandten Statistik.* Oldenbourg Verlag, München, 1984.

[15] Bleymüller. J., G. Gehlert, and H. Gülicher. *Statistik für Wirtschaftswissenschaftler.* Vahlen, München, 13. edition, 2002.

[16] M. Kendall and A. Stuart. *Advanced Theory of Statistics*, volume 1. Griffin, London, 1963.

[17] E. Kreyszig. *Statistische Methoden und ihre Anwendungen.* Vandenhœk, Göttingen, 1972.

[18] J. Lehn and H. Wegmann. *Einführung in die Statistik.* Teubner, Stuttgart, 4. edition, 2004.

[19] W. Masing. *Handbuch der Qualitätssicherung.* Carl Hanser Verlag, München, 1980.

[20] G. Pflug. *Stochastische Modelle in der Informatik.* Teubner Verlag, Stuttgart, 1986.

[21] C. Reimann, P. Filzmoser, R.G. Garrett, and R. Dutter. *Statistical Data Analysis Explained. Applied Environmental Statistics with R.* Wiley & Sons, New York, 2008.

[22] L. Sachs. *Angewandte Statistik.* Springer, Berlin, 11. edition, 2004.

[23] K. Sarkadi and I. Vince. *Mathematical Methods of Statistical Quality Control.* Academic Press, New York, 1974.

[24] A.H. Schaafsma and F.G. Willemze. *Moderne Qualitätskontrolle.* Philips Technische Bibliothek, Eindhoven, 1961.

[25] E. Schindowsky and O. Schürz. *Statistische Qualitätskontrolle.* VEB Verlag, Berlin, 1974.

[26] W. Shewhart. *Economic Control of Quality of Manufactured Product.* Van Nostrand, Princeton, 1931.

[27] W.A. Stahel. *Statistische Datenanalyse. Eine Einführung für Naturwissenschaftler.* Vieweg, Braunschweig, 4. edition, 2002.

[28] W. Timischl. *Biostatistik. Eine Einführung für Biologen und Mediziner.* Springer, Wien, 2. edition, 2000.

[29] J.W. Tukey. *Exploratory Data Analysis.* Addison-Wesley, Reading, Mass., 1977.

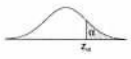
[30] R. Viertl. *Einführung in die Stochastik. Mit Elementen der Bayes-Statistik und Ansätzen für die Analyse unscharfer Daten.* Springer, Wien, 3. edition, 2003.

[31] B.J. Winer. *Statistical Principles in Experimental Design.* McGraw-Hill Comp., New York, 1971.

Anhang A

Tabellen von Verteilungen: Quantile, kritische Werte

Table A.1: $N(0,1)$-Verteilung. $\alpha = P(Z \geq z_\alpha) = 1 - G(z_\alpha)$.

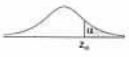


z_α	**0.00**	**0.01**	**0.02**	**0.03**	**0.04**	**0.05**	**0.06**	**0.07**	**0.08**	**0.09**
.0	.5000	.4960	.4920	.4880	.4840	.4801	.4761	.4721	.4681	.4641
.1	.4602	.4562	.4522	.4483	.4443	.4404	.4364	.4325	.4286	.4247
.2	.4207	.4168	.4129	.4090	.4052	.4013	.3974	.3936	.3897	.3859
.3	.3821	.3783	.3745	.3707	.3669	.3632	.3594	.3557	.3520	.3483
.4	.3446	.3409	.3372	.3336	.3300	.3264	.3228	.3192	.3156	.3121
.5	.3085	.3050	.3015	.2981	.2946	.2912	.2877	.2843	.2810	.2776
.6	.2743	.2709	.2676	.2643	.2611	.2578	.2546	.2514	.2483	.2451
.7	.2420	.2389	.2358	.2327	.2296	.2266	.2236	.2206	.2177	.2148
.8	.2119	.2090	.2061	.2033	.2005	.1977	.1949	.1922	.1894	.1867
.9	.1841	.1814	.1788	.1762	.1736	.1711	.1685	.1660	.1635	.1611
1.0	.1587	.1562	.1539	.1515	.1492	.1469	.1446	.1423	.1401	.1379
1.1	.1357	.1335	.1314	.1292	.1271	.1251	.1230	.1210	.1190	.1170
1.2	.1151	.1131	.1112	.1093	.1075	.1056	.1038	.1020	.1003	.0985
1.3	.0968	.0951	.0934	.0918	.0901	.0885	.0869	.0853	.0838	.0823
1.4	.0808	.0793	.0778	.0764	.0749	.0735	.0721	.0708	.0694	.0681
1.5	.0668	.0655	.0643	.0630	.0618	.0606	.0594	.0582	.0571	.0559
1.6	.0548	.0537	.0526	.0516	.0505	.0495	.0485	.0475	.0465	.0455
1.7	.0446	.0436	.0427	.0418	.0409	.0401	.0392	.0384	.0375	.0367
1.8	.0359	.0351	.0344	.0336	.0329	.0322	.0314	.0307	.0301	.0294
1.9	.0287	.0281	.0274	.0268	.0262	.0256	.0250	.0244	.0239	.0233
2.0	.0228	.0222	.0217	.0212	.0207	.0202	.0197	.0192	.0188	.0183
2.1	.0179	.0174	.0170	.0166	.0162	.0158	.0154	.0150	.0146	.0143
2.2	.0139	.0136	.0132	.0129	.0125	.0122	.0119	.0116	.0113	.0110
2.3	.0107	.0104	.0102	.0099	.0096	.0094	.0091	.0089	.0087	.0084
2.4	.0082	.0080	.0078	.0075	.0073	.0071	.0069	.0068	.0066	.0064
2.5	.0062	.0060	.0059	.0057	.0055	.0054	.0052	.0051	.0049	.0048
2.6	.0047	.0045	.0044	.0043	.0041	.0040	.0039	.0038	.0037	.0036
2.7	.0035	.0034	.0033	.0032	.0031	.0030	.0029	.0028	.0027	.0026
2.8	.0026	.0025	.0024	.0023	.0023	.0022	.0021	.0021	.0020	.0019
2.9	.0019	.0018	.0018	.0017	.0016	.0016	.0015	.0015	.0014	.0014
3.0	.0013	.0013	.0013	.0012	.0012	.0011	.0011	.0011	.0010	.0010
3.1	.0010	.0009	.0009	.0009	.0008	.0008	.0008	.0008	.0007	.0007
3.2	.0007	.0007	.0006	.0006	.0006	.0006	.0006	.0005	.0005	.0005
3.3	.0005	.0005	.0005	.0004	.0004	.0004	.0004	.0004	.0004	.0003
3.4	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0002

Die Tabelle enthält Werte von α. Die Zeilen sind in Stufen von 0.1 von z_α zu verstehen, die Spalten in Stufen von 0.01. Z.B. wird α für $z_\alpha = 1.96$ aus der Zeile 1.9 und Spalte .06 gelesen, d.h. $\alpha = .025$. [1]

[1] R: `1-pnorm(`z_α`)`

Table A.2: Student-t-Verteilung. Rechte Quantile $t_{n;\alpha}$; $\alpha = P(T \geq t_{n;\alpha})$



	α					
FG	**.25**	**.1**	**.05**	**.025**	**.01**	**.005**
1	1.000	3.078	6.314	12.706	31.824	63.659
2	.816	1.886	2.920	4.303	6.965	9.925
3	.765	1.638	2.353	3.182	4.541	5.841
4	.741	1.533	2.132	2.776	3.747	4.604
5	.727	1.476	2.015	2.571	3.365	4.032
6	.718	1.440	1.943	2.447	3.143	3.707
7	.711	1.415	1.895	2.365	2.998	3.499
8	.706	1.397	1.860	2.306	2.896	3.355
9	.703	1.383	1.833	2.262	2.821	3.250
10	.700	1.372	1.812	2.228	2.764	3.169
11	.697	1.363	1.796	2.201	2.718	3.106
12	.695	1.356	1.782	2.179	2.681	3.055
13	.694	1.350	1.771	2.160	2.650	3.012
14	.692	1.345	1.761	2.145	2.624	2.977
15	.691	1.341	1.753	2.131	2.602	2.947
16	.690	1.337	1.746	2.120	2.583	2.921
17	.689	1.333	1.740	2.110	2.567	2.898
18	.688	1.330	1.734	2.101	2.552	2.878
19	.688	1.328	1.729	2.093	2.539	2.861
20	.687	1.325	1.725	2.086	2.528	2.845
21	.686	1.323	1.721	2.080	2.518	2.831
22	.686	1.321	1.717	2.074	2.508	2.819
23	.685	1.319	1.714	2.069	2.500	2.807
24	.685	1.318	1.711	2.064	2.492	2.797
25	.684	1.316	1.708	2.060	2.485	2.787
26	.684	1.315	1.706	2.056	2.479	2.779
27	.684	1.314	1.703	2.052	2.473	2.771
28	.683	1.313	1.701	2.048	2.467	2.763
29	.683	1.311	1.699	2.045	2.462	2.756
30	.683	1.310	1.697	2.042	2.457	2.750
31	.682	1.309	1.696	2.040	2.453	2.744
32	.682	1.309	1.694	2.037	2.449	2.738
33	.682	1.308	1.692	2.035	2.445	2.733
34	.682	1.307	1.691	2.032	2.441	2.728
35	.682	1.306	1.690	2.030	2.438	2.724
40	.681	1.303	1.684	2.021	2.423	2.704
45	.680	1.301	1.679	2.014	2.412	2.690
50	.679	1.299	1.676	2.009	2.403	2.678
55	.679	1.297	1.673	2.004	2.396	2.668
60	.679	1.296	1.671	2.000	2.390	2.660
65	.678	1.295	1.669	1.997	2.385	2.654
70	.678	1.294	1.667	1.994	2.381	2.648
75	.678	1.293	1.665	1.992	2.377	2.643
80	.678	1.292	1.664	1.990	2.374	2.639
85	.677	1.292	1.663	1.988	2.371	2.635
90	.677	1.291	1.662	1.987	2.368	2.632
95	.677	1.291	1.661	1.985	2.366	2.629
100	.677	1.290	1.660	1.984	2.364	2.626

Die Tabelle enthält die rechten α-Quantile der t-Verteilung mit n Freiheitsgraden (FG). Die Zeilen sind die FG, die Spalten beziehen sich auf α.

R: `1-qt(alpha,n)`

Table A.3: Chi-Quadrat-Verteilung. Rechte Quantile $\chi^2_{n;\alpha}$; $\alpha = P(\chi^2 \geq \chi^2_{n;\alpha})$

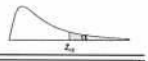
	α										
FG	**.995**	**.99**	**.975**	**.95**	**.9**	**.5**	**.1**	**.05**	**.025**	**.01**	**.005**
1	.000	.000	.001	.004	.016	.455	2.706	3.841	5.024	6.635	7.879
2	.010	.020	.051	.103	.211	1.386	4.605	5.991	7.378	9.210	10.597
3	.072	.115	.216	.352	.584	2.366	6.251	7.815	9.348	11.345	12.838
4	.207	.297	.484	.711	1.064	3.357	7.779	9.488	11.143	13.277	14.860
5	.412	.554	.831	1.145	1.610	4.351	9.236	11.070	12.832	15.086	16.750
6	.676	.872	1.237	1.635	2.204	5.348	10.645	12.592	14.449	16.812	18.548
7	.989	1.239	1.690	2.167	2.833	6.346	12.017	14.067	16.013	18.475	20.278
8	1.344	1.647	2.180	2.733	3.490	7.344	13.362	15.507	17.535	20.090	21.955
9	1.735	2.088	2.700	3.325	4.168	8.343	14.684	16.919	19.023	21.666	23.589
10	2.156	2.558	3.247	3.940	4.865	9.342	15.987	18.307	20.483	23.209	25.188
11	2.603	3.053	3.816	4.575	5.578	10.341	17.275	19.675	21.920	24.725	26.757
12	3.074	3.571	4.404	5.226	6.304	11.340	18.549	21.026	23.337	26.217	28.299
13	3.565	4.107	5.009	5.892	7.042	12.340	19.812	22.362	24.736	27.688	29.819
14	4.075	4.660	5.629	6.571	7.790	13.339	21.064	23.685	26.119	29.141	31.319
15	4.601	5.229	6.262	7.261	8.547	14.339	22.307	24.996	27.488	30.578	32.801
16	5.142	5.812	6.908	7.962	9.312	15.338	23.542	26.296	28.845	32.000	34.267
17	5.697	6.408	7.564	8.672	10.085	16.338	24.769	27.587	30.191	33.409	35.718
18	6.265	7.015	8.231	9.390	10.865	17.338	25.989	28.869	31.526	34.805	37.156
19	6.844	7.633	8.907	10.117	11.651	18.338	27.204	30.144	32.852	36.191	38.582
20	7.434	8.260	9.591	10.851	12.443	19.337	28.412	31.410	34.170	37.566	39.997
21	8.034	8.897	10.283	11.591	13.240	20.337	29.615	32.671	35.479	38.932	41.401
22	8.643	9.543	10.982	12.338	14.041	21.337	30.813	33.924	36.781	40.289	42.796
23	9.261	10.196	11.689	13.091	14.848	22.337	32.007	35.172	38.076	41.638	44.181
24	9.886	10.856	12.401	13.848	15.659	23.337	33.196	36.415	39.364	42.980	45.559
25	10.520	11.524	13.120	14.611	16.473	24.337	34.382	37.652	40.646	44.314	46.928
26	11.160	12.198	13.844	15.379	17.292	25.336	35.563	38.885	41.923	45.642	48.290
27	11.808	12.879	14.573	16.151	18.114	26.336	36.741	40.113	43.195	46.963	49.645
28	12.461	13.565	15.308	16.928	18.939	27.336	37.916	41.337	44.461	48.278	50.994
29	13.121	14.257	16.047	17.708	19.768	28.336	39.087	42.557	45.722	49.588	52.336
30	13.787	14.953	16.791	18.493	20.599	29.336	40.256	43.773	46.979	50.892	53.672
31	14.458	15.655	17.539	19.281	21.434	30.336	41.422	44.985	48.232	52.191	55.003
32	15.134	16.362	18.291	20.072	22.271	31.336	42.585	46.194	49.480	53.486	56.328
33	15.815	17.074	19.047	20.867	23.110	32.336	43.745	47.400	50.725	54.776	57.648
34	16.501	17.789	19.806	21.664	23.952	33.336	44.903	48.602	51.966	56.061	58.964
35	17.192	18.509	20.569	22.465	24.797	34.336	46.059	49.802	53.203	57.342	60.275
40	20.707	22.164	24.433	26.509	29.051	39.335	51.805	55.758	59.342	63.691	66.766
45	24.311	25.901	28.366	30.612	33.350	44.335	57.505	61.656	65.410	69.957	73.166
50	27.991	29.707	32.357	34.764	37.689	49.335	63.167	67.505	71.420	76.154	79.490
55	31.735	33.570	36.398	38.958	42.060	54.335	68.796	73.311	77.380	82.292	85.749
60	35.534	37.485	40.482	43.188	46.459	59.335	74.397	79.082	83.298	88.379	91.952
65	39.383	41.444	44.603	47.450	50.883	64.335	79.973	84.821	89.177	94.422	98.105
70	43.275	45.442	48.758	51.739	55.329	69.334	85.527	90.531	95.023	100.425	104.215
75	47.206	49.475	52.942	56.054	59.795	74.334	91.061	96.217	100.839	106.393	110.286
80	51.172	53.540	57.153	60.391	64.278	79.334	96.578	101.879	106.629	112.329	116.321
85	55.170	57.634	61.389	64.749	68.777	84.334	102.079	107.522	112.393	118.236	122.325
90	59.196	61.754	65.647	69.126	73.291	89.334	107.565	113.145	118.136	124.116	128.299
95	63.250	65.898	69.925	73.520	77.818	94.334	113.038	118.752	123.858	129.973	134.247
100	67.328	70.065	74.222	77.929	82.358	99.334	118.498	124.342	129.561	135.807	140.169

Die Tabelle enthält die rechten α-Quantile der χ^2-Verteilung mit n Freiheitsgraden (FG). Die Zeilen sind die FG, die Spalten beziehen sich auf α. Z.B. wird für $\alpha = 0.025$ und 10 FG ein Wert $\chi^2_{10;0.025} = 20.483$ abgelesen. [2]

[2] ℝ: `1-qt(alpha,n)`

Table A.4: F-Verteilung. Rechte Quantile $F_{m,n;\alpha}$; $\alpha = P(F \geq F_{m,n;\alpha})$

$\alpha = .1$



FG	n = 1	2	3	4	5	6	7	8	9
m = 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	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	n = 10	12	15	20	30	60	120	200	500
m = 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

Die Tabellen enthalten die rechten α-Quantile der F-Verteilung mit m (Zeilen) und n (Spalten) Freiheitsgraden (FG). Die beiden obigen Tabellen sind für $\alpha = 0.1$, nachfolgende Tabellen für andere Werte von α. Z.B. wird für $\alpha = 0.1$ und $m = 10$ und $n = 12$ FG ein Wert $F_{10,12;0.1} = 2.188$ abgelesen.

R: `1-qf(alpha,m,n)`

Tabelle A.4: F-Verteilung. Fortsetzung

$\alpha = .05$

FG	n = 1	2	3	4	5	6	7	8	9
m = 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	n = 10	12	15	20	30	60	120	200	500
m = 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-Verteilung. Fortsetzung

$\alpha = .025$

FG	n = 1	2	3	4	5	6	7	8	9
m = 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	n = 10	12	15	20	30	60	120	200	500
m = 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-Verteilung. Fortsetzung

$\alpha = .01$

FG	n = 1	2	3	4	5	6	7	8	9
m = 1	4052.192	98.505	34.116	21.198	16.258	13.745	12.246	11.259	10.561
2	4998.686	99.002	30.817	18.000	13.274	10.925	9.546	8.649	8.022
3	5402.648	99.169	29.457	16.694	12.060	9.779	8.451	7.591	6.992
4	5623.821	99.252	28.710	15.977	11.392	9.148	7.847	7.006	6.422
5	5763.357	99.300	28.237	15.522	10.967	8.746	7.460	6.632	6.057
6	5858.054	99.335	27.911	15.207	10.672	8.466	7.191	6.371	5.802
7	5927.838	99.359	27.672	14.976	10.455	8.260	6.993	6.178	5.613
8	5980.675	99.376	27.489	14.799	10.289	8.101	6.840	6.029	5.467
9	6021.547	99.389	27.347	14.659	10.157	7.976	6.719	5.911	5.351
10	6055.443	99.400	27.229	14.546	10.051	7.874	6.620	5.814	5.257
12	6105.356	99.416	27.052	14.374	9.888	7.718	6.469	5.667	5.111
15	6156.220	99.434	26.872	14.198	9.722	7.559	6.314	5.515	4.962
20	6208.075	99.452	26.690	14.020	9.552	7.396	6.155	5.359	4.808
30	6259.915	99.468	26.506	13.838	9.379	7.228	5.992	5.198	4.649
60	6312.735	99.484	26.316	13.652	9.202	7.056	5.824	5.032	4.483
120	6338.517	99.491	26.221	13.558	9.111	6.969	5.737	4.946	4.398
200	6349.377	99.495	26.183	13.520	9.075	6.934	5.702	4.911	4.363
500	6358.308	99.499	26.148	13.486	9.042	6.902	5.671	4.880	4.332

FG	n = 10	12	15	20	30	60	120	200	500
m = 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-Verteilung. Fortsetzung

α = **.005**

FG	n = 1	2	3	4	5	6	7	8	9
m = 1	16205.232	198.502	55.553	31.333	22.785	18.635	16.235	14.688	13.614
2	19991.950	199.000	49.803	26.284	18.314	14.544	12.404	11.042	10.107
3	21606.355	199.167	47.473	24.259	16.530	12.916	10.882	9.596	8.717
4	22491.330	199.250	46.196	23.157	15.556	12.027	10.050	8.805	7.956
5	23046.763	199.301	45.394	22.456	14.940	11.463	9.522	8.301	7.471
6	23428.396	199.333	44.838	21.975	14.513	11.073	9.155	7.952	7.134
7	23705.137	199.358	44.436	21.622	14.200	10.786	8.885	7.694	6.885
8	23915.941	199.376	44.131	21.352	13.961	10.565	8.678	7.496	6.693
9	24081.789	199.390	43.882	21.139	13.772	10.391	8.514	7.339	6.541
10	24215.665	199.401	43.692	20.967	13.618	10.250	8.380	7.211	6.417
12	24417.562	199.417	43.388	20.705	13.384	10.034	8.176	7.015	6.227
15	24620.402	199.402	43.085	20.438	13.146	9.814	7.967	6.814	6.032
20	24826.230	199.450	42.777	20.167	12.903	9.588	7.754	6.608	5.832
30	25034.044	199.467	42.467	19.891	12.656	9.358	7.534	6.396	5.625
60	25243.835	199.484	42.149	19.611	12.402	9.122	7.309	6.177	5.410
120	25348.584	199.492	41.989	19.468	12.274	9.001	7.193	6.065	5.300
200	25391.425	199.495	41.925	19.411	12.222	8.952	7.147	6.019	5.255
500	25429.346	199.498	41.867	19.359	12.175	8.908	7.104	5.978	5.215

FG	n = 10	12	15	20	30	60	120	200	500
m = 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: Binomialverteilung. $F(x) = P(X \leq x) = \sum_{i=0}^{x} \binom{n}{i} p^i (1-p)^{n-i}$.

		p									
n	**x**	**0.01**	**0.05**	**0.1**	**0.15**	**0.2**	**0.25**	**0.3**	**0.35**	**0.4**	**0.45**
8	**0**	.9227	.6634	.4305	.2725	.1678	.1001	.0576	.0319	.0168	.0084
	1	.9973	.9428	.8131	.6572	.5033	.3671	.2553	.1691	.1064	.0632
	2	.9999	.9942	.9619	.8948	.7969	.6785	.5518	.4278	.3154	.2201
	3	1.0000	.9996	.9950	.9786	.9437	.8862	.8059	.7064	.5941	.4770
	4	1.0000	1.0000	.9996	.9971	.9896	.9727	.9420	.8939	.8263	.7396
	5	1.0000	1.0000	1.0000	.9998	.9988	.9958	.9887	.9747	.9502	.9115
	6	1.0000	1.0000	1.0000	1.0000	.9999	.9996	.9987	.9964	.9915	.9819
	7	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	.9999	.9998	.9993	.9983
	8	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
9	**0**	.9135	.6302	.3874	.2316	.1342	.0751	.0404	.0207	.0101	.0046
	1	.9966	.9288	.7748	.5995	.4362	.3003	.1960	.1211	.0705	.0385
	2	.9999	.9916	.9470	.8591	.7382	.6007	.4628	.3373	.2318	.1495
	3	1.0000	.9994	.9917	.9661	.9144	.8343	.7297	.6089	.4826	.3614
	4	1.0000	1.0000	.9991	.9944	.9804	.9511	.9012	.8283	.7334	.6214
	5	1.0000	1.0000	.9999	.9994	.9969	.9900	.9747	.9464	.9006	.8342
	6	1.0000	1.0000	1.0000	1.0000	.9997	.9987	.9957	.9888	.9750	.9502
	7	1.0000	1.0000	1.0000	1.0000	1.0000	.9999	.9996	.9986	.9962	.9909
	8	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	.9999	.9997	.9992
	9	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
10	**0**	.9044	.5987	.3487	.1969	.1074	.0563	.0282	.0135	.0060	.0025
	1	.9957	.9139	.7361	.5443	.3758	.2440	.1493	.0860	.0464	.0233
	2	.9999	.9885	.9298	.8202	.6778	.5256	.3828	.2616	.1673	.0996
	3	1.0000	.9990	.9872	.9500	.8791	.7759	.6496	.5138	.3823	.2660
	4	1.0000	.9999	.9984	.9901	.9672	.9219	.8497	.7515	.6331	.5044
	5	1.0000	1.0000	.9999	.9986	.9936	.9803	.9527	.9051	.8338	.7384
	6	1.0000	1.0000	1.0000	.9999	.9991	.9965	.9894	.9740	.9452	.8980
	7	1.0000	1.0000	1.0000	1.0000	.9999	.9996	.9984	.9952	.9877	.9726
	8	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	.9999	.9995	.9983	.9955
	9	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	.9999	.9997
	10	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
11	**0**	.8953	.5688	.3138	.1673	.0859	.0422	.0198	.0088	.0036	.0014
	1	.9948	.8981	.6974	.4922	.3221	.1971	.1130	.0606	.0302	.0139
	2	.9998	.9848	.9104	.7788	.6174	.4552	.3127	.2001	.1189	.0652
	3	1.0000	.9984	.9815	.9306	.8389	.7133	.5696	.4256	.2963	.1911
	4	1.0000	.9999	.9972	.9841	.9496	.8854	.7897	.6683	.5328	.3971
	5	1.0000	1.0000	.9997	.9973	.9883	.9657	.9218	.8513	.7535	.6331
	6	1.0000	1.0000	1.0000	.9997	.9980	.9924	.9784	.9499	.9006	.8262
	7	1.0000	1.0000	1.0000	1.0000	.9998	.9988	.9957	.9878	.9707	.9390
	8	1.0000	1.0000	1.0000	1.0000	1.0000	.9999	.9994	.9980	.9941	.9852
	9	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	.9998	.9993	.9978
	10	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	.9998
	11	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000

Tabelle A.5: Binomialverteilung. Fortsetzung.

		p									
n	**x**	**0.50**	**0.55**	**0.6**	**0.65**	**0.7**	**0.75**	**0.8**	**0.85**	**0.9**	**0.95**
8	0	.0039	.0017	.0007	.0002	.0001	.0000	.0000	.0000	.0000	.0000
	1	.0352	.0181	.0085	.0036	.0013	.0004	.0001	.0000	.0000	.0000
	2	.1445	.0885	.0498	.0253	.0113	.0042	.0012	.0002	.0000	.0000
	3	.3633	.2604	.1737	.1061	.0580	.0273	.0104	.0029	.0004	.0000
	4	.6367	.5230	.4059	.2936	.1941	.1138	.0563	.0214	.0050	.0004
	5	.8555	.7799	.6846	.5722	.4482	.3215	.2031	.1052	.0381	.0058
	6	.9648	.9368	.8936	.8309	.7447	.6329	.4967	.3428	.1869	.0572
	7	.9961	.9916	.9832	.9681	.9424	.8999	.8322	.7275	.5695	.3366
	8	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
9	0	.0020	.0008	.0003	.0001	.0000	.0000	.0000	.0000	.0000	.0000
	1	.0195	.0091	.0038	.0014	.0004	.0001	.0000	.0000	.0000	.0000
	2	.0898	.0498	.0250	.0112	.0043	.0013	.0003	.0000	.0000	.0000
	3	.2539	.1658	.0994	.0536	.0253	.0100	.0031	.0006	.0001	.0000
	4	.5000	.3786	.2666	.1717	.0988	.0489	.0196	.0056	.0009	.0000
	5	.7461	.6386	.5174	.3911	.2703	.1657	.0856	.0339	.0083	.0006
	6	.9102	.8505	.7682	.6627	.5372	.3993	.2618	.1409	.0530	.0084
	7	.9805	.9615	.9295	.8789	.8040	.6997	.5638	.4005	.2252	.0712
	8	.9980	.9954	.9899	.9793	.9596	.9249	.8658	.7684	.6126	.3698
	9	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
10	0	.0010	.0003	.0001	.0000	.0000	.0000	.0000	.0000	.0000	.0000
	1	.0107	.0045	.0017	.0005	.0001	.0000	.0000	.0000	.0000	.0000
	2	.0547	.0274	.0123	.0048	.0016	.0004	.0001	.0000	.0000	.0000
	3	.1719	.1020	.0548	.0260	.0106	.0035	.0009	.0001	.0000	.0000
	4	.3770	.2616	.1662	.0949	.0473	.0197	.0064	.0014	.0001	.0000
	5	.6230	.4956	.3669	.2485	.1503	.0781	.0328	.0099	.0016	.0001
	6	.8281	.7340	.6177	.4862	.3504	.2241	.1209	.0500	.0128	.0010
	7	.9453	.9004	.8327	.7384	.6172	.4744	.3222	.1798	.0702	.0115
	8	.9893	.9767	.9536	.9140	.8507	.7560	.6242	.4557	.2639	.0861
	9	.9990	.9975	.9940	.9865	.9718	.9437	.8926	.8031	.6513	.4013
	10	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
11	0	.0005	.0002	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000
	1	.0059	.0022	.0007	.0002	.0000	.0000	.0000	.0000	.0000	.0000
	2	.0327	.0148	.0059	.0020	.0006	.0001	.0000	.0000	.0000	.0000
	3	.1133	.0610	.0293	.0122	.0043	.0012	.0002	.0000	.0000	.0000
	4	.2744	.1738	.0994	.0501	.0216	.0076	.0020	.0003	.0000	.0000
	5	.5000	.3669	.2465	.1487	.0782	.0343	.0117	.0027	.0003	.0000
	6	.7256	.6029	.4672	.3317	.2103	.1146	.0504	.0159	.0028	.0001
	7	.8867	.8089	.7037	.5744	.4304	.2867	.1611	.0694	.0185	.0016
	8	.9673	.9348	.8811	.7999	.6873	.5448	.3826	.2212	.0896	.0152
	9	.9941	.9861	.9698	.9394	.8870	.8029	.6779	.5078	.3026	.1019
	10	.9995	.9986	.9964	.9912	.9802	.9578	.9141	.8327	.6862	.4312
	11	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000

Tabelle A.6: Kolmogorov-Smirnov-Verteilung.

Die Tabelle gibt Quantile der Statistiken K_n^+, K_n^- für den einseitigen Test an.

	α				
n	**0.1**	**0.05**	**0.025**	**0.01**	**0.005**
1	.9000	.9500	.9750	.9900	.9950
2	.6838	.7764	.8419	.9000	.9293
3	.5648	.6360	.7076	.7846	.8290
4	.4927	.5652	.6239	.6889	.7342
5	.4470	.5094	.5633	.6272	.6685
6	.4104	.4680	.5193	.5774	.6166
7	.3815	.4361	.4834	.5384	.5758
8	.3583	.4096	.4543	.5065	.5418
9	.3391	.3875	.4300	.4796	.5133
10	.3226	.3687	.4092	.4566	.4889
11	.3083	.3524	.3912	.4367	.4677
12	.2958	.3382	.3754	.4192	.4490
13	.2847	.3255	.3614	.4036	.4325
14	.2748	.3142	.3489	.3897	.4176
15	.2659	.3040	.3376	.3771	.4042
16	.2578	.2947	.3273	.3657	.3920
17	.2504	.2863	.3180	.3553	.3809
18	.2436	.2785	.3094	.3457	.3706
19	.2373	.2714	.3014	.3369	.3612
20	.2316	.2647	.2941	.3287	.3524
21	.2262	.2586	.2872	.3210	.3443
22	.2212	.2528	.2809	.3139	.3367
23	.2165	.2475	.2749	.3073	.3295
24	.2120	.2424	.2693	.3010	.3229
25	.2079	.2377	.2640	.2952	.3166
26	.2040	.2332	.2591	.2896	.3106
27	.2003	.2290	.2544	.2844	.3050
28	.1968	.2250	.2499	.2794	.2997
29	.1935	.2212	.2457	.2747	.2947
30	.1903	.2176	.2417	.2702	.2899
31	.1873	.2141	.2379	.2660	.2853
32	.1844	.2108	.2342	.2619	.2809
33	.1817	.2077	.2308	.2580	.2768
34	.1791	.2047	.2274	.2543	.2728
35	.1766	.2018	.2242	.2507	.2690
36	.1742	.1991	.2212	.2473	.2653
37	.1719	.1965	.2183	.2440	.2618
38	.1697	.1939	.2154	.2409	.2584
39	.1675	.1915	.2127	.2379	.2552
40	.1655	.1891	.2101	.2349	.2521
Approximation für $n > 40$	$\frac{1.07}{\sqrt{n}}$	$\frac{1.22}{\sqrt{n}}$	$\frac{1.36}{\sqrt{n}}$	$\frac{1.52}{\sqrt{n}}$	$\frac{1.63}{\sqrt{n}}$

Anhang B

Wichtige parametrische Tests bei Normalverteilung und nichtparametrische Tests

Tests auf μ $(\overline{X}-\text{Test})$

Gegeben n unabhängige Stichprobenvariablen $X_1, ..., X_n \sim N(\mu, \sigma^2)$.

(1) σ bekannt

H_0	H_1	Entscheidung gegen H_0, falls	kritische Werte
$\mu = \mu_0$	$\mu > \mu_0$	$\overline{X} > c_1$	$c_1 = \mu_0 + \frac{\sigma}{\sqrt{n}} z_{1-\alpha}$
$\mu = \mu_0$	$\mu < \mu_0$	$\overline{X} < c_2$	$c_2 = \mu_0 - \frac{\sigma}{\sqrt{n}} z_{1-\alpha}$
$\mu = \mu_0$	$\mu \neq \mu_0$	$\overline{X} < c_3$ oder $\overline{X} > c_4$	$c_3 = \mu_0 - \frac{\sigma}{\sqrt{n}} z_{1-\alpha/2}$ $c_4 = \mu_0 + \frac{\sigma}{\sqrt{n}} z_{1-\alpha/2}$

Dabei ist z_p definiert durch $\Phi(z_p) = p$.

(2) σ unbekannt $(t - Test)$

H_0	H_1	Entscheidung gegen H_0, falls	kritische Werte
$\mu = \mu_0$	$\mu > \mu_0$	$\overline{X} > c_1$	$c_1 = \mu_0 + \frac{S}{\sqrt{n}} t_{n-1;1-\alpha}$
$\mu = \mu_0$	$\mu < \mu_0$	$\overline{X} < c_2$	$c_2 = \mu_0 - \frac{S}{\sqrt{n}} t_{n-1;1-\alpha}$
$\mu = \mu_0$	$\mu \neq \mu_0$	$\overline{X} < c_3$ oder $\overline{X} > c_4$	$c_3 = \mu_0 - \frac{S}{\sqrt{n}} t_{n-1;1-\alpha/2}$ $c_4 = \mu_0 + \frac{S}{\sqrt{n}} t_{n-1;1-\alpha/2}$

$t_{n-1;p}$ ist das p-Quantil der t-Verteilung mit n-1 Freiheitsgraden und $S = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2}$.

Test auf σ^2 (χ^2-Test)

Gegeben: n unabhängige Stichprobenvariablen $X_1, ..., X_n \sim N(\mu, \sigma^2)$.

(1) μ bekannt

H_0	H_1	Entscheidung gegen H_0, falls	kritische Werte
$\sigma^2 = \sigma_0^2$	$\sigma^2 > \sigma_0^2$	$T > c_1$	$c_1 = \sigma_0^2 \chi^2_{n;1-\alpha}$
$\sigma^2 = \sigma_0^2$	$\sigma^2 < \sigma_0^2$	$T < c_2$	$c_2 = \sigma_0^2 \chi^2_{n;\alpha}$
$\sigma^2 = \sigma_0^2$	$\sigma^2 \neq \sigma_0^2$	$T < c_3$ oder $T > c_4$	$c_3 = \sigma_0^2 \chi^2_{n;\alpha/2}$ $c_4 = \sigma_0^2 \chi^2_{n;1-\alpha/2}$

$T = \sum_{i=1}^{n}(X_i - \mu)^2$. $\chi^2_{n;p}$ ist das p-Quantil der χ^2-Verteilung mit n Freiheitsgraden.

(2) μ unbekannt

H_0	H_1	Entscheidung gegen H_0, falls	kritische Werte
$\sigma^2 = \sigma_0^2$	$\sigma^2 > \sigma_0^2$	$T > c_1$	$c_1 = \sigma_0^2 \chi^2_{n-1;1-\alpha}$
$\sigma^2 = \sigma_0^2$	$\sigma^2 < \sigma_0^2$	$T < c_2$	$c_2 = \sigma_0^2 \chi^2_{n-1;\alpha}$
$\sigma^2 = \sigma_0^2$	$\sigma^2 \neq \sigma_0^2$	$T < c_3$ oder $T > c_4$	$c_3 = \sigma_0^2 \chi^2_{n-1;\alpha/2}$ $c_4 = \sigma_0^2 \chi^2_{n-1;1-\alpha/2}$

$T = \sum_{i=1}^{n}(X_i - \overline{X})^2$. $\chi^2_{n-1;p}$ ist das p-Quantil der χ^2-Verteilung mit $n-1$ Freiheitsgraden.

Test auf Gleichheit zweier Erwartungswerte (t-Test)

Gegeben: n_1 unabhängige Stichprobenvariablen $X_1, ..., X_{n_1} \sim N(\mu_X, \sigma_X^2)$ und n_2 unabhängige Stichprobenvariablen $Y_1, ..., Y_{n_2} \sim N(\mu_Y, \sigma_Y^2)$.

(1) unverbundene Stichproben

Die Variablen X_i und Y_j sind ebenfalls unabhängig, $\sigma_X = \sigma_Y$ (unbekannt).

H_0	H_1	Entscheidung gegen H_0, falls	kritische Werte
$\mu_X = \mu_Y$	$\mu_X > \mu_Y$	$T > c_1$	$c_1 = t_{n_1+n_2-2;1-\alpha}$
$\mu_X = \mu_Y$	$\mu_X < \mu_Y$	$T < c_2$	$c_2 = -t_{n_1+n_2-2;1-\alpha}$
$\mu_X = \mu_Y$	$\mu_X \neq \mu_Y$	$T < c_3$ oder $T > c_4$	$c_3 = -t_{n_1+n_2-2;1-\alpha/2}$ $c_4 = t_{n_1+n_2-2;1-\alpha/2}$

$$T = \frac{\overline{X} - \overline{Y}}{\sqrt{\frac{(n_1-1)S_X^2 + (n_2-1)S_Y^2}{n_1+n_2-2}\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}},$$

$t_{n_1+n_2-2;p}$ ist das p-Quantil der t-Verteilung mit $(n_1 + n_2 - 2)$ Freiheitsgraden.

(2) verbundene Stichproben

Die Variablen X_i und Y_i sind abhängig (paarweise an einem Merkmalsträger erhoben), $n_1 = n_2 = n$. Die Variablen $D_i = X_i - Y_i \quad (i = 1, ..., n)$ sind unabhängig.

H_0	H_1	Entscheidung gegen H_0, falls	kritische Werte
$\mu_X = \mu_Y$	$\mu_X > \mu_Y$	$T > c_1$	$c_1 = t_{n-1;1-\alpha}$
$\mu_X = \mu_Y$	$\mu_X < \mu_Y$	$T < c_2$	$c_2 = -t_{n-1;1-\alpha}$
$\mu_X = \mu_Y$	$\mu_X \neq \mu_Y$	$T < c_3$ oder $T > c_4$	$c_3 = -t_{n-1;1-\alpha/2}$ $c_4 = t_{n-1;1-\alpha/2}$

$T = \frac{\overline{D}}{S_D}\sqrt{n}$. Weiters ist $t_{n-1;p}$ das p-Quantil der t-Verteilung mit $n - 1$ Freiheitsgraden und

$$\overline{D} = \frac{1}{n}\sum_{i=1}^{n} D_i, \quad S_D = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(D_i - \overline{D})^2}.$$

Test auf Gleichheit zweier Varianzen (F-Test)

Gegeben: n_1 unabhängige Stichprobenvariablen $X_1, ..., X_{n_1} \sim N(\mu_X, \sigma_X^2)$ und n_2 unabhängige Stichprobenvariablen $Y_1, ..., Y_{n_2} \sim N(\mu_Y, \sigma_Y^2)$. Die Variablen X_i und Y_j sind ebenfalls unabhängig.

H_0	H_1	Entscheidung gegen H_0, falls	kritische Werte
$\sigma_X^2 = \sigma_Y^2$	$\sigma_X^2 > \sigma_Y^2$	$T > c_1$	$c_1 = F_{n_1-1,n_2-1;1-\alpha}$
$\sigma_X^2 = \sigma_Y^2$	$\sigma_X^2 < \sigma_Y$	$T < c_2$	$c_2 = F_{n_1-1,n_2-1;\alpha}$
$\sigma_X^2 = \sigma_Y^2$	$\sigma_X^2 \neq \sigma_Y^2$	$T < c_3$ oder $T > c_4$	$c_3 = F_{n_1-1,n_2-1;\alpha/2}$ $c_4 = F_{n_1-1,n_2-1;1-\alpha/2}$

$T = \frac{S_X^2}{S_Y^2}$. Weiters ist $F_{n_1-1,n_2-1;p}$ das p-Quantil der F-Verteilung mit $n_1 - 1$ und $n_2 - 1$ Freiheitsgraden.

Test auf Unkorreliertheit (Unabhängigkeit)

Gegeben: Paarige, unabhängige Stichprobenvariablen $(X_1, Y_1), ..., (X_n, Y_n) \sim N_2(\mu_x, \mu_y, \sigma_X^2, \sigma_Y^2, \rho)$.

H_0	H_1	Entscheidung gegen H_0, falls	kritische Werte
$\rho = 0$	$\rho \neq 0$	$\lvert R\rvert\sqrt{\frac{n-2}{1-R^2}} > c$	$c = t_{n-2;1-\alpha/2}$

$R = \frac{\sum_{i=1}^n (X_i - \overline{X})(Y_i - \overline{Y})}{\sqrt{\sum_{i=1}^n (X_i - \overline{X})^2 \sum_{i=1}^n (Y_i - \overline{Y})^2}}$.

$t_{n-2;p}$ ist das p-Quantil der t-Verteilung mit $n - 2$ Freiheitsgraden.

Wichtige nichtparametrische Tests

Anpassungstests

(1) χ^2-Test

Gegeben seien n unabhängige Stichprobenvariablen $X_1, ..., X_n$ und eine vollständig spezifizierte Verteilungsfunktion F.

H_0	H_1	Entscheidung gegen H_0, falls	kritische Werte
$F(x) = F_0(x)\ \forall x$	$F(x) \neq F_0(x)$ für mindestens ein x	$T > c$	$c = \chi^2_{k-1;1-\alpha}$

Dabei ist $T = \sum_{i=1}^{k}(h_i - e_i)^2/e_i$ mit k Klassen der Daten, h_i den absoluten Häufigkeiten und e_i den theoretischen, absoluten Häufigkeiten aus F_0. Ist F_0 nicht vollständig bestimmt, d.h. nur bis auf r Parameter, die aus den Daten geschätzt werden, dann sind die kritischen Werte $c = \chi^2_{k-r-1;1-\alpha}$.

(2) Kolmogorov-Smirnov-Test

Gegeben seien n unabhängige Stichprobenvariablen $X_1, ..., X_n$ mit stetiger Verteilungsfunktion F.

H_0	H_1	Entscheidung gegen H_0, falls	kritische Werte
$F(x) = F_0(x)$ $\forall x$	$F(x) > F_0(x)$ mindestens ein x	$sup(F_n(x) - F_0(x)) > c_1$	$c_1 = k^-_{1-\alpha}$
$F(x) = F_0(x)$ $\forall x$	$F(x) < F_0(x)$ mindestens ein x	$sup(F_0(x) - F_n(x)) > c_2$	$c_2 = k^+_{1-\alpha}$
$F(x) = F_0(x)$ $\forall x$	$F(x) \neq F_0(x)$ mindestens ein x	$sup\lvert F_n(x) - F_0(x)\rvert > c_3$	$c_3 = k_{1-\alpha}$

Test auf Quantil Q_p

(1) Binomialtest

Gegeben n unabhängige Stichprobenvariablen $X_1, ..., X_n$.

$$Y_i = \begin{cases} 1 & \text{falls } X_i \leq Q_p \\ 0 & \text{sonst} \end{cases}$$

$\Rightarrow$ falls $P(X_i \leq Q_p) = p$, $i = 1, ..., n$, ist $\sum Y_i \sim Bi(n, p)$.

H_0	H_1	Entscheidung gegen H_0, falls	kritische Werte
$p = p_0$	$p > p_0$	$\overline{Y} > c_1$	$c_1 = min_c\{\frac{c}{n} : P(B \leq c) \geq 1 - \alpha\}$
$p = p_0$	$p < p_0$	$\overline{Y} < c_2$	$c_2 = max\{\frac{c}{n} : P(B \geq c) \geq 1 - \alpha\}$
$p = p_0$	$p \neq p_0$	$\overline{Y} < c_3$ oder $\overline{Y} > c_4$	$c_3 = max\{\frac{c}{n} : P(B \geq c) \geq 1 - \alpha_1\}$ $c_4 = min\{\frac{c}{n} : P(B \leq c) \geq 1 - \alpha_2\}$ mit $\alpha = \alpha_1 + \alpha_2$

Dabei ist B binomialverteilt $Bi(n, p_0)$.

(2) Wilcoxons Vorzeichen-Rangtest für den Median.
Gegeben n unabhängige Stichprobenvariablen $X_1, ..., X_n$ mit symmetrischer Verteilung um Median m.

H_0	H_1	Entscheidung gegen H_0, falls	kritische Werte
$m = m_0$	$m > m_0$	$W^+ > c_1$	$c_1 = min\{c : P(W^+ \leq c) \geq 1 - \alpha\}$
$m = m_0$	$m < m_0$	$W^+ < c_2$	$c_2 = max\{c : P(W^+ \geq c) \geq 1 - \alpha\}$
$m = m_0$	$m \neq m_0$	$W^+ < c_3$ oder $W^+ > c_4$	$c_3 = max\{c : P(W^+ \geq c) \geq 1 - \alpha/2\}$ $c_4 = min\{c : P(W^+ \leq c) \geq 1 - \alpha/2\}$

Dabei ist $W^+ = \sum_{i=1}^{n} I_i \, r(|D_i|)$ mit $I_i = 1$, wenn $D_i > 0$, sonst 0, und $r(|D_i|)$ der Rang von $|D_i|$ mit $D_i = X_i - m_0$.

Test auf Zufälligkeit

Wald-Wolfowitz-Iterationstest.
Gegeben seien n unabhängige Stichprobenvariablen $X_1, ..., X_n$ mit nur zwei Merkmalsausprägungen.

H_0	H_1	Entscheidung gegen H_0, falls	kritische Werte
Reihenfolge zufällig	Reihenfolge nicht zufällig	$R > c_1$ oder $R < c_2$	$c_1 = min\{c : P(R \geq c) \geq 1 - \alpha/2\}$ $c_2 = max\{c : P(R \leq c) \geq 1 - \alpha/2\}$

Dabei ist R die Anzahl von Iterationen, d.h. geschlossenen Gruppen, denen ein anderes Element oder keines vorsteht, in der Folge der Beobachtungen.

Index