Duale Hochschule Baden-Württemberg Stuttgart

Vorlesungsskript

Statistik und Wahrscheinlichkeitstheorie

Dozent: Giles-Arnaud Nzouankeu Nana

Sommersemester

Inhaltsvezeichnis

Deskriptive Statistik 7

1. Einführung 7

1.1 Aufgaben der deskriptiven Statistik 7

1.2 Grundgesamtheit und Stichprobe 8

1.3 Merkmale und Skalenniveaus 9

2. Eindimensionale Häufigkeitsverteilungen 12

2.1 Häufigkeitsverteilungen bei diskreten Merkmalen 12

2.1.1 Absolute und relative Häufigkeitsverteilung 12

2.1.2 Graphische Darstellung 14

2.2 Häufigkeitsverteilungen bei stetigen Merkmalen 16

2.2.1 Prinzip der Klassenbildung 16

2.2.2 Histogramme 18

2.3 Empirische Verteilungsfunktion 19

2.4 Statistische Maßzahlen 21

2.4.1 Lagemaße 21

2.4.1.1 Modus 21

2.4.1.2 Median 22

2.4.1.3 Quantile 23

2.4.1.4 Arithmetisches Mittel 24

2.4.1.5 Vergleich der Lagemaße 25

2.4.2 Streuungsmaße 26

2.4.2.1 Quartilsabstand 27

2.4.2.2 Varianz und Standardabweichung 27

2.4.3 Formmaße 28

2.4.3.1 Schiefe 28

2.4.3.2 Wölbung 29

2.4.4 Box-Plots 29

3. Zweidimensionale Häufigkeitsverteilungen 31

3.1 Kontingenztabelle 31

3.2 Graphische Darstellung 32

3.2.1 Stabdiagramm und Säulendiagramm 32

3.2.2 Streudiagramme und zweidimensionale Histogramme 33

3.3 Korrelationsanalyse 34

3.3.1 Kovarianz ... 35
3.3.2 Korrelationskoeffizient ... 38
3.4 Regressionsanalyse ... 40
3.4.1 Schätzung der Regressionskoeffizienten ... 42
3.4.2 Prognose ... 43
3.4.3 Güte der Anpassung ... 44
Wahrscheinlichkeitstheorie ... 46
4. Das Rechnen mit Wahrscheinlichkeiten ... 46
4.1 Zufallsvorgänge und deren Beschreibung ... 46
4.1.1 Zufallsvorgang ... 46
4.1.2 Elementarereignis ... 46
4.1.3 Ergebnisraum ... 46
4.1.4 Ereignis ... 47
4.1.5 Sicheres Ereignis ... 48
4.1.6 Unmögliches Ereignis ... 48
4.2 Die Verknüpfung von Ereignissen ... 48
4.2.1 Vereinigung ... 48
4.2.2 Durchschnitt ... 48
4.2.3 Differenz ... 49
4.2.4 Komplement ... 49
4.2.5 Disjunkte Ereignisse ... 49
4.3 Die Axiome von Kolmogoroff ... 50
4.3.1 Axiome der Wahrscheinlichkeitsrechnung ... 50
4.3.2 Laplace-Experiment ... 51
4.4 Zufallsauswahl und Kombinatorik ... 51
4.4.1 Zufallsauswahl und Urnenmodell ... 51
4.4.2 Modell mit Zurücklegen ... 52
4.4.3 Permutationen ... 52
4.4.4 Kombinationen ... 52
4.5 Bedingte Wahrscheinlichkeiten und Unabhängigkeit von Ereignissen ... 53
4.5.1 Bedingte Wahrscheinlichkeit ... 53
4.5.2 Multiplikationssatz für unabhängige Ereignisse ... 54
4.6 Totale Wahrscheinlichkeit ... 55
5. Diskrete Zufallsvariable ... 56
5.1 Verteilung einer diskreten Zufallsvariable ... 56

5.1.1 Wahrscheinlichkeitsfunktion ... 56
5.1.2 Verteilungsfunktion ... 58
5.2 Unabhängigkeit von diskreten Zufallsvariablen ... 60
5.3 Parameter von diskreten Zufallsvariablen ... 61
5.3.1 Erwartungswert ... 61
5.3.2 Varianz ... 62
5.4 Spezielle diskrete Verteilunge ... 63
5.4.1 Die Binomialverteilung ... 63
5.4.2 Die Poisson-Verteilung ... 65
5.4.3 Die hypergeometrische Verteilung ... 66
6. Stetige Zufallsvariable ... 68
6.1 Definition und Verteilung ... 68
6.1.1 Dichtefunktion ... 68
6.1.2 Verteilungsfunktion ... 70
6.2 Unabhängigkeit von stetigen Zufallsvariablen ... 71
6.3 Parameter von stetigen Zufallsvariablen ... 71
6.3.1 Erwartungswert stetiger Zufallsvariablen ... 72
6.3.2 Varianz stetiger Zufallsvariablen ... 72
6.4 Die Normalverteilung ... 73
6.5 Die Exponentialverteilung ... 76
6.6 Sätze der Wahrscheinlichkeitsrechnung ... 78
6.6.1 Gesetz der großen Zahlen ... 78
6.6.2 Zentraler Grenzwertsatz ... 78
6.7 Prüfverteilungen ... 79
6.7.1 χ^2-Verteilung ... 79
6.7.2 t-Verteilung ... 80
6.7.3 F-Verteilung ... 81
7. Zweidimensionale Zufallsvariablen ... 82
7.1 Diskrete zweidimensionale Zufallsvariablen ... 82
7.1.1 Gemeinsame Wahrscheinlichkeitsfunktion ... 82
7.1.2 Bedingte Wahrscheinlichkeitsfunktion ... 83
7.1.3 Gemeinsame Verteilungsfunktion ... 83
7.2 Stetige zweidimensionale Zufallsvariablen ... 83
7.2.1 Gemeinsame Dichtefunktion ... 83
7.2.2 Randdichten ... 83

7.2.3 Gemeinsame Verteilungsfunktion 84
7.3 Eigenschaften zweidimensionaler Zufallsvariablen 84
7.3.1 Unabhängigkeit 84
7.3.2 Kovarianz 84
7.3.3 Kovarianz 85
7.3.4 Korrelationskoeffizient 86
8. Punktschätzung von Parametern 87
8.1 Der Begriff der Punktschätzung 87
8.2 Stichprobe und Schätzer 87
8.2.1 Eigenschaften von guten Schätzfunktionen 88
8.3 Spezielle Schätzfunktionen 89
8.3.1 Schätzen von Anteilswerten 89
8.3.2 Schätzen von Mittelwerten 90
8.3.3 Schätzen der Varianz 90
9. Intervallschätzung 91
9.1 Bedeutung und Definition des Konfidenzintervalls 91
9.2 Konfidenzintervalle für den Erwartungswert 91
9.2.1 Konfidenzintervall für $\boldsymbol{\mu}$ bei bekanntem $\boldsymbol{\sigma 2}$ 92
9.2.2 Konfidenzintervall für $\boldsymbol{\mu}$ bei unbekanntem $\boldsymbol{\sigma 2}$ 92
9.3 Konfidenzintervall für die Varianz 93
9.4 Konfidenzintervalle für eine Wahrscheinlichkeit 93
10. Statistischer Test 94
10.1 Der Binomial-Test und Gaußtest 94
10.1.1 Binomial-Test 94
10.1.2 Gauß-test 97
10.2 11.2 Fehlentscheidungen 98
10.3 Spezielle Testverfahren 99
10.3.1 **t**-Tests (Lagetests) 99
10.3.1.1 Einfacher **t**-Test 100
10.3.1.2 Doppelter **t**-Test 101
10.4 Testen von Anteilswerten 102
10.5 Unabhängigkeitstest 103
11. Quantile-Tabellen 105
Quantile der Standardnormalverteilung 106
Quantile der $\boldsymbol{\chi 2}$-Verteilung 107

Quantile der t −Verteilung 108
Quantile der F −Verteilung 109
Literatur 119

Deskriptive Statistik

1. Einführung

Die deskriptive Statistik stellt Methoden bereit, um Untersuchungsgegenstände, die aus einer Vielzahl von einzelnen Objekten bestehen, zu beschreiben. Hinter dieser abstrakten Beschreibung stehen vielfältige unterschiedliche Anwendungen. Beispielsweise soll das Wahlverhalten der deutschen Bürger beschrieben werden oder die Nebenwirkungen bei der Einnahme eines Medikamentes oder das Interesse der Konsumenten an einem neuen Produkt. Die drei genannten Beispiele haben gemeinsam, dass das konkrete Interesse auf eine Gesamtheit bezogen ist. Dies sind die wahlberechtigten Personen in Deutschland, die Patienten, die das Medikament einnehmen oder der Kundenkreis des neuen Produkts. Diese Gesamtheiten sollen durch die statistische Analyse quantitativ beschrieben werden. Dazu werden an den Einheiten - in den genannten Beispielen sind dies Personen - Untersuchungsmerkmale betrachtet. Der Schritt einer deskriptiven Auswertung sollte das Untersuchungsziel und auch die Gesamtheit genauer beschreiben. Wichtige Begriffe, die dabei nützlich sind, werden in diesem Kapitel behandelt.

1.1 Aufgaben der deskriptiven Statistik

Die deskriptive Statistik beschäftigt sich mit der Aufbereitung und Komprimierung von Daten, die in den verschiedensten Bereichen anfallen oder erhoben werden. Mithilfe der deskriptiven Statistik können Sachgebiete quantitativ beschrieben werden. Dabei werden viele Daten im Hinblick auf die Eigenschaften, die von Interesse sind, zusammengefasst. Zum Aufgabengebiet der deskriptiven Statistik zählen im Einzelnen:

- das Zusammenfassen und Ordnen der Daten in Tabellen,
- das Erstellen von Diagrammen und
- das Berechnen charakteristischer Kenngrößen oder Maßzahlen.

Beispiel:

- Der Absatz der verschiedenen Modelle eines Kfz-Herstellers soll graphisch dargestellt werden.
- Die Einkommensstruktur der Bundesbürger soll komprimiert dargestellt werden. Von Interesse ist hier der Anteil der Bevölkerung, die zu einer Einkommensklasse gehören, an der Gesamtbevölkerung, wobei das Einkommen in verschiedene Größenklassen aufgeteilt ist.

Deskriptive Statistiken werden in unterschiedlichsten Bereichen benötigt. So stellen etwa die staatlichen Statistikämter unter anderen Statistiken der Bevölkerungsstruktur, der Erwerbstätigkeit, des produzierenden Gewerbes, der Sozialleistungen, für Handel und Verkehr oder für das Gesundheitswesen zur Verfügung. Im Bereich Betriebsstatistik erfassen Unternehmen etwa Auftragseingänge, Produktion, Erzeugerpreise sowie Lohn- und Materialkosten. In der Forschung

werden mittels Statistik Forschungsgegenstände wie z.B. die Verträglichkeit von Medikamenten, das Wahlverhalten, die Belastungsfähigkeitunterschiedlicher Materialien, das Kaufverhalten der Geschlechter oder das Armutsrisiko von Patchworkfamilien beschrieben.

1.2 Grundgesamtheit und Stichprobe

Ausgangspunkt einer statistischen Fragestellung ist ein Untersuchungsgegenstand, der sich in der Regel auf viele einzelne Objekte bezieht. Dabei sollte der Untersuchungsgegenstand präzis formuliert und außerdem sachlich, räumlich und zeitlich abgegrenzt sein, damit eindeutig feststeht, auf welche konkreten Objekte er sich bezieht. Für die Vielzahl an möglichen Untersuchungsgegenständen sollen nun exemplarisch zwei Fälle illustriert werden.

Beispiel:

> Eine Firma überlegt, ein neues Produkt einzuführen. Sie möchte wissen, wie es auf dem Markt ankommen würde. Um das herauszufinden, führt das Unternehmen eine Umfrage durch. Untersuchungsgegenstand ist also die Akzeptanz eines neuen Produkts auf dem Markt. Dazu muss der potenzielle Absatzmarkt festgelegt werden. Welche geographische Größe ist relevant, ist der Markt national, europaweit oder global definiert? Sollen nur Ballungszentren bedient werden oder die ganze Fläche? Welches Geschlecht und welche Altersgruppen kommen infrage? Welche Einkommensklassen? Welche Vertriebswege sollen genutzt werden?

Beispiel

> Ein Kfz-Versicherer möchte die Schadensfälle des abgelaufenen Geschäftsjahres nach Schadenshöhe und geografischer Häufigkeit charakterisieren.
> Untersuchungsgegenstand ist die Gesamtheit der im ab gelaufenen Geschäftsjahr eingegangenen Schadensfälle. Somit müssen Beginn und Ende des Geschäftsjahres durch konkrete Zeitpunkte markiert werden. Sollen sämtliche Versicherungstarife analysiert werden oder gilt das Interesse nur einer Auswahl?

Wie an den Beispielen deutlich wird, bezieht sich der Untersuchungsgegenstand stets auf eine Menge von Objekten. Diese werden in der Statistik als Untersuchungseinheiten bezeichnet. Die Menge aller Untersuchungseinheiten bildet dann die Grundgesamtheit.

Untersuchungseinheit
Untersuchungseinheiten sind die Objekte, auf die sich die statistische Analyse bezieht.

Grundgesamtheit
Die Menge aller Untersuchungseinheiten wird als Grundgesamtheit oder Population bezeichnet.

Meist werden in der Praxis nicht alle Untersuchungseinheiten, auf die sich eine Fragestellung bezieht, analysiert. Dies wäre aus organisatorischen und zeitlichen Gründen oft viel zu aufwendig oder sogar vollkommen unmöglich. Häufig ist das auch gar nicht notwendig. Die moderne Statistik stellt nämlich Methoden zur Verfügung, die es ermöglichen, basierend auf einer relativ kleinen Auswahl von Untersuchungseinheiten allgemein gültige Aussagen bezüglich einer weitaus größeren Grundgesamtheit herzuleiten.

Stichprobe
Eine Stichprobe bezeichnet eine Auswahl an Untersuchungseinheiten einer Grundgesamtheit.

Die Vorgehensweise, Ergebnisse einer Teilgesamtheit auf eine übergeordnete Gesamtheit zu verallgemeinern, ist Aufgabe der induktiven Statistik. Die deskriptive Statistik dient ausschließlich nur zur Beschreibung einer vollständig bekannten Grundgesamtheit oder zur Beschreibung einer Stichprobe.

1.3 Merkmale und Skalenniveaus

Bei einer Untersuchung werden an den Untersuchungseinheiten Merkmale oder Variable betrachtet, die dem jeweiligen Interesse entsprechen. Merkmale werden mit Großbuchstaben X, Y, Z, ... notiert. Allgemein sind an einer statistischen Einheit in der Regel mehrere Merkmale beobachtbar.

Beispiel

Sind die Untersuchungseinheiten Personen, so könnten interessierende Merkmale sein:

X = das Monatseinkommen,
Y = die bei der letzten Bundestagswahl gewählte Partei,
Z = das Geschlecht.

Neben dem Begriff Merkmal wird häufig auch der Begriff Variable verwendet. Beide Begriffe werden synonym angewandt. Der Begriff Variable deutet schon darauf hin, dass die Variable oder das Merkmal für die verschiedenen Untersuchungseinheiten verschiedene Werte annehmen kann. Die Werte, die die Merkmale annehmen, werden als Merkmalsausprägungen bezeichnet. Die möglichen Ausprägungen eines Merkmals bilden den Merkmals- oder Zustandsraum.

Beispiel:

Im obigen Beispiel wurden Personen als Einheiten betrachtet. Die genannten Merkmale könnten die Werte
X = Monatseinkommen: 1 700 €, 2 100 €, 1 900 €, ...
Y = gewählte Partei: CDU, SPD, Grüne, FDP, ...
Z = Geschlecht: männlich, weiblich, neutral
annehmen.

Das Beispiel illustriert, dass die Art der Ausprägungen unterschiedlich sein kann. Im Folgenden sollen die verschiedenen möglichen Arten von Merkmalsausprägungen beschrieben werden. Die verwendeten Begriffe sind:

Qualitative Merkmale

Die Ausprägungen unterscheiden sich in ihrer Art. Sie können nicht durch eine Größenordnung beschrieben werden, aber in verschiedene Kategorien eingeteilt werden.

Quantitative Merkmale

Die Ausprägungen lassen sich durch Zahlen beschreiben. Sie besitzen eine Ausprägung, bei der die Größe interpretiert werden kann.

Oft werden die Kategorien von qualitativen Merkmalen im praktischen Gebrauch mit Zahlen belegt. So stehen z.B. Steuer- oder Kundennummern für verschiedene Personen oder Steuerklassen für unterschiedliche Personengruppen. Bankleitzahlen stehen für Banken, Postleitzahlen für Orte. Wichtig für das Verständnis des Begriffs des qualitativen Merkmals ist hierbei, dass die jeweiligen Zahlen in diesen Fällen nur der Unterscheidung dienen. Eine Anordnung der Merkmale im Sinne einer Größe ist hier nicht sinnvoll.

Beispiel:

- Qualitative Merkmale: Geschlecht, Familienstand, Konfession, Steuerklasse, Farbe eines Autos
- Quantitative Merkmale: Körpergröße, Alter, Einkommen, Umsatz, Temperatur.

Quantitative Merkmale lassen sich weiter unterteilen, je nachdem, ob die Merkmalsausprägungen nur diskrete Werte oder kontinuierliche Werte annehmen können.

Diskrete Merkmale
Die Merkmalsausprägungen nehmen nur bestimmte, separate Zahlenwerte an. Es werden nur endlich viele oder abzählbar unendlich viele Werte angenommen.

Stetige Merkmale
Es können Werte aus einem reellen Zahlenintervall angenommen werden. Die Werte sind kontinuierlich. Man stellt sich vor, dass sie auf beliebig viele Nachkommastellen messbar sind.

Bemerkung:

- Alle qualitativen Merkmale sind trivialerweise diskret. Quantitative Merkmale sind dann diskret, wenn die Merkmalsausprägungen durch einen Zählvorgang ermittelt werden können.
- Merkmale, die sich sehr fein unterteilen lassen, aber im Prinzip diskrete Merkmale sind, werden auch quasi-stetige Merkmale genannt und zu den stetigen Merkmalen gezählt.

Beispiel

- Diskrete Merkmale: Anzahl der Semester eines Studierenden, die Anzahl der Angestellten in einem Betrieb, Anzahl der Fehltage eines Arbeitnehmers.
- Stetige Merkmale: Körpergröße, Temperatur, Zeit, Benzinverbrauch.
- Quasi-stetige Merkmale: Einkommen eines Haushalts, Umsatz einer Firma,

Je nach Art des betrachteten Merkmals können die Merkmalsausprägungen anhand verschiedener Skalen gemessen werden. Im Folgenden sind die verschiedenen Skalenniveaus, geordnet nach dem Informationsgehalt, beginnend beim niedrigsten, beschrieben.

Skalenniveaus

- Nominales Skalenniveau:
  Bei einem nominalen Skalenniveau lassen sich die verschiedenen Ausprägungen des Merkmals lediglich unterscheiden. Es gibt keine natürliche Anordnung der Merkmals-ausprägungen. Man spricht von einem nominalen Merkmal.
- Ordinales Skalenniveau:
  Es gibt eine Rangfolge bzw. Ordnung innerhalb der Ausprägungen des Merkmals. Der Abstand zwischen den Ausprägungen ist aber nicht sinnvoll interpretierbar. Man spricht von

einem ordinalen Merkmal.

- Metrisches Skalenniveau:
  Die Ausprägungen des Merkmals lassen sich der Größe nach anordnen. Zudem sind die Abstände der Ausprägungen interpretierbar. Man spricht von einem quantitativen bzw. metrischen Merkmal.

Beispiel: Skalen

- Beispiele für Merkmale, die auf einer nominalen Skala gemessen werden, sind: Geschlecht, Familienstand, Konfession, Augenfarbe, Rechtsform eines Unternehmens, Branche eines Unternehmens.
- Beispiele für Merkmale, die auf einer ordinalen Skala gemessen werden, sind: Leistungsbeurteilung, Bewertung bei einem Schönheitswettbewerb, Zufriedenheit mit einem Produkt mit beispielsweise den Ausprägungen: sehr zufrieden, zufrieden, unzufrieden, sehr unzufrieden.
- Beispiele für Merkmale, die auf einer metrischen Skala gemessen werden, sind: Körpergröße, Wartezeit auf den Bus, Gewinn eines Unternehmens, Einkommen eines Arbeitnehmers.

Für die metrischen Merkmale sind weitere Unterteilungen möglich.

Arten metrischer Skalen

- Intervall skala:
  Nur die Differenzen zwischen Ausprägungen können interpretiert werden.
- Verhältnisskala:
  Das Merkmal besitzt einen Nullpunkt, der inhaltlich ausgezeichnet ist. Als Folge davon ergibt eine Aussage wie „die eine Ausprägung ist doppelt so hoch wie die andere“ Sinn. Allgemein sind die Verhältnisse der verschiedenen Ausprägungen interpretierbar.

Beispiel: Metrische Skalen

- Ein Merkmal, das auf einer Intervallskala gemessen wird, ist die Temperatur. Abstände sind hier interpretierbar. Eine Aussage wie „heute ist es 3 Grad wärmer als gestern“ ist sinnvoll. Bei null Grad Celsius handelt es sich nicht um einen natürlichen Nullpunkt. Angenommen heute sind 6°C und gestern 3 °C gemessen worden, so ist es nicht sinnvoll zu sagen, dass es heute doppelt so warm ist wie gestern.
- Ein Beispiel für ein Merkmal, das auf einer Verhältnisskala gemessen wird, ist dasEinkommen eines Arbeitnehmers. Hier bietet es sich an, Verhältnisse zu bilden. Aussagen wie „Peter verdient doppelt so viel wie Dieter“ oder „Petra verdient 15 % mehr als Susanne“ sind zulässig.

2. Eindimensionale Häufigkeitsverteilungen

Um sich einen Überblick bezüglich wesentlicher Eigenschaften eines Merkmals anzueignen, beginnt man mit der Häufigkeitsverteilung. Diese Verteilung beschreibt, wie häufig die einzelnen Merkmalsausprägungen in der Grundgesamtheit zu finden sind. In diesem Kapitel werden für diskrete und stetige Merkmale Häufigkeitsbegriffe definiert und graphische Darstellungen vorgestellt. Ferner werden Methoden erarbeitet, mit denen sich die charakteristischen Eigenschaften eines einzelnen Merkmals beschreiben lassen. Man unterscheidet bei diesen statistischen Kenngrößen oder Maßzahlen hierbei Lagemaße und Streuungsmaße.

2.1 Häufigkeitsverteilungen bei diskreten Merkmalen

2.1.1 Absolute und relative Häufigkeitsverteilung

Zu den diskreten Merkmalen können wir hier alle qualitativen sowie die quantitativ-diskreten Merkmale zählen. So gehören beispielsweise zum qualitativen Merkmal „Geschlecht“ die drei Ausprägungen „männlich“, „weiblich“ und „neutral“. Durch einfaches Abzählen lässt sich ermitteln, wie häufig die beiden Ausprägungen in der Grundgesamtheit vertreten sind.

Wir bezeichnen mit

- N den Umfang der Grundgesamtheit, bestehend aus N Untersuchungseinheiten,
- x_i die Merkmalsausprägung des Merkmals X, die bei der i-ten Untersuchungseinheit beobachtet wurde.

Nummeriert man die Untersuchungseinheiten fortlaufend von 1 bis *N* durch, dann enthält die Urliste die statistischen Daten dergestalt, dass jeder Untersuchungseinheit *i* die Merkmalsausprägung x_i zugeordnet ist.

Beispiel: Haushaltsgröße von Privathaushalten

Für *N = 40* Privathaushalte liegen für das Merkmal „Anzahl Personen“ folgende Daten vor:

3 5 2 3 3 2 3 5 3 5 5 3 2 1 4 3 4 4 2 4
3 3 1 4 5 3 2 4 1 4 1 2 2 3 1 3 1 4 2 2

Die Daten liegen in Form einer Urliste vor, die man auch in standardisierter Form mit 40 Zeilen und zwei Spalten notieren könnte, wobei die erste Spalte lediglich zur Numerierung der Untersuchungseinheiten (Objekte) dient. Einem Objekt entspricht ein Haushalt, der das Merkmal *X* = „Anzahl Personen“ besitzt. Beispielsweise besitzt der erste Haushalt (erstes Objekt) die Merkmalsausprägung $x_1 = 3$, der zweite Haushalt (zweites Objekt) die Merkmalsausprägung $x_2 = 5$, usw., wie man der ersten Zeile in der obigen Urliste entnimmt.

Um Fragen der Form „Wie viele Haushalte sind 4-Personen-Haushalte in dieser Grundgesamtheit?“ beantworten zu können, ordnet man den in der Grundgesamtheit vorkommenden unterschiedlichen Merkmalsausprägungen ihre absoluten Häufigkeiten zu, d.h. man ermittelt durch einfaches Abzählen,

wie viele 4-Personen-Haushalte es in der Grundgesamtheit gibt.

Allgemein formuliert man diesen Sachverhalt folgendermaßen:
Ein diskretes Merkmal X habe $M \leq N$ verschiedene Ausprägungen $x_1, \dots, x_M$. Die absolute Häufigkeit einer Ausprägung x_j wird mit f_j bezeichnet. Der Buchstabe j ist der sogenannte Laufindex, der zwischen 1 und M variiert. Die Summe aller absoluten Häufigkeiten f_j entspricht der Anzahl der Untersuchungseinheiten.

Absolute Häufigkeiten

$f(x_j) = f_j$	absolute Häufigkeit der Ausprägung x_j,
$f_1, \dots, f_M$	absolute Häufigkeitsverteilung,
$0 \leq f_j \leq N$	$j = 1, \dots, M$,
$f_1 + \cdots + f_M = \sum_{j=1}^{M} f_j = N.$	

Will man den Anteil der 4-Personen-Haushalte in der Grundgesamtheit ermitteln, so benötigt man die Anzahl N der Untersuchungseinheiten. Man spricht dann von der relativen Häufigkeit, mit der die Merkmalsausprägung in der Grundgesamtheit vorkommt.

Relative Häufigkeiten

$h(x_j) = h_j$	relative Häufigkeit der Ausprägung x_j,
$h_j = \frac{f_j}{N}$	realtive Häufigkeitsverteilung,
$f_1, \dots, f_M$	
$0 \leq h_j \leq 1$	$j = 1, \dots, M$,
$h_1 + \cdots + h_M = \sum_{j=1}^{M} h_j = 1.$	

In der Praxis gewinnt man die Häufigkeiten am einfachsten durch das Erstellen einer Strichliste oder - weniger mühsam - mittels einer geeigneten Statistiksoftware.

Beispiel Privathaushalte (Fortsetzung)

Für die Daten der 40 Privathaushalte aus dem obigen Beispiel ergeben sich folgende Häufigkei
ten:

Haushaltsgröße x_j	**absolute Häufigkeiten f_j**	**relative Häufigkeiten h_j**
1	6	6/40 = 0,150
2	9	9/40 = 0,225
3	12	12/40 = 0,300
4	8	8/40 = 0,200
5	5	5/40 = 0,125
Summe	**40**	**1**

Die Anzahl der 4-Personen-Haushalte unter den betrachteten 40 Privathaushalten beträgt also 8, bzw. 20 % der Haushalte sind 4-Personen-Haushalte.

Beispiel:

Die Wahl zum 17. Deutschen Bundestag fand am 27. September 2009 statt. Es ergab sich folgendes Ergebnis (gültige abgegebene Zweitstimmen):

Partei	Anzahl Zweitstimmen	Zwitstimmen in %
CDU/CSU	14 658 515	33,8
SPD	9 990 488	23,0
FDP	6 316 080	14,6
GRÜNE	3 977 125	10,7
DIE LINKE	5 155 933	11,9
PIRATEN	847 870	2,0
SONSTIGE	1 759 032	4,0
SUMME	43 371 190	100,0

Objekt = wahlberechtigter Bürger, der eine gültige Zweitstimme abgegeben hat
Grundgesamtheit = alle 43 371 190 wahlberechtigte Bürger, die eine gültige Zweitstimme abgegeben haben
Untersuchungsmerkmal (qualitativ, nominalskaliert). X = gewählte Partei
Merkmalsausprägungen: x_1 = CDU/CSU, x_2 = SPD, x_3 = FDP, x_4 = GRÜNE, x_5 = DIE LINKE, x_7 = = PIRATEN, x_7 = SONSTIGE.
Die zu den Merkmalsausprägungen $x_1, \dots, x_8$ gehörenden absoluten Häufigkeiten $f_1, \dots, f_8$ findet man in der Spalte Anzahl Zweitstimmen, die relativen Häufigkeiten (in Prozent) $h_1, \dots, h_8$ in der Spalte Zweitstimmen in %.

Bemerkung:
Die relative Häufigkeit wird oft in Prozentwerten angegeben.

2.1.2 Graphische Darstellung

Für die Darstellung der Häufigkeitsverteilungen $f_1, \dots f_M$ bzw. $h_1, \dots, h_M$ sind Tabellen oder Graphiken üblich. Da sich Menschen oft leichter von visuellen Eindrücken als von Zahlentabellen überzeugen lassen, werden insbesondere bei Präsentationen in der Praxis häufig graphische Darstellungen des Zahlenmaterials verwendet.
Die bekanntesten Darstellungsformen qualitativer oder quantitativ-diskreter Merkmale sind Stab-, Säulen-, Balken- und Kreisdiagramme. Darüber hinaus gibt es weitere optisch ansprechende Möglichkeiten z.B. Piktogramme.

Stabdiagramm

Die Ausprägungen $x_1, \dots, x_M$ des Merkmals X werden auf der horizontalen Merkmalsachse aufgetragen und darüber jeweils ein zur Merkmals-Achse senkrechter Strich (Stab) mit der Höhe $f_1, \dots f_M$ bzw. $h_1, \dots, h_M$.

Säulendiagramm

Modifikation des Stabdiagramms. Es werden die Stäbe durch Rechtecke ersetzt, die mittig über die Ausprägungen gezeichnet werden.

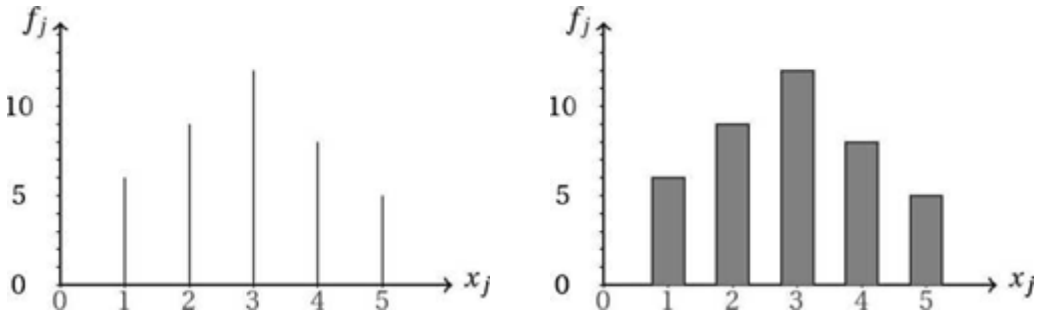


Abbildung 1: Stab- und Säulendiagramm

Balkendiagramm

Das Balkendiagramm ergibt sich als weitere Variante direkt aus dem Säulendiagramm, indem man die Ausprägungen auf der vertikalen Achse abträgt.

Kreisdiagramme eignen sich besonders zur Darstellung von Häufigkeiten qualitativer Merkmale, insbesondere von Häufigkeiten nominalskalierter Merkmale, da bei einem Kreisdiagramm keine Ordnung in den Daten darstellen werden kann. Die Aufteilung des Kreises in die einzelnen Sektoren, die die Merkmalsausprägungen repräsentieren, ist dabei proportional zu den relativen Häufigkeiten. Die Größe eines Kreissektors, also sein Winkel, kann damit aus den relativen Häufigkeiten h_j gemäß Winkel Kreissektor $j = h_j * 360^{\circ}$ berechnet werden.

Kreisdiagramm

Die Flächen der Kreissektoren im Kreisdiagramm sind proportional zu den Häufigkeiten. Für den Winkel α_j des j-ten Kreissektors gilt: $\alpha_j = h_j * 360^{\circ}$.

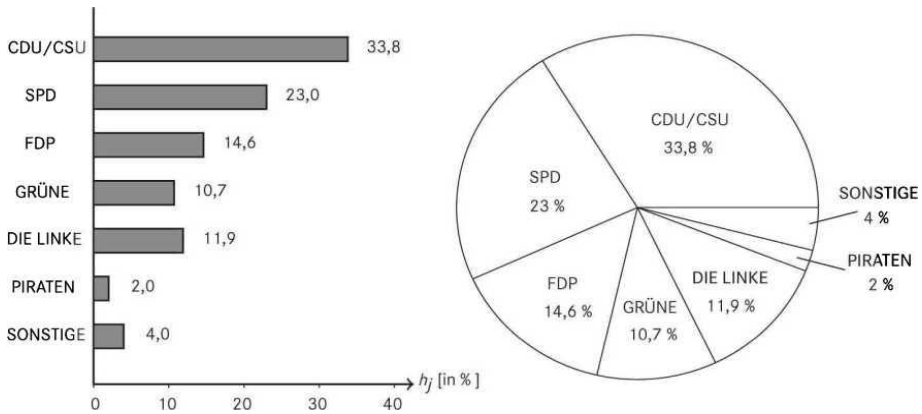


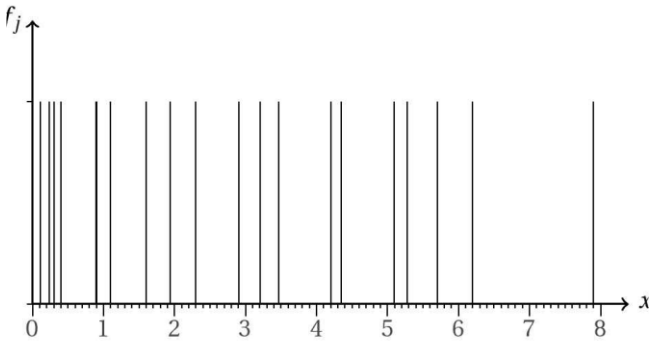
Abb. 1: Balken- und Kreisdiagramm für das Merkmal „gewählte Partei“ bei den gültigen Zweitstimmen der Bundestagswahl 2009

Beispiel: Bedienzeiten

Im Rahmen einer Kundenzufriedenheitsanalyse werden bei 20 Kunden die Bedienzeiten X[min] an einem Postschalter gemessen:

2.30	1.94	0.11	5.70	5.28	2.91	0.89	4.20	0.30	0.23
5.09	7.90	3.47	1.60	0.40	6.20	0.90	4.35	3.21	1.10

Einem Objekt entspricht ein Kunde, der das Merkmal X = „Bedienzeit“ besitzt. Da die Bedienzeit im Prinzip beliebig genau messbar ist, ist es keine Überraschung, dass keine zwei Kunden exakt gleich lange bedient werden. Die Merkmalsausprägungen sind also alle verschieden, sie treten alle mit der absoluten Häufigkeit $f_j = 1$ bzw. der relativen Häufigkeit $h_j = 1/20 = 0{,}05$ für $j = 1, \dots, N = 1, \dots, 20$ auf. Die Tabelle der absoluten Häufigkeiten stellt die Urliste nur in geordneter Form dar, die Darstellung als Stabdiagramm zeigt nur gleich hohe Stäbe:



Man erkennt zwar hier noch, wo die Daten dichter zusammenliegen, aber ansonsten ist das Bild wenig informativ.

Das Beispiel zeigt eine für stetige oder quasi-stetige Merkmale typische Situation auf. Es erweist sich in solchen Fällen als sinnvoll, mehrere nebeneinander liegende Ausprägungen zusammenzufassen. Dies besprechen wir im folgenden Abschnitt.

2.2 Häufigkeitsverteilungen bei stetigen Merkmalen

2.2.1 Prinzip der Klassenbildung

In vielen Fällen, insbesondere bei stetigen und quasi-stetigen Merkmalen, ist es oft nicht möglich, die N beobachteten Merkmalswerte der Urliste auf eine deutlich kleinere Menge von $x_1, \dots, x_M$ (unterschiedlichen) Werten zu komprimieren. Häufigkeitsverteilungen, die mit den Methoden des vorhergehenden Abschnitts erstellt werden, haben dann nur eine geringe Aussagekraft, da jede beobachtete Merkmalsausprägung die gleiche Häufigkeit besitzt.
Um eine interpretierbare und überschaubare Häufigkeitsverteilung zu erhalten, fasst man mehrere Merkmalsausprägungen zu einer Klasse zusammen, d.h. die Werte der Urliste werden M verschiedenen Klassen $K_1, K_2, \dots, K_M$ zugeordnet. Dabei sind unterschiedliche Klassenbreiten erlaubt. Das ursprüngliche Merkmal, wie etwa Zeit, Größe oder Gewicht, wird so zu einem diskreten Merkmal, das nur noch die M verschiedenen „Werte“ K_j annehmen kann. Der Preis, den man mit einer Diskretisierung bzw. Klassenbildung bezahlt, ist ein Informationsverlust, da die Verteilung der Werte innerhalb einer Klasse nicht mehr berücksichtigt wird.

Wir führen folgende Bezeichnungen ein:

M Anzahl der Klassen,
a_j untere Klassengrenze der Klasse K_j,
b_j obere Klassengrenze der Klasse K_j,
$b_j - a_j$ Klassenbreite der Klasse K_j.

Damit lassen sich die absoluten und relativen Häufigkeiten je Klasse wie bei den diskreten Merkmalen berechnen.

Absolute Häufigkeiten

f_j Anzahl der Beobachtungen x_j in der Klasse K_j
$a_j \leq x_j < b_j,\ j = 1, \dots, M$

Relative Häufigkeiten

$h_{j=} {}^{f_j}/_N$ Anteil der Beobachtungen x_j in der Klasse K_j
$a_j \leq x_j < b_j,\ j = 1, \dots, M$

Also werden alle Merkmalswerte, die größer oder gleich der Klassenuntergrenze a_j und kleiner als die Klassenobergrenze b_j sind, in der Klasse *j* gezählt. Die Häufigkeitstabelle enthält somit die Klassennummer K_j, für die wir kürzer einfach j schreiben, die Klassengrenze a_j und b_j und manchmal auch die Klassenbreite $b_j - a_j$, die absoluten Häufigkeiten f_j und die relativen Häufigkeiten h_j der Klassen. Dieser Aufbau ist in der folgenden Tabelle dargestellt:

Beispiel: Bedienzeiten (Fortsetzung)

Für die im Rahmen einer Kundenzufriedenheitsanalyse an 20 Kunden gemessenen Bedienzeiten *X* [min] aus dem obigen Beispiel ergibt sich die folgende klassierte Häufigkeitstabelle:

Klassennr j	**Klasse j**	**Klassenbreite $b_j - a_j$**	f_j	h_j
1	$0 \leq \cdots < 1$	$1 - 0 = 1$	6	${}^{6}/_{20} = 0{,}30$
2	$1 \leq \cdots < 2$	$2 - 1 = 1$	3	${}^{3}/_{20} = 0{,}15$
3	$2 \leq \cdots < 5$	$5 - 2 = 3$	6	${}^{6}/_{20} = 0{,}30$
4	$5 \leq \cdots < 8$	$8 - 5 = 3$	5	${}^{5}/_{20} = 0{,}25$
Summe			20	1,00

25% der Kunden hatten also z.B. eine Bedienzeit von 5 bis unter 8 Minuten. Es ist jedoch nicht mehr erkennbar, wie sich die Zeiten innerhalb einer Klasse verteilen

Für die Anzahl der Klassen und die Klassenbreite gibt es kaum feste Regeln. Bei sehr vielen schmalen Klassen ist die Darstellung unübersichtlich und die Struktur der Verteilung schwer erkennbar.

Dagegen ist eine geringe Anzahl von breiten Klassen mit einem hohen Informationsverlust verbunden und charakteristische Eigenschaften der Verteilung werden eventuell verdeckt. Am besten ist es, wenn es sachlogische Zusammenhänge gibt, die die Klassengrenzen definieren.

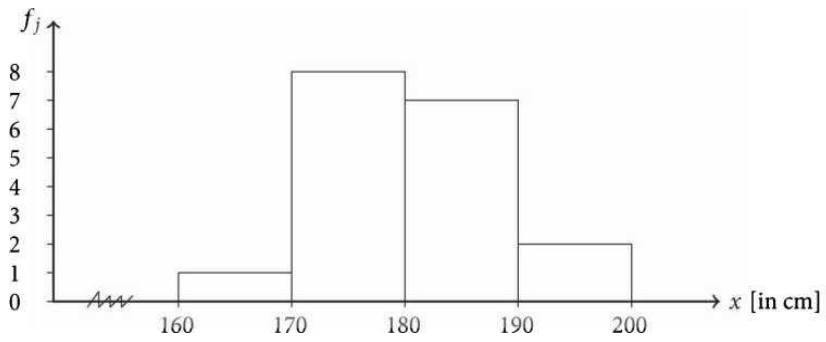
2.2.2 Histogramme

Die klassierte Häufigkeitstabelle stellt für stetige und quasi-stetige Merkmale eine erste Möglichkeit zur Veranschaulichung der Verteilung der Daten dar. Eine weitere Möglichkeit ist das Histogramm. Bei der Darstellung des Histogramms werden die Merkmalsausprägungen jeder Klasse durch ein Rechteck repräsentiert, dessen Fläche proportional zu der jeweiligen absoluten bzw. relativen Klassenhäufigkeit ist.
Für die Höhe des Rechtecks über der j-ten Klasse gilt dann:

$$f_j^* = \frac{f_j}{b_j - a_j} \quad bzw. \quad h_j^* = \frac{h_j}{b_j - a_j}$$

Beispiel: Körpergröße männlicher Studierender

Für das Merkmal Körpergröße der 18 männlichen Studierenden aus obigem Beispiel und der obigen Klasseneinteilung mit 4 Klassen jeweils der Breite 10 [cm] ergibt sich folgendes Histogramm:



■

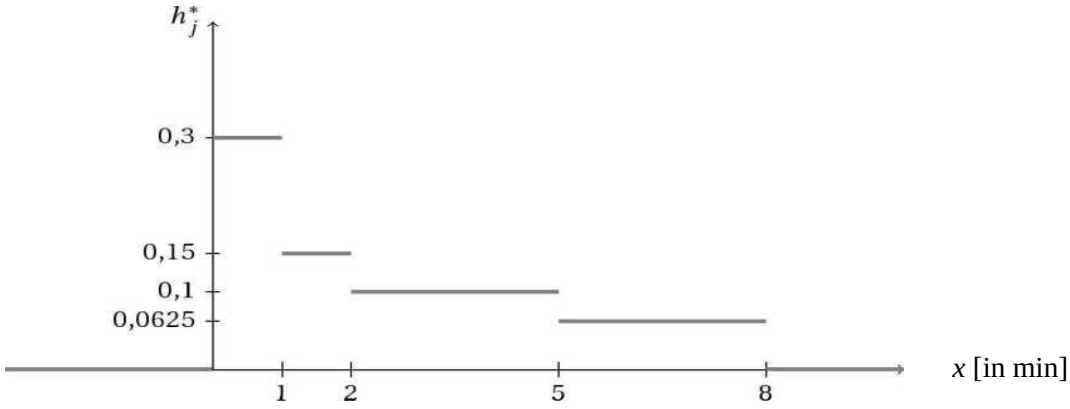
Die mathematische Funktion, die ein Histogramm beschreibt, bezeichnet man als empirische Dichte. Dabei ordnet man jedem x-Wert in eindeutiger Weise die jeweilige Höhe des darüberliegenden Rechtecks zu.

$$h^*(x) = \begin{cases} \frac{h_j}{b_j - a_j} & \text{für} \quad aj < x < bj \\ 0 & \text{sonst} \end{cases}$$

Dabei sind *aj*, *bj* die Klassengrenzen, *M* die Anzahl der Klassen. Die Gesamtfläche unter der Funktion *h** (x) bzw. die Fläche aller Rechtecke ergibt in der Summe immer den Wert eins.

Beispiel: Bedienzeiten (Fortsetzung)

Graph der Dichtefunktion für das Merkmal Bedienzeiten *X* [min] aus obigem Beispiel



Der Graph von h^* (x) zeigt einen stufigen Verlauf und vermittelt so den vereinfachenden Eindruck, dass sich die Werte innerhalb einer Klasse gleichmäßig verteilen. In der Wahrscheinlichkeitstheorie werden wir diese Vereinfachung aufheben und Dichten zulassen, die einen kurvigen, zusammenhängenden Verlauf aufweisen können.

2.3 Empirische Verteilungsfunktion

Bei quantitativen oder ordinalskalierten Merkmalen mag es sinnvoll sein, die Häufigkeiten beginnend bei der kleinsten Ausprägung in aufsteigender Reihenfolge aufzuaddieren. Dadurch erhält man die Anzahl der Daten, die eine bestimmte obere Grenze nicht überschreiten. Diese Häufigkeiten nennt man kumulierte oder Summenhäufigkeiten. Nimmt man an, dass die Ausprägungen des Merkmals *X* der Größe nach geordnet sind mit $x_1 \leq x_2 \leq \cdots \leq x_M$ mit $M \leq N$.

Die empirische Verteilungsfunktion an der Stelle *x* ist die kumulierte relative Häufigkeit aller Merkmalausprägungen x_j, die kleiner oder gleich *x* sind:

$$H(x) = \sum_{x_j \leq x} h(x_j)$$

bzw.

$$H(x) = \begin{cases} 0 & f\ddot{u}r\ x < x_1 \\ \sum_{x_k \leq x_j} h(x_k) & f\ddot{u}r\ x_j \leq x \leq x_{j+1}\ \ j = 1, \ldots, M-1 \\ 1 & f\ddot{u}r\ x \geq x_M \end{cases}$$

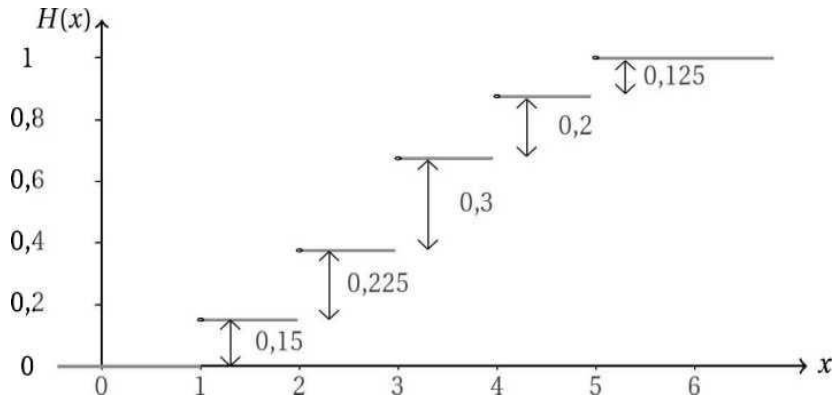
wobei *M* wieder die Anzahl der unterschiedlichen Merkmalsausprägungen bezeichnet.

Beispiel: Haushaltsgröße von Privathaushalten

Für die Daten der 40 Privathaushalte aus obigem Beispiel ergibt sich die folgende Verteilungsfunktion:

$$H(x) = \begin{cases} 0 & f\ddot{u}r\ x < 1 \\ 0{,}15 & f\ddot{u}r\ 1 \leq x < 2 \\ 0{,}375 & f\ddot{u}r\ 2 \leq x < 3 \\ 0{,}675 & f\ddot{u}r\ 3 \leq x < 4 \\ 0{,}875 & f\ddot{u}r\ 4 \leq x < 5 \\ 1 & f\ddot{u}r\ x > 5 \end{cases}$$

und als graphische Darstellung



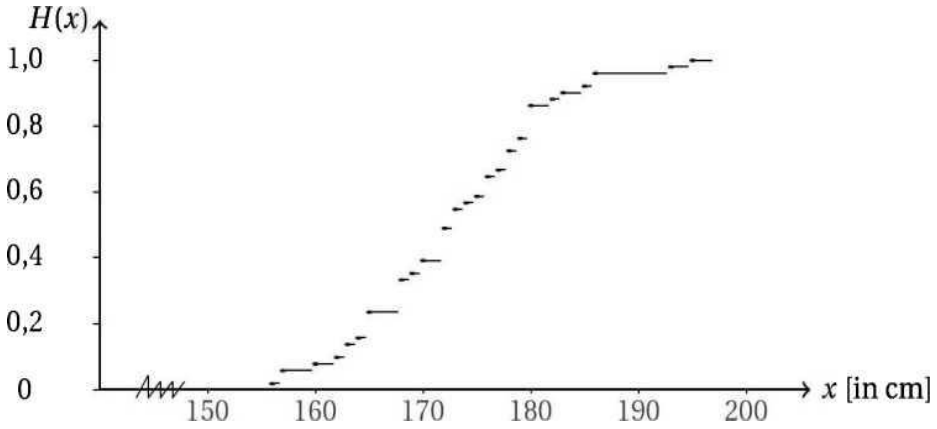
Eigenschaften der empirischen Verteilungsfunktion (HM)

- HM ist eine Treppenfunktion,
- HM = 0 für alle x, die kleiner als die Merkmalsausprägung x_{min} sind,
- HM wächst monoton ab x_{min} von 0 bis 1,
- $HM = 1$ ab dem größten Wert x_{max}.

Bei einem stetigen oder quasi-stetigen Merkmal mit vielen Ausprägungen kann die empirische Verteilung im Prinzip, wie vorher gezeigt, berechnet und dargestellt werden. Man erhält dann eine Treppenfunktion mit einer sehr großen Anzahl von Treppen und, wenn jeder Merkmaiswert nur einmal vorkommt, Stufen der Höhe *1/N*. Die Funktion nähert sich einer glatten Kurve.

Beispiel: Körpergröße der Studierenden

Für das Merkmal Körpergröße der 51 Studierenden aus obigem Beispiel erhält man folgenden Graphen der empirischen Verteilungsfunktion



2.4 Statistische Maßzahlen

In diesem Abschnitt werden Methoden vorgestellt, mit denen sich die charakteristischen Eigenschaften eines einzelnen Merkmals durch aussagekräftige statistische Kennzahlen oder Maßzahlen beschreiben lassen. Man unterscheidet hierbei Lagemaße, Streuungsmaße und Formmaße.

2.4.1 Lagemaße

Verteilungen eines Merkmals geben detaillierte Informationen, welche Merkmalswerte wie häufig in einer Grundgesamtheit anzutreffen sind. Lageparameter hingegen dienen dazu, die Eigenschaften einer Verteilung in komprimierter Form wiederzugeben, indem sie alle Merkmalswerte auf einen einzigen repräsentativen Wert reduzieren, der stellvertretend für alle Merkmalswerte steht. Insbesondere sind Lagemaße beim Vergleich mehrerer Grundgesamtheiten beliebt.

2.4.1.1 Modus

Dieses Lagemaß gibt an, welche Merkmalsausprägung am häufigsten vorkommt. Bei einer stetigen oder klassifizierten Variablen ist der Modus die Klasse, in der die Werte am dichtesten liegen, also die Dichte den größten Wert annimmt.

X ist ein diskretes Merkmal:

x_{mod} = häufigster Wert des Merkmals X,

X ist ein stetiges bzw. klassiertes Merkmal mit der Dichte h_j^*:

x_{mod} = Klasse K_j mit größter Häufigkeitsdichte h_j^*.

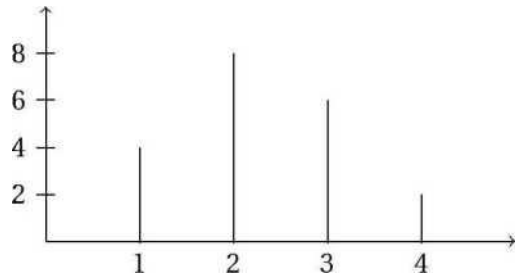
Bemerkung:
Wenn man bei einem klassierten Merkmal nur einen einzelnen Wert und nicht die ganze Klasse als Modus angeben möchte, wählt man in der Regel die Klassenmitte stellvertretend für die ganze Klasse.

Der Modus ist das wichtigste Lagemaß für kategoriale Merkmale und bereits auf Nominalskalenniveau sinnvoll. In der Darstellung durch Stab- oder Säulendiagramme ist der Modus die Merkmalsausprägung mit dem höchsten Stab bzw. mit der höchsten Säule. Bei stetigen Merkmalen ist der Modus die Klasse mit der größten Dichte.

Beispiel:

Eine Kleinstadt beabsichtigt, ihren Marktplatz neu zu gestalten. Dazu wird eine Umfrage unter 20 Marktständen durchgeführt. Unter anderem wird nach der Zahl X der Beschäftigten gefragt. Das Ergebnis ist in der folgenden Tabelle bzw. in dem folgenden Diagramm zusammengefasst.

Beschäftigte x_j	$f(x_j)$
1	4
2	8
3	6
4	2
Gesamt	20



Für das Merkmal *X* ist x_{mod} = 2, d.h. Verkaufsstände mit 2 Beschäftigten kommen unter den 20 befragten Marktständen am häufigsten vor.

2.4.1.2 Median

Der Median oder Zentralwert ist ein Merkmalswert, der die Grundgesamtheit in zwei Hälften teilt, wobei in der einen Hälfte die Objekte mit den größeren Merkmalswerten, und in der anderen Hälfte die kleineren Merkmalswerte liegen. Um diese Maßzahl zu ermitteln, sind die Beobachtungswerte der Größe nach zu sortieren. Die geordneten Werte werden mit tief gestellten, in eckigen Klammern gesetzten Indizes versehen, sodass gilt:

$$x_{[1]} \leq x_{[2]} \leq \cdots \leq x_{[N]}.$$

Demnach ist $x_{[1]}$ der kleinste Wert der Beobachtungsreihe; $x_{[N]}$ ist der größte Wert. Die sortierten Beobachtungswerte nennt man auch Rangliste. Das dazugehörige Merkmal muss mindestens ordinal skaliert sein.

Beispiel:

Fünf Angestellte einer Firma vergleichen ihr Bruttomonatseinkommen. Folgende Einkommen wurden festgestellt:

Angestellter Nr. *i*	1	2	3	4	5
Einkommen x_t in €	3 400	3 800	4100	3 700	3 200

Für die geordnete Reihe der Merkmalswerte ergibt sich:

$x_{[1]}$	$x_{[2]}$	$x_{[3]}$	$x_{[4]}$	$x_{[5]}$
3200	3 400	3700	3 800	4100

Die Grundgesamtheit umfasst 5 Objekte, d.h. eine ungerade Zahl von Objekten. Eine Aufteilung in zwei gleich große Hälften zu jeweils exakt 50% ist nicht möglich. Der Merkmalswert 3700 € mit der Ordnungsnummer 3, der in der „Mitte“ steht, kommt der Idee des Medians am nächsten. Wir stellen fest, dass 60% der Merkmalswerte kleiner oder gleich $x_{[3]}$ = 3700 € und 60% der Merkmalswerte größer oder gleich $x_{[3]}$ = 3700 € sind. Daher ist $x_{[3]}$ der Median.

Beispiel

Zu den fünf Angestellten aus dem obigen Beispiel kommt ein weiterer Angestellter hinzu. Er verdient 3300€. Die Grundgesamtheit umfasst jetzt 6 Objekte, d.h. eine gerade Zahl von Objekten. Für die geordnete Reihe der Merkmalswerte erhält man:

$x_{[1]}$	$x_{[2]}$	$x_{[3]}$	$x_{[4]}$	$x_{[5]}$	$x_{[6]}$
3200	3300	3400	3700	3800	4100

Zwar ist eine Aufteilung in zwei gleich große Hälften zu jeweils exakt 50% möglich, jedoch gibt es diesmal keinen Merkmalswert, der genau in der „Mitte“ steht. Für jeden Wert x aus dem Intervall [3400, 3700], d.h. für jedes x mit $x_{[3]} \leq x \leq x_{[4]}$ gilt, dass
50% der Merkmalswerte kleiner oder gleich x und 50% der Merkmalswerte größer oder gleich x sind. Dies zeigt, dass der Median in bestimmten Fällen nicht eindeutig ist. Es ist üblich, in solchen Fällen die Mitte des Intervalls als Median zu nehmen.

Berechnung des Medians bei vorliegender Urliste

Sortiere die Urliste nach aufsteigenden Merkmalswerten: $x_{[1]} \leq x_{[2]} \leq \cdots \leq x_{[N]}$,

dann erhält man:

$$x_{med} = \begin{cases} x_{\left[\frac{N+1}{2}\right]}, & \text{für } N \text{ ungerade,} \\ \frac{1}{2} * (x_{\left[\frac{N}{2}\right]} + x_{\left[\frac{N+1}{2}\right]}), & \text{für } N \text{ gerade.} \end{cases}$$

2.4.1.3 Quantile

Der Median versucht eine Grundgesamtheit möglichst gut in zwei gleich große Hälften zu je 50% aller Objekte aufzuteilen. Bei einem *p*-Prozent-Quantil verhält es sich ähnlich, jedoch können diesmal die beiden Teile der Grundgesamtheit auch unterschiedlich groß sein. Sei *p* eine Zahl zwischen Null und Eins und $x_{[1]} \leq x_{[2]} \leq \cdots \leq x_{[N]}$ die der Größe nach geordnete Urliste. Wir definieren ähnlich wie beim Median: Das *p*-Prozent-Quantil x_p ist dadurch charakterisiert, dass mindestens *p**100% der Merkmalsausprägungen einen Wert kleiner oder gleich dem Wert x_p und mindestens (1 - p) * 100 % einen Wert größer oder gleich dem Wert x_p annehmen.

$$x_p = \begin{cases} x_{[<Np>]}, & \text{wenn } Np \text{ nicht ganzzahlig,} \\ \frac{1}{2} * (x_{[Np]} + x_{[Np+1]}), & \text{wenn } Np \text{ ganzzahlig.} \end{cases}$$

Das Symbol < *Np* > bedeute „nächstgrößere ganze Zahl an *Np*“.

Spezielle Quantile sind das

- Untere oder erste Quartil $x_{0,25}$. Dieses besagt, dass mindestens 25% der Merkmalswerte kleiner oder gleich $x_{0,25}$ sind, während dementsprechend mindestens 75 % der Werte größer oder gleich $x_{0,25}$ sind.

- Obere oder dritte Quartil $x_{0,75}$. Dieses besagt, dass mindestens 75 % der Merkmalswerte kleiner oder gleich $x_{0,75}$ sind, während dementsprechend mindestens 25 % der Werte größer oder gleich $x_{0,75}$ sind.
- Mittlere oder zweites Quartil $x_{0,50}$. Dieses entpricht dem Median x_{med}

Beispiel:

Nachfolgende Tabelle zeigt die der Größe nach geordnete Urliste der Körpergröße der männlichen Studierenden aus obigem Beispiel:

$x_{[1]}$	$x_{[2]}$	$x_{[3]}$	$x_{[4]}$	$x_{[5]}$	$x_{[6]}$	$x_{[7]}$	$x_{[8]}$	$x_{[9]}$
165	172	175	176	177	178	178	179	179

$x_{[10]}$	$x_{[11]}$	$x_{[12]}$	$x_{[13]}$	$x_{[14]}$	$x_{[15]}$	$x_{[16]}$	$x_{[17]}$	$x_{[18]}$
180	180	182	183	185	186	186	193	196

Wir bestimmen mithilfe obiger Rangliste einige Quantile bezüglich der Körpergröße nach der obigen Formel:

1.Quartil: $Np = 0{,}25 * 18 = 4{,}5$
$x_{0,25} = x_{[<4,5>]} = x_{[5]} = 177cm$

3. Quartil: $Np = 0{,}75 * 18 = 13{,}5$
$x_{0,75} = x_{[<13,5>]} = x_{[14]} = 185cm$

9. Dezil: $Np = 0{,}9 * 18 = 16{,}2$
$x_{0,9} = x_{[<16,2>]} = x_{[17]} = 193cm$

Daraus folgt, dass ein 175 cm großer Student bezüglich seiner Körpergröße im unteren Viertel liegt, während ein 196 cm großer Student den oberen 10 % angehört.

Liegen die Daten nur klassiert vor, so erfolgt die Bestimmung des p-Prozent-Quantils analog zur Bestimmung des Medians.

2.4.1.4 Arithmetisches Mittel

Das am häufigsten benutzte Lagemaß der Verteilung eines quantitativen Merkmals ist das arithmetische Mittel, das umgangssprachlich oft einfach als Mittelwert bezeichnet wird. Zu seiner Berechnung werden alle Merkmalswerte addiert und deren Summe, der gesamte Merkmalsbetrag, durch die Zahl der Merkmalswerte *N* dividiert. Jeder Wert x, geht mit dem gleichen Gewicht 1/N in die Berechnung ein. Das arithmetische Mittel darf nur dann berechnet werden, wenn die Summe bzw. die Differenz zwischen zwei Ausprägungen definiert ist. Dies setzt quantitative Merkmale voraus.

Wir bezeichnen mit

N die Anzahl der Elemente der Grundgesamtheit bzw. die Zahl der Merkmalswerte
x_i die Merkmalsausprägung des i-ten Elements, $i = 1, \dots, N$

Dann ist

$$\bar{x} = \frac{x_1 + x_2 + \cdots + x_N}{N} = \frac{1}{N}\sum_{i=1}^{N} x_i$$

Beispiel:

Für die Einkommen der fünf Angestellten einer Firma aus dem obigen Beispiel erhält man:

$$\bar{x} = \frac{3400 + 3800 + 4100 + 3700 + 3200}{5} = \frac{18200}{5} = 3640€.$$

Die fünf Angestellten verdienen also zusammen 18 200 €, das Durchschnittseinkom¬men beträgt 3 640 €

Falls die Daten bereits in Form einer Häufigkeitstabelle vorliegen, vereinfacht sich die Berechnung von $\bar{x}$ und die obige Formel lässt sich mit den Bezeichnungen:

M die Zahl der unterschiedlichen Merkmalsausprägungen
x_j die j-te unterschiedliche Merkmalsausprägung, $j = 1, \ldots, M$
f_j die der Merkmalsausprägung x_j zugeordnete absolute Häufigkeit, j = 1,..., M
h_j die der Merkmalsausprägung x_j zugeordnete relative Häufigkeit, j = 1,..., M

in folgender Form schreiben

$$\bar{x} = \frac{x_1 * f_1 + x_2 * f_2 + \cdots + x_N * f_N}{N} = \frac{1}{N}\sum_{i=1}^{N} x_i * f_i = \sum_{i=1}^{N} x_i * h_i$$

Beispiel:

Im obigen Beispiel mit den Verkaufsständen egibt sich als arithmetischer Mittelwert:

$$\bar{x} = \frac{1 * 4 + 2 * 8 + 3 * 6 + 4 * 2}{4} = \frac{46}{20} = 2{,}3.$$

Wir stellen fest, dass bei den 20 Marktständen insgesamt 46 Personen beschäftigt sind und die durchschnittliche Beschäftigtenzahl 2,3 beträgt, d. h. 2 Vollzeitbeschäftigte und eine Teilzeitkraft mit 30 % der Regelarbeitszeit.

2.4.1.5 Vergleich der Lagemaße

Die Lagemaße, die vorgestellt wurden, werden auf sehr unterschiedliche Weise ermittelt, woraus unterschiedliche Eigenschaften folgen. Der Ersteller einer Statistik ist also gezwungen, darüber zu entscheiden, welches Lagemaß für seine Problemstellung jeweils am sinnvollsten ist.

Modalwerte werden hauptsächlich angegeben:

- bei nominalen Merkmalen, da andere Lagemaße bei diesem Skalenniveau nicht zulässig sind,
- bei ordinalen und quantitativen Merkmalen, wenn es sich um einen ausgeprägten Gipfel handelt.

Die Angabe des Medians ist sinnvoll:

- bei ordinal skalierten Daten,
- bei quantitativen Merkmalen, die schief verteilt sind,
- bei Verdacht auf Extremwerte (Ausreißer),
- wenn der Median und alle sonstigen Quantile sich über die empirische Verteilungsfunktion beschreiben und graphisch abschätzen lassen.

Das arithmetische Mittel

- darf nur für quantitative Merkmale (nicht ordinalskalierte) berechnet werden,
- ist vor allem bei symmetrischen, eingipfligen Verteilungen sinnvoll,
- nutzt im Gegensatz zu anderen Lagemaßen alle Daten.

Beispiel: Arithmetisches Mittel und Ausreißer

Wir betrachten acht Angestellte einer Firma mit folgenden Bruttomonatseinkommen in€:

Angestellter Nr. i	1	2	3	4	5	6	7	8
Einkommen x_i	3 200	3 400	3 600	3 700	3 800	4100	4 700	25 000

Das Durchschnittseinkommen der ersten sieben Angestellten beträgt 3785,71 €, der Median liegt bei 3 700 €. Durch den achten Angestellten, mit seinem extrem hohen Einkommen von 25000 €, steigt das durchschnittliche Einkommen auf 6437,50 €, was ein „falsches“ Bild der Einkommensverteilung liefert. Der Median dagegen repräsentiert mit 3750 € nach wie vor das Zentrum der Verteilung, er ist gegenüber dem Ausreißer unempfindlich.

Wir erkennen, dass sowohl bei Ausreißern als auch durch die Asymmetrie oder Schiefe einer Verteilung die Interpretation von Lagemaßen erschwert werden kann. Es empfiehlt sich also die Verwendung weiterer Parameter zur Beschreibung einer Verteilung.

2.4.2 Streuungsmaße

Die bisher behandelten Lageparameter dienten lediglich der Kennzeichnung des Zentrums einer Verteilung. Oft ist jedoch von Interesse, wie stark die Einzelwerte vom Zentrum abweichen, d.h. wie eng oder weit sie um das Zentrum der Verteilung streuen. Zur Gewinnung dieser Information definiert man Streuungsparameter. Wir setzen dabei stets ein metrisches Skalenniveau voraus.

Die Streuungsparameter, die wir im Folgenden behandeln, lassen sich grob in zwei Gruppen einteilen. Zur ersten zählt der Quartilsabstand. Bei diesen Maßen dienen die Abstände zwischen speziellen Beobachtungen der Verteilung als Maß für die Streuung. Zur zweiten Gruppe zählt die Varianz. Hier werden die Abweichungen aller Merkmalswerte von einem Lagemaß für die Beurteilung der Streuung herangezogen. Je kleiner diese Kennzahlen sind, desto stärker sind die Merkmalswerte um den jeweiligen Lageparameter konzentriert.

2.4.2.1 Quartilsabstand

Der Quartilsabstand ist die Länge des Interquartilbereichs $[x_{0,25};\ x_{0,75}]$

$$D_Q = x_{0,75} - x_{0,25}.$$

Beispiel:

Für die Körpergrößen der Studierenden aus obigem Beispiel erhalten wir:

$$D_Q = x_{0,75} - x_{0,25} = x_{[39]} - x_{[13]} = 179 - 168 = 11\ cm.$$

2.4.2.2 Varianz und Standardabweichung

Die bekannteste Maßzahl für die Streuung einer Verteilung ist die Standardabweichung bzw. ihr Quadrat, die Varianz. Die Varianz lässt sicht folgendermaßen berechnen

$$s^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})^2 = \frac{1}{N}\sum_{i=1}^{N}x_i^2 - \left(\frac{1}{N}\sum_{i=1}^{N}x_i\right)^2 = \overline{x^2} - \bar{x}^2$$

Die Standardabweichung s ist die Wurzel aus der Varianz:

$$s = \sqrt{s^2}.$$

Beispiel:

Ein Zug A und ein Zug B haben jeweils 4 Fahrten unternommen. Bei jeder Fahrt wurde das Merkmal „Fahrgastaufkommen *X* [Pers./Fahrt]“ erhoben. Die Merkmalswerte sind durch folgende zwei Urlisten gegeben:

ZugA: 200, 250, 350, 400
Zug B: 50, 100, 500, 550

In beiden Grundgesamtheiten beträgt das arithmetische Mittel jeweils $\bar{x} = 300$, d.h. jeder Zug hat im Schnitt das gleiche Fahrgastaufkommen pro Fahrt. Während aber beim Zug A alle Fahrten fast gleich viele, nämlich ungefähr 300 Passagiere aufweisen, ist das Fahrgastaufkommen bei Zug B deutlich unterschiedlicher. Wie wird dieser Sachverhalt durch die Standardabweichung beschrieben?
Wir berechnen für jeden Zug jeweils die Varianz der Variablen *X* [Pers./Fahrt]:

ZugA:

$s^2 = ((200 - 300)^2 + (250 - 300)^2 + (350 - 300)^2 + (400 - 300)^2)/4$
$= (100^2 + 50^2 + 50^2 + 100^2)/4$
$= 6250$ [Pers.^{2}/Fahrt^2]

ZugB:

$s^2 = ((50 - 300)^2 + (100 - 300)^2 + (500 - 300)^2 + (550 - 300)^2)/4$
$= (250^2 + 200^2 + 200^2 + 250^2)/4$
$= 51250$ [Pers.^{2}/Fahrt^2]

Die Einheit der Varianz [Pers.^{2}/Fahrt^2] ist das Quadrat der ursprünglichen Einheit zu *X*. Da die Standardabweichung die Wurzel der Varianz ist, ergibt sich:

ZugA: $s = \sqrt{6250} = 79{,}06$ [Pers./Fahrt]
ZugB: $s = \sqrt{51250} = 226{,}27$ [Pers./Fahrt]

Während bei Zug A die Passagierzahlen nur eine Standardabweichung von 79,09 Personen pro Fahrt aufweisen, beträgt dieser Wert bei Zug B 226,27 Personen pro Fahrt.

Falls die Daten bereits in Form einer Häufigkeitstabelle vorliegen, dann vereinfacht sich die Berechnung von s^2 in die folgende Form:

$$s^2 = \frac{1}{N}\sum_{i=1}^{M}(x_i - \bar{x})^2 * f_i = \sum_{i=1}^{M}(x_i - \bar{x})^2 * h_i$$

Eigenschaften Varianz:

- Die Varianz s^2 ist stets größer oder gleich 0. Nimmt s^2 den Wert Null an, so liegt überhaupt keine Streuung vor, d.h. alle Merkmalswerte sind identisch.
- Es gilt der sogenannte Verschiebesatz:

$$s^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})^2 = \frac{1}{N}\sum_{i=1}^{N}x_i^2 - \left(\frac{1}{N}\sum_{i=1}^{N}x_i\right)^2 = \overline{x^2} - \bar{x}^2$$

  Er ist vor allem zur schnellen Berechnung von s^2 geeignet: Man bildet die Summe der quadrierten Werte x^2, mittelt diese und zieht das bereits vorher berechnete quadrierte arithmetische Mittel $\bar{x}^2$ ab.

- Werden die Merkmalswerte x_i in der Form $y_i = a + b * x_i$ mit a, b reelle Zahlen, dann gilt

$$s_y^2 = b^2 * s_x^2 \quad bzw. \quad s_y = |b| * s_x.$$

2.4.3 Formmaße

Einige statistische Methoden setzen eine bestimmte Verteilungsform voraus. Einen ersten Eindruck diesbezüglich liefert die graphische Darstellung der Verteilung. Sie lässt erkennen, ob eine Verteilung einen oder mehrere Gipfel hat, ob sie symmetrisch, linkssteil oder rechtssteil, ob sie schwach oder stark gewölbt ist. Neben den Lagemaßen und Streuungsmaßen dienen die Formmaße dazu, die Verteilungsform weiter quantitativ zu beschreiben.

2.4.3.1 Schiefe

Die Schiefe ist ein Formmaß, das bei eingipfligen Verteilungen die Symmetrie bzw. Asymmetrie kennzeichnet. Sie ist definiert als:

$$g_1 = \frac{\frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})^3}{s^3}.$$

Die Schiefe ist dimensionslos und kann sowohl positive als auch negative Werte annehmen. Falls sich positive und negative Abweichungen der Werte vom Mittelwert ausgleichen, ergibt sich für die Schiefe der Wert 0.

Für eingipflige Verteilungen gilt:

$g_1 = 0$ für symmetrische Verteilungen,
$g_1 < 0$ für linkssteile Verteilungen,
$g_1 > 0$ für rechtssteile Verteilungen.

2.4.3.2 Wölbung

Die Wölbung (auch Exzess genannt) beschreibt die Massenanhäufungen an den Enden bzw. um den Mittelwert der Verteilung. Sie gibt an, ob bei (gleicher Varianz) das absolute Maximum der Häufigkeitsverteilung größer oder kleiner ist als das der Normalverteilung. Die Normalverteilung ist eine wichtige symmetrische Verteilung, die in der Wahrscheinlichkeitstheorie und induktiven Statistik ausführlich behandelt wird. Die Wölbung wird folgendermaßen definiert:

$$g_2 = \frac{\frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})^4}{s^4} - 3.$$

Für eingipflige Verteilungen gilt:

$g_2 = 0$ bei Normalverteilung,
$g_2 < 0$ bei spitzeren Verteilungen,
$g_2 > 0$ bei flacheren Verteilungen.

2.4.4 Box-Plots

Bei der deskriptiven Analyse von Daten, insbesondere von größeren Datenmengen, bedient man sich neben der Berechnung von Maßzahlen häufig graphischer Methoden. Sie sollen einen Eindruck vom Verhalten der Daten wie Konzentration, Ausdehnung oder Symmetrie vermitteln. Neben anderen Darstellungen hat sich in der Praxis der sogenannte Box-Plot als diagnostisches Instrument bewährt. Die Box-Plots fassen die in einem Datenbestand enthaltenen Informationen mithilfe von 5 Zahlen zusammen, der sogenannten Fünf-Punkte-Zusammenfassung einer Verteilung:

$$x_{min}, \quad x_{0,25}, \quad x_{med}, \quad x_{0,75}, \quad x_{max}$$

und stellen sie graphisch dar.

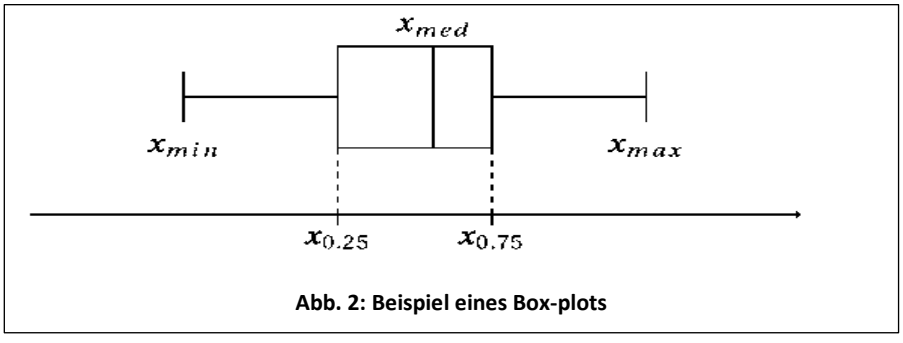


Abb. 2: Beispiel eines Box-plots

Das Zentrum von mit Box-Plots dargestellten Verteilungen ist durch den Median x_{med} gegeben. Die beiden Werte x_{min} und x_{max} informieren über den Datenausdehnungsbereich. Die Box zeigt den zentralen 50%-Anteil der Daten, die Länge der Box ist der Quartilsabstand $D_Q = x_{0,75} - x_{0,25}$.

Bei der Interpretation von Box-Plots ist zu berücksichtigen, dass die Länge der waagerechten Striche von der Box zu den beiden Werten x_{min} und x_{max} durch wenige Ausreißer stark beeinflusst werden kann, was zu einer starken Streckung der Graphik führen kann. Deswegen werden Ausreißer meist gesondert behandelt. Dazu werden sogenannte Ausreißerzäune definiert, die zur Identifikation von Ausreißern dienen sollen. Die Grenzen des inneren Zauns lauten:

$$\left[x_{0,25} - 1{,}5 * D_Q, x_{0,75} + 1{,}5 * D_Q\right],$$

die des äußeren Zauns:

$$\left[x_{0,25} - 3 * D_Q, x_{0,75} + 3 * D_Q\right].$$

Die Werte außerhalb der Box werden wie folgt behandelt:

- Werte, die außerhalb des äußeren Zauns liegen, d.h. mehr als 3 Box-Längen vom linken bzw. rechten Rand der Box entfernt sind, werden Extremwerte genannt und durch einen „x“ und den Zahlenwert wiedergegeben.
- Werte, die zwischen $x_{0,25} - 3 * D_Q$ und $x_{0,25} - 1{,}5 * D_Q$ bzw. $x_{0,75} + 1{,}5 * D_Q$ und $x_{0,75} + 3 * D_Q$ liegen, d. h. zwischen 1,5 und 3 Boxlängen vom linken bzw. rechten Rand der Box entfernt sind, werden Ausreißer genannt und durch einen „o“ wiedergegeben.

Der kleinste und größte beobachtete Wert x_u und x_o, die nicht als Ausreißer eingestuft werden, werden in den Box-Plot eingetragen und durch senkrechte Striche markiert.

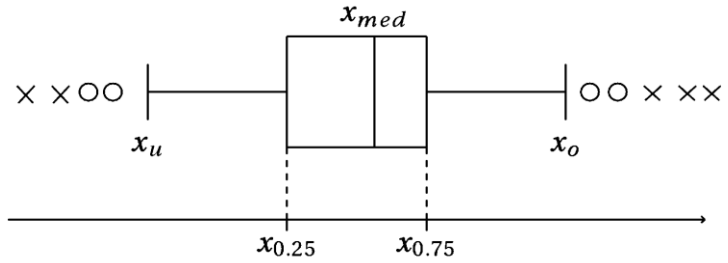


Abb. 3: Modifizierter Box-Plot.

3. Zweidimensionale Häufigkeitsverteilungen

In vielen statistischen Anwendungen ist man nicht nur an einem einzigen, sondern an vielen Merkmalen interessiert, die an den Untersuchungseinheiten gleichzeitig erhoben werden. Durch die gleichzeitige Betrachtung zweier Merkmale, ergeben sich neue statistische Charakteristika. Von besonderem Interesse ist dabei, ob ein Zusammenhang zwischen zwei Merkmalen besteht.

3.1 Kontingenztabelle

In diesem Abschnitt werden Methoden zur Darstellung der gemeinsamen Verteilung von zwei Merkmalen dargestellt, die nur wenige Ausprägungen aufweisen bzw. in nur wenigen Kategorien auftreten. Dabei kann es sich um qualitative Merkmale wie Geschlecht oder Familienstand, oder um quantitative Merkmale wie die Anzahl von Beschäftigten in Kleinbetrieben handeln.

Beispiel: Rauchen und Geschlecht

In einer Untersuchung zu den Rauchgewohnheiten von Studierenden wurden die Merkmale „Rauchen“ und „Geschlecht“ u. a. bei den Studierenden des obigen Kapitels erhoben. Es ergibt sich folgende Tabelle:

	Raucher	Nichtraucher	Summe
männlich	4	14	18
weiblich	6	27	33
Summe	10	41	51

Aus der Tabelle geht hervor, dass sich die Menge der 51 Studierenden aus 10 Rauchern und 41 Nichtrauchern zusammensetzt bzw. aus 18 Männern und 33 Frauen. Weiter erkennt man, dass 22% (4/18 = 0,22) der Männer rauchen und 18% (6/33 = 0,18) der Frauen, und dass die Raucher zu 40% (4/10 = 0,4) männlich und zu 60% (6/10 = 0,6) weiblich sind. Man interessiert sich dafür, ob zwischen den Merkmalen „Rauchen“ und „Geschlecht“ ein Zusammenhang besteht.

Die Zusammenfassung der Daten zweier Merkmale in Tabellenform wie im obigen Beispiel wollen wir nun verallgemeinern. Ausgangspunkt sind zwei Merkmale *X* und *Y* mit den unterschiedlichen Ausprägungen

$$x_1, .., x_j, \dots, x_M \quad j = 1, \dots, M \text{ und } y_1, .., y_k, \dots, y_L \quad k = 1, \dots, L.$$

Bei der gleichzeitigen Betrachtung beider Merkmale können $M \bullet L$ verschiedene Kombinationen der M unterschiedlichen Ausprägungen für das Merkmal *X* und der L unterschiedlichen Ausprägungen für das Merkmal *Y* gebildet werden. In völliger Analogie zur eindimensionalen Häufigkeitstabelle bildet man nun die absoluten Häufigkeiten

$$f_{jk} = f(x_j, y_k)$$

und die relativen Häufigkeiten

$$h_{jk} = \frac{f(x_j, y_k)}{M * L}$$

für jede Merkmalkombiantion (x_j, y_k).

	y_1	$\cdots$	y_k	$\cdots$	y_L	$f_{j\cdot}$
x_1	f_{11}	$\cdots$	f_{1k}	$\cdots$	f_{1L}	$f_{1\cdot}$
$\vdots$	$\vdots$		$\vdots$		$\vdots$	$\vdots$
x_j	f_{j1}	$\cdots$	f_{jk}	$\cdots$	f_{jL}	$f_{j\cdot}$
$\vdots$	$\vdots$		$\vdots$		$\vdots$	$\vdots$
x_M	f_{M1}	$\cdots$	f_{Mk}	$\cdots$	f_{ML}	$f_{M\cdot}$
$f_{\cdot k}$	$f_{\cdot 1}$	$\cdots$	$f_{\cdot k}$	$\cdots$	$f_{\cdot L}$	N

	y_1	$\cdots$	y_k	$\cdots$	y_L	$h_{j\cdot}$
x_1	h_{11}	$\cdots$	h_{1k}	$\cdots$	h_{1L}	$h_{1\cdot}$
$\vdots$	$\vdots$		$\vdots$		$\vdots$	$\vdots$
x_j	h_{j1}	$\cdots$	h_{jk}	$\cdots$	h_{jL}	$h_{j\cdot}$
$\vdots$	$\vdots$		$\vdots$		$\vdots$	$\vdots$
x_M	h_{M1}	$\cdots$	h_{Mk}	$\cdots$	h_{ML}	$h_{M\cdot}$
$h_{\cdot k}$	$h_{\cdot 1}$	$\cdots$	$h_{\cdot k}$	$\cdots$	$h_{\cdot L}$	1

Tabelle 3.1 Zweidimensionale Häufigkeitstabellen

Zweidimensionale Häufigkeitstabellen werden in der Regel durch die Zeilen und Spaltensummen ergänzt.

Die Zeilensummen ergeben die Randhäufigkeiten des Merkmals X und werden bezeichnet durch:

$$f_{j\cdot} = f_{j1} + \cdots + f_{jL}, \qquad j = 1, \ldots, M$$

Die Randhäufigkeiten $f_{j\cdot}$ sind die eindimensionalen Häufigkeiten, mit denen das Merkmal X die Merkmalsausprägungen x_j annimmt, wenn das Merkmal Y nicht berücksichtigt wird.

Die Spaltensummen ergeben entsprechend die Randhäufigkeiten des Merkmals *Y:*

$$f_{\cdot k} = f_{1k} + \cdots + f_{Mk}, \qquad k = 1, \ldots, L$$

Die Randhäufigkeiten $f_{\cdot k}$ von Y sind also die eindimensionalen Häufigkeiten, mit denen das Merkmal Y die Merkmalsausprägungen y_k annimmt, wenn das Merkmal X nicht berücksichtigt wird.

3.2 Graphische Darstellung

Die Kontingenztabellen enthalten zwar genaue Informationen bezüglich der Häufigkeiten, sie sind jedoch weniger geeignet um den Grad eines Zusammenhangs zu erfassen. Zu diesem Zweck bedient man sich u.a. graphischer Darstellungen.

3.2.1 Stabdiagramm und Säulendiagramm

Die Zusammenhänge zweier qualitativer Merkmale lassen sich mittels eines Säulendiagram darstellen. Die Längen der Säulen repräsentieren die Häufigkeiten der Ausprägungen des ersten Merkmals. Außerdem ist jede Säule entsprechend der Ausprägungen des zweiten Merkmals unterteilt. Eine andere Möglichkeit besteht darin, für jede Merkmalskombination eine dreidimensionale Säule oder einen dreidimensionalen Stab zu erstellen, der die jeweilige Häufigkeit f_{jk} *bzw* h_{jk} repräsentiert, und die $M \bullet L$ Stäbe bzw. Säulen in räumlicher Perspektive anzuordnen.

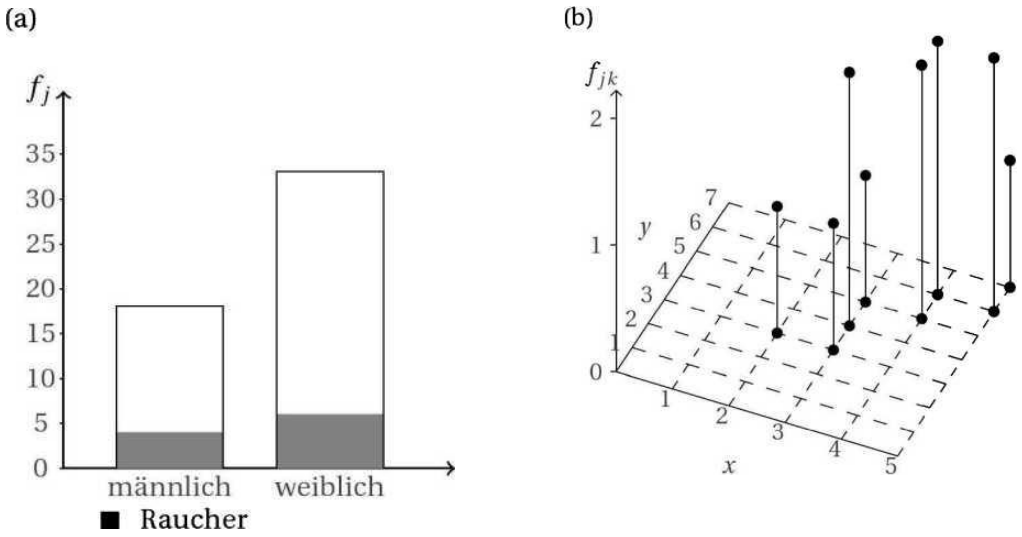


Abb. 4: (a) Säulendiagramm, (b) dreidimensionales Stabdiagramm

3.2.2 Streudiagramme und zweidimensionale Histogramme

Für die Darstellung quantitativer Merkmale mit vielen Ausprägungen empfiehlt es sich die bisherigen Methoden, die auf qualitative Merkmale abgestellt sind, durch zusätzliche Methoden zu ergänzen. Da in der Kontingenztafelanalyse nur das nominale Skalenniveau benutzt wird, werden quantitative Merkmale als qualitativ behandelt. Im Folgenden wird explizit ein metrisches Skalenniveau vorausgesetzt.

Die einfachste Darstellung der beobachteten Merkmalsausprägungen $(x_i, y_i), i = 1, \dots, N$ zweier quantitativer Merkmale ist das Streudiagramm, in dem die Messwerte in einem *x*-*y*- Koordinatensystem als Punkte, Kreuze oder sonstige Symbole dargestellt werden.

Streudiagramm

Die Darstellung der Beobachtungswerte $(x_1, y_1), (x_2, y_2), \dots, (x_N, y_N)$ im *x*-*y*-Koordinatensystem heißt Streudiagramm.

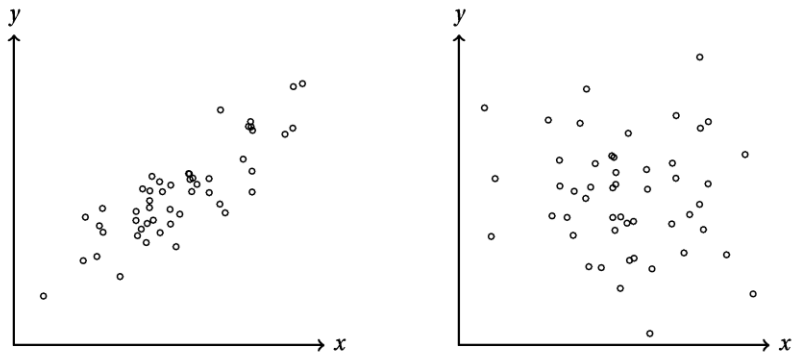


Abb. 5: Beispile von Streudiagrammen

Zweidimensionales Histogramm

Die Veranschaulichung der Häufigkeiten f_{jk} bzw h_{jk} im j-ten Intervall des Merkmals X und im k-ten Intervall von Y heißt zweidimensionales Histogramm. Dazu werden Quader über den Rechtecksklassen

$$[a_i, b_i) \times [c_k, d_k), j = 1, \dots, M, k = 1, \dots, L,$$

mit der Höhe

$$f_{jk}^* = \frac{f_{jk}}{(b_j - a_j)(d_k - c_k)} \quad bzw \quad h_{jk}^* = \frac{h_{jk}}{(b_j - a_j)(d_k - c_k)}.$$

errichtet.

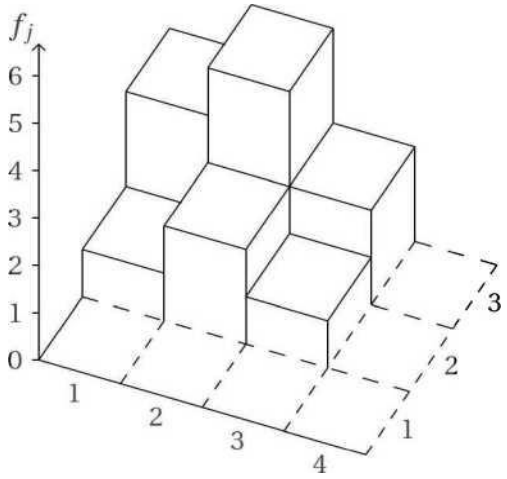


Abb. 6: Zweidimensionales Histogramm

3.3 Korrelationsanalyse

Mithilfe der Korrelation wird die lineare Abhängigkeit zweier metrischer Merkmale bestimmt. Genauer handelt es sich bei einem linearen Zusammenhang um einen Zusammenhang, der durch eine lineare Funktion dargestellt werden kann:

$$y = a + bx.$$

In den statistischen Anwendungen liegt meist kein exakter linearer Zusammenhang vor. Die Datenpunkte liegen dann in der Nähe einer fiktiven Gerade. So kann der Korrelationskoeffizient beispielsweise angewandt werden, um folgende beispielsweise Annahmen zu prüfen:

- Die Werbeausgaben steigern den Gewinn eines bestimmten Unternehmens.
- Je höher das Einkommen ist, desto höher sind die Konsumausgaben.
- Je größer die Wohnfläche ist, desto höher ist die Miete der Wohnung.

3.3.1 Kovarianz

Um später den Korrelationskoeffizienten nach Bravais-Pearson einführen zu können, benötigt man die empirische Kovarianz. Die Kovarianz ist in der Statistik eine grundlegende Größe für die Untersuchung des Zusammenhangs zweier metrischer Merkmale
Für zwei Merkmale *X* und *Y* liege eine Urliste $(x_i, y_i), i = 1, \ldots, N$ vor. Zudem seien $\bar{x}$ und $\bar{y}$ die jeweiligen arithmetischen Mittel.

Kovarianz

Die Kovarianz ist durch die folgende Formel gegeben:

$$cov(x,y) = \frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x}) * (y_i - \bar{y})$$

Im Falle einer Häufigkeitstabelle erhalten wir für die Kovarianz:

$$cov(x,y) = \frac{1}{N}\sum_{j=1}^{M}\sum_{k=1}^{L}(x_j - \bar{x}) * (y_k - \bar{y}) * f_{jk}.$$

Um zu verstehen, was die Kovarianz ausdrückt, wollen wir die Berechnung der Größe genauer betrachten. Das untenstehende Bild enthält ein typisches Streudiagramm einer zweidimensionalen Häufigkeitsverteilung. Das Streudiagramm wurde durch den Schwerpunkt (x, y) des Datensatzes ergänzt. Es entstehen so in dem Diagramm 4 Quadranten I, II, III und IV. Betrachten wir jeweils einen Punkt (x_i, y_i) in den verschiedenen Quadranten.

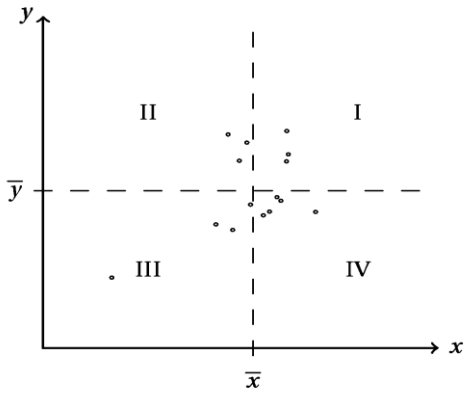


Abb. 7: Interpretation der Kovarianz

Für die Abweichungsprodukte $(x_i - \bar{x}) * (y_i - \bar{y})$, die in ihrer Summe der wesentliche Bestandteil der Kovarianz sind, gilt Folgendes: Liegt ein Beobachtungspunkt (x_i, y_i) in

Quadrant I so gilt $x_i > \bar{x},\ y_i > \bar{y}$ und damit $(x_i - \bar{x}) * (y_i - \bar{y}) > 0$,
Quadrant II so gilt $x_i < \bar{x},\ y_i > \bar{y}$ und damit $(x_i - \bar{x}) * (y_i - \bar{y}) < 0$,
Quadrant III so gilt $x_i < \bar{x},\ y_i < \bar{y}$ und damit $(x_i - \bar{x}) * (y_i - \bar{y}) > 0$,
Quadrant IV so gilt $x_i > \bar{x},\ y_i < \bar{y}$ und damit $(x_i - \bar{x}) * (y_i - \bar{y}) < 0$.

Daraus ergibt sich, dass alle Beobachtungswerte aus dem ersten und dritten Quadranten einen positiven, Beobachtungswerte aus dem zweiten und vierten Quadranten einen negativen Beitrag zur Kovarianz liefern. Bei Punktewolken, die vor allem im ersten und dritten Quadranten liegen, wird demnach die

Kovarianz $cov(x,y) > 0$. Bei derartigen Punktewolken finden sich bei großen x-Werten gehäuft große *y*-Werte, bei kleinen x-Werten aber gehäuft kleine *y*-Werte. Es liegt ein sog. positiver oder gleichsinniger Zusammenhang vor.

Punktewolken, die vorwiegend im zweiten und vierten Quadranten liegen, führen entsprechend zu einer negativen Kovarianz $cov(x,y) < 0$. Kleine x-Werte gehen hier mit großen *y*-Werten einher und umgekehrt. Es besteht ein sog. negativer oder gegensinniger Zusammenhang.

Sind alle Quadranten einigermaßen gleich stark besetzt, so heben sich negative und positive Beiträge gegenseitig auf, und man erhält eine Kovarianz in der Nähe von null, d.h. $cov(x,y) \approx 0$.

Eigenschaften der Kovarianz:

- Vertauscht man bei der Kovarianz die Variablen *X* und *Y*, so ändert sich der Wert der Kovarianz nicht. Die Kovarianz ist also symmetrisch in x und *y*, d.h. es gilt:

$$cov(x,y) = cov(y,x).$$

- Setzt man speziell $X = Y$, so berechnet man die Kovarianz der Variablen *X* mit sich selbst. In diesem Fall ist die Kovarianz identisch mit der gewöhnlichen Varianz s^2.

$$cov(x,x) = \frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x}) * (x_i - \bar{x}) = \frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})^2 = s_x^2$$

- Ebenso wie für die Varianz gilt auch für die Kovarianz ein Verschiebesatz:

$$cov(x,y) = \frac{1}{N}\sum_{i=1}^{N} x_i y_i - \overline{xy} = \overline{x * y} - \bar{x} * \bar{y}$$

- Werden die Merkmalswerte x_i in der Form $u_i = a_1 + b_1 x_i$ und in der Form $v_i = a_2 + b_2 y_i$ linear transformiert, so ergibt sich folgende Transformationsregel:

$$cov(u,v) = b_1 * b_2 * cov(x,y)$$

d.h. die additiven Konstanten a_1 und a_2, die eine Verschiebung der Punktewolke bewirken, haben keinen Einfluss auf die Kovarianz.

Beispiel: Rabatt und Stückzahl

Ein Unternehmen hat in den letzten Wochen eine Rabattaktion gestartet, bei der die Höhe des gegebenen Rabatts variiert. Neben der Höhe des Rabatts *X* wurden die verkauften Stückzahlen *Y* notiert. Um die Aktion auszuwerten, sollen die Daten der folgenden Urliste graphisch in Form eines Streudiagramms dargestellt werden und die Kovarianz berechnet werden.

X: Höhe des Rabatts in %
Y: Verkaufte Stückzahl

X	3	5	3	7	4	9	1	0	3	6
Y	1121	1345	1 113	1 405	1098	1587	1042	992	1 121	1374

Wir erhalten das folgende Streudiagramm:

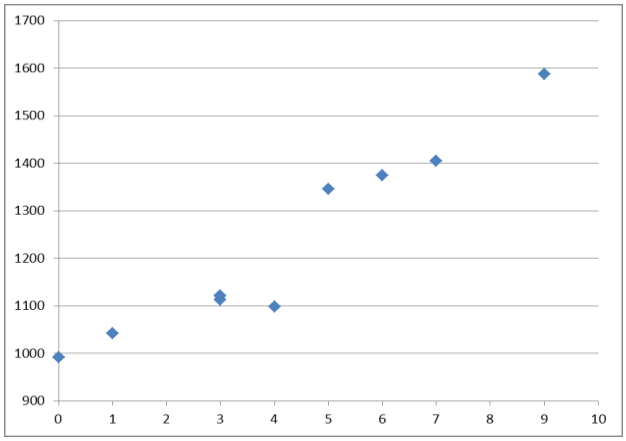


Abb. 8: Punktediagramm zwischen X und Y.

Im Streudiagramm ist zu erkennen, dass die abgesetzte Stückzahl mit zunehmendem Rabatt steigt, d.h. ein positiver Zusammenhang zwischen zunehmenden Rabatt und abgesetzter Stückzahl vorliegt. Wir werden also für die Kovarianz einen positiven Wert erhalten. Aus der Urliste ergibt sich $\bar{x}$ = 4,1 und $\bar{y}$ = 1219,8 und damit

$cov(x, y)$ = ((3 - 4,1) (1121 - 1219,8) + (5 - 4,1) (1345 - 1219,8)
+ (3 - 4,1) (1113 - 1219,8) + (7 - 4,1) (1405-1219,8)
+ (4 - 4,1) (1098 - 1219,8) + (9 - 4,1) (1587-1219,8)
+ (1 - 4,1) (1042 - 1219,8) + (0 - 4,1) (992 - 1219,8)
+ (3 - 4,1) (1121 - 1219,8) + (6 - 4,1) (1347 - 1219,8)) = 457,42

Beispiel: Handwerksbetriebe

Für **die Merkmale** X = „Anzahl Beschäftigte“ und Y = „Umsatz in 100 000 €“ aus Handwerksbetriebsbeispiel mit der dazugehörigen zweidimensionalen Häufigkeitstabelle

		Y					
		2	**3**	**5**	**6**	**7**	$f_{j.}$
X	**2**	1	-	-	-	-	1
	3	1	2	1	-	-	4
	4	-	-	2	2	-	4
	5	-	-	-	2	1	3
	$f_{.k}$	2	2	3	4	1	12

und den Mittelwerten $\bar{x}$ = 3,75 und $\bar{y}$ = 5 erhalten wir mit der Kovarianz-Formel die folgende Arbeitstabelle:

(x_j, y_k)	f_{jk}	$(x_j - \bar{x})$	$(y_k - \bar{y})$	$(x_j - \bar{x}) * (y_k - \bar{y}) * f_{jk}$
(2,3)	1	-1,75	-2	3,5
(3,3)	1	-0,75	-2	1,5
(3,4)	2	-0,75	-1	1,5
(3,5)	1	-0,75	0	0
(4,5)	2	0,25	0	0
(4,6)	2	0,25	1	0,5
(5,6)	2	1,25	1	2,5
(5,7)	1	1,25	2	2,5
				12

Die Kovrianz

$$cov(x, y) = \frac{12}{12} = 1 \quad (in\ 100.000€)$$

d.h. es besteht ein positiver Zusammenhang zwischen Anzahl der Beschäftigten und Höhe des Umsatzes.

Die Kovarianz ist in der Lage, den empirischen Zusammenhang zweier Merkmale *X* und *Y* aufzuzeigen. $cov(x, y) > 0$ zeigt einen positiven, $cov(x, y) < 0$ einen negativen Zusammenhang. Die Kovarianz kann also die Richtung, aber nicht die Stärke des Zusammenhangs messen. Ein sehr großer positiver Wert bedeutet beispielsweise nicht zwangsläufig, dass ein sehr starker positiver Zusammenhang vorliegt. Die Kovarianz ist nämlich eine dimensionsbehaftete Größe, die allein durch die Änderung der Maßeinheit kleiner oder größer werden kann. Um diesem Problem zu begegnen, wird eine normierte Kovarianz als Maßzahl verwendet.

3.3.2 Korrelationskoeffizient

Der Korrelationskoeffizient normiert die Kovarianz, indem er sie durch die Standardabweichungen der beiden Merkmale teilt. Daraus ergibt sich ein normiertes Maß.

Korrelationskoeffizient (nach Bravais-Pearson)

Der Korrelationskoeffizient nach Bravais-Pearson ergibt sich aus den Daten $(x_i, y_i), i = 1, \ldots, N$ durch

$$r_{xy} = \frac{\sum_{i=1}^{N}(x_i - \bar{x}) * (y_i - \bar{y})}{\sqrt{\sum_{i=1}^{N}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{N}(y_i - \bar{y})^2}} \; bzw. r_{xy} = \frac{cov(x.y)}{s_x * s_y}$$

mit dem Wertebereich

$$-1 \leq r_{xy} \leq +1 \; bzw. \; |r_{xy}| \leq +1.$$

Die Art des gemessenen Zusammenhangs wird deutlich, wenn man die Extremwerte von r_{xy} betrachtet. $r_{xy} = \pm 1$ gilt genau dann, wenn zwischen *X* und *Y* eine exakte lineare Beziehung $y_i = a + b * x_i$ besteht.

Eigenschaften des Korrelationskoeffizienten

Für die folgenden Werte gilt:

- $|r_{xy}| = 1 \; => \;$ *X* und *Y* sind perfekt korreliert. Die Werte (x_i, y_i) liegen exakt auf einer Gera den mit positiver oder negativer Steigung.
- $r_{xy} > 0 \; => \;$ *X* und *Y* sind positiv korreliert. Die Werte (x_i, y_i) liegen um eine Gerade mit positiver Steigung.
- $r_{xy} < 0 => \;$ *X* und *Y* sind negativ korreliert. Die Werte (x_i, y_i) liegen um eine Gerade mit negativer Steigung.
- $r_{xy} = 0 \; => \;$ *X* und *Y* sind unkorreliert. Die Werte (x_i, y_i) streuen in der Regel gleichmäßig um den Schwerpunkt $(\bar{x}, \bar{y})$. Es besteht kein linearer Zusammenhang zwischen *X* und *Y*.

Im folgenden Beispiel sind einige Streudiagramme dargestellt und die sich ergebenden Korrelationskoeffizienten quantifiziert.

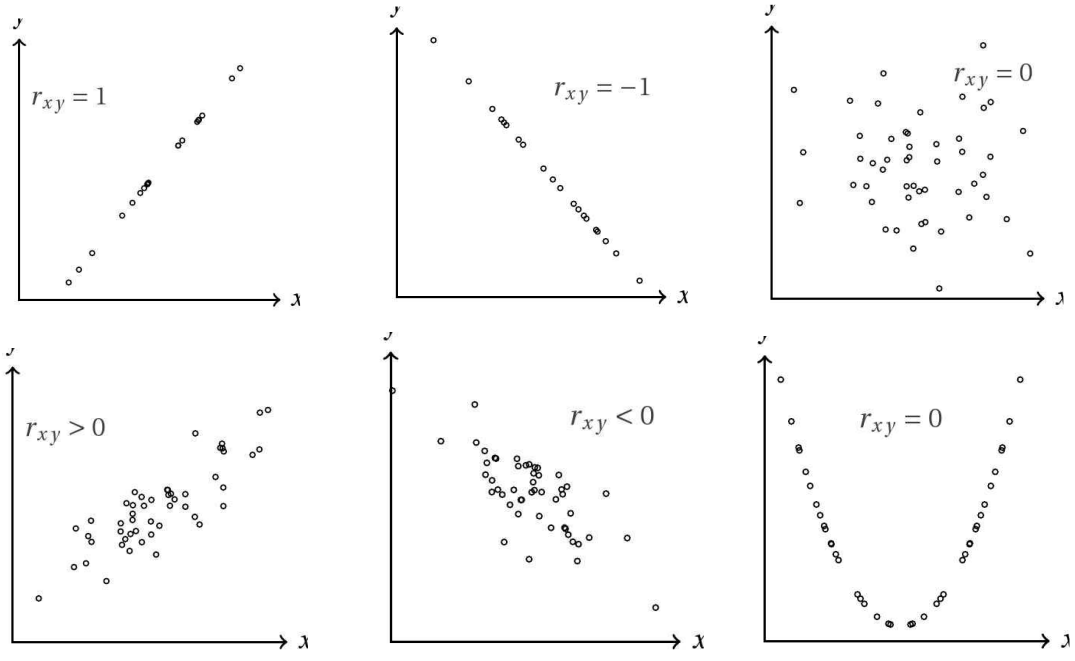


Abb. 9: Streudiagramme und Korrelationskoeffizienten

Von besonderem Interesse ist das letzte Streudiagramm, wo man zwischen *X* und *Y* einen deutlichen Zusammenhang feststellt. Da dieser Zusammenhang nicht linear ist, wird er vom Korrelationskoeffizienten nicht erfasst. Der Korrelationskoeffizient ergibt null. Dies zeigt auch das folgende Zahlenbeispiel:

Beispiel:

Wir betrachten folgende Urliste:

X_i	-3	-2	-1	0	1	2	3
Y_i	9	4	1	0	1	4	9

Mithilfe des Verschiebesatzes-Formels erhalten wir

$$cov(x,y) = \frac{1}{N}\sum_{i=1}^{N} x_i y_i - \overline{xy} = 0 - 0 = 0$$

und damit $r_{xy} = 0$. Zwischen den beiden Merkmalen besteht also keine lineare Beziehung. Der Zusammenhang zwischen *X* und *Y* ist quadratisch, da $y_i = x_i^2$ gilt. Dieser nicht-lineare Zusammenhang wird vom Korrelationskoeffizienten nicht aufgedeckt.

Beispiel: Handwerksbetriebe

Für das Handwerksbetriebsbeispiel soll der Korrelationskoeffizient berechnet werden. Mit den Standardabweichungen von *X* und *Y* sowie der Kovarianz zwischen *X* und *Y*

$$cov(x,y) = 1, \qquad s_x = 0{,}92, \qquad s_y = 1{,}22$$

erhalten wir

$$r_{xy} = \frac{1}{0{,}92 * 1{,}22} = 0{,}89.$$

Beschäftigtenzahl und Umsatz sind also sehr stark positiv korreliert.

3.4 Regressionsanalyse

Die Korrelation, die im letzten Abschnitt behandelt wurde, lässt offen, ob eine der betrachteten Variablen auf die andere wirkt, oder ob lediglich ein ungerichteter Zusammenhang vorliegt. Im Rahmen der Regressionsanalyse beschäftigt man sich mit der Wirkung von einer oder mehreren Größen - den Einflussgrößen oder unabhängigen Variablen - auf eine andere Größe, die abhängige Variable. Es liegt also eine Richtung der Abhängigkeit vor.

Im einigen obigen Beispiele wurden zwischen den Variablen X und Y einen hohen

Korrelationskoeffizient berechnet und damit einen sehr starken linearen Zusammenhang nachgewiesen. Ein solcher Zusammenhang ließe sich durch eine lineare Funktion

$$Y = a + b * X$$

beschreiben. Allerdings sehen wir an den obigen Streudiagrammen deutlich, dass ein so klarer funktionaler Zusammenhang nicht vorliegt, da die Beobachtungen nicht auf einer Geraden liegen. Dieser Tatsache trägt man dadurch Rechnung, dass man den linearen Zusammenhang nicht als exakt annimmt, sondern durch einen Fehlerterm *u* ergänzt:

$$Y = a + b * X + u.$$

Bei *u* handelt es sich dann um einen Fehler, einen Teil der Werte von *Y*, der durch den funktionalen Zusammenhang nicht erklärt wird.

Lineare Regression

Seien $(x_i\ y_i)$, $i = 1, \ldots, N$, Beobachtungen der Merkmale *X* und *Y*, dann heißt

$$y_i = a + b * x_i + u_i$$

lineare Regression, wobei *a* den Achsenabschnitt, *b* die Steigung der Geraden und u_i die Fehler bezeichnen.
a und *b* stehen für konkrete Werte, die aus gegebenen Daten geschätzt werden können.

In der nachstehenden Abbildung ist ein Streudiagramm zu sehen, in dem die zu den Variablen gehörende exakte lineare Beziehung *Y = a+ bX*, die Beobachtungswerte $(x_i\ y_i)$ sowie die Fehler u_i eingezeichnet sind.

Wichtige Themen im Zusammenhang mit der linearen Regression sind die Bestimmung von Schätzungen für die Koeffizienten *a* und *b* und die Untersuchung der Güte der Regression. Die Regression hat auch die wichtige Anwendung, Prognosen für eine Größe *Y* anhand einer Merkmalsausprägung x_{prog} des Merkmals *X* zu liefern, wenn der Wert der Variable *Y* für x_{prog} nicht bekannt ist.

Beispiel:

Beispiele für die Anwendung einer Regression sind:

- Einfluss des Verkaufspreises auf die abgesetzte Menge.
- Einfluss der Höhe der Investitionen auf zukünftige Gewinne.
- Einfluss der konsumierten Kalorien auf die Gewichtszunahme.
- Einfluss der Wohnfläche auf die Höhe der Miete.

Im Folgenden wollen wir uns mit der Schätzung der Regressionskoeffizienten befassen.

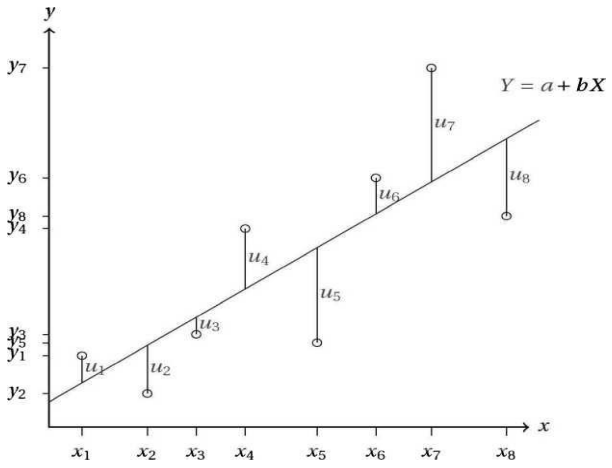


Abb. 10: Lineare Regression

3.4.1 Schätzung der Regressionskoeffizienten

Die gängige Methode, die Regressionskoeffizienten zu schätzen, ist die Methode der kleinsten Quadrate. Dadurch wird versucht, den Fehler u_i zu minimieren. Die Kleinste-Quadrate-Schätzer (KQ-Schätzer) für die Regressionskoeffizienten *a* und *b* sind gegeben durch

$$\hat{b} = \frac{cov(x,y)}{s_x^2}$$

$$\hat{a} = \bar{y} - \hat{b} * \bar{x}$$

Beispiel 3.27 Unternehmensgewinne und Umsatz

Die Höhe von Unternehmensgewinnen soll durch den Umsatz erklärt werden. Dazu liegt ein Datensatz der US-Industrie vor. Er beinhaltet Umsätze der gesamten verarbeitenden US-Industrie *X* und Gewinne *Y* je Einheit des S&P-500-Index in US-Dollar für die Jahre 2000 bis 2011. Der Index deckt 75% des gesamten Markts ab.

Jahr	Umsatz in Bio. US-$	Gewinn pro Einheit S&P500 in US-$
2000	4,2	56,1
2001	3,97	38,9
2002	3,92	46
2003	4,02	54,7
2004	4,29	67,9
2005	4,74	76,5
2006	5,02	87,7
2007	5,32	82,5
2008	5,45	65,4
2009	4,44	60,8
2010	4,82	83,7
2011	5,37	97,1

$$\bar{x} = 4{,}63 \;\; \bar{y} = 68{,}11 \;\; s_x^2 = 0{,}3 \;\; cov(x, y) = 7{,}54.$$

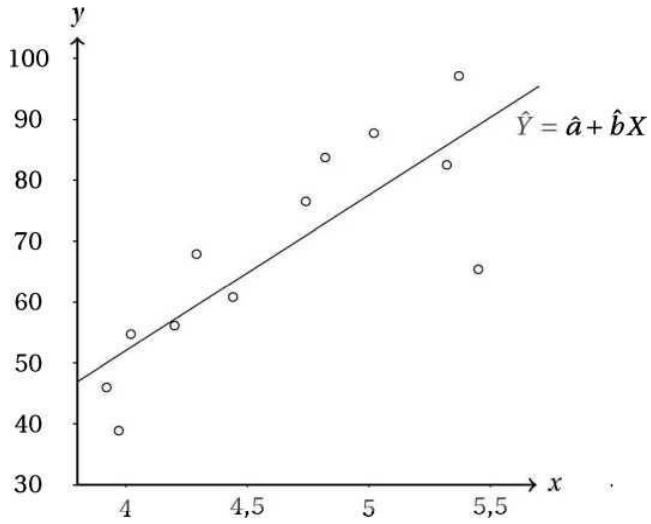
Als Schätzungen für die Koeffizienten ergeben sich:

$$\hat{b} = \frac{7{,}54}{0{,}3} = 25{,}52 \qquad \hat{a} = 68{,}11 - 25{,}52 * 4{,}63 = -50{,}05$$

Die Regressionsgleichung lautet:

$$\hat{y} = -50{,}05 + 25{,}52 * x$$

Der Koeffizient *b* kann so interpretiert werden: „Mit je einer Bio. US-Dollar höherem Umsatz wird im Durchschnitt pro Einheit 25,52 US-$ mehr Gewinn erzielt“.
In nachfolgendem Bild sind die Beobachtungen und die geschätzte Regressionsgerade abgebildet.



3.4.2 Prognose

Liegen Schätzungen $\hat{a}$ und $\hat{b}$ für die Regressionskoeffizienten vor, so kann zu einem gegebenen Wert x_{prog} der Wert der abhängigen Variable *y* prognostiziert werden. Dies geschieht einfach, indem man Wert x_{prog} in die Gleichung des geschätzten linearen Zusammenhangs $\hat{y} = \hat{a} + \hat{b} * x$ einsetzt.

Prognose für einen Wert

Zu einem Wert x_{prog} für die unabhängige Variable *X* ist die Prognose für die abhängige Variable *Y* durch

$$y_{prog} = = \hat{a} + \hat{b} * x_{prog}$$

gegeben.

Beispiel Unternehmensgewinne und Umsatz (Fortsetzung)

Im obigen Beispiel ergeben sich für Umsätze von 3, 4, 5 und 6 Bio. US-Dollar folgende prognostizierte Gewinne pro Einheit:

Umsatz in Bio. US-$ x_{prog}	Prognose y_{prog}
3	-50,05 + 25,52*3 = 26,51
4	-50,05 + 25,52*4 = 52,03
5	-50,05 + 25,52*5 = 77,55
6	-50,05 + 25,52 * 6 = 103,07

Prognosen durch ein Regressionsmodell zu erhalten, ist für die Anwendung von großer Bedeutung. In der Praxis ist es dabei wichtig, einschätzen zu können, wie genau die jeweiligen Prognosen sind. Anhaltspunkte hierfür liefert das Bestimmtheitsmaß R^2, das im nächsten Abschnitt behandelt wird.

3.4.3 Güte der Anpassung

Durch eine Regression wird nur ein Teil der Werte der abhängigen Variable erklärt. Der restliche Teil bildet den Fehler, der nicht erklärt werden kann. Für die Praxis ist es wünschenswert, eine Maßzahl zur Verfügung zu haben, die angibt, wie gut eine Regression die Zielgröße erklärt. Eine solche Maßzahl stellt das Bestimmtheitsmaß dar.

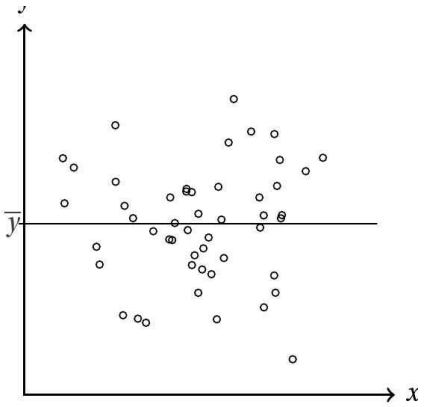
Bestimmtheitsmaß ist durch

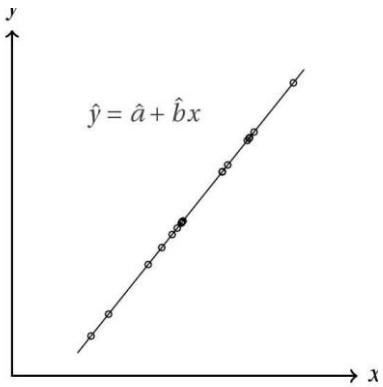
$$R^2 = \frac{\sum_{i=1}^{N}(\hat{y} - \bar{y})^2}{\sum_{i=1}^{N}(y - \bar{y})^2} \quad bzw. R^2 = \frac{s_{\hat{y}}^2}{s_y^2}$$

gegeben.
Da in der Vorlesung nur zwei Dimensionen bhandelt werden gilt $R^2 = r_{xy}^2$

Eigenschaften des Bestimmtheitsmaßes:

- $0 \leq R^2 \leq 1$
- Falls $R^2 = 0$, dann verläuft die Regressionsgerade waagrecht und besitzt keinerlei Erklärungswert: die abhängige Variable hängt nicht (jedenfalls soweit ihr Mittelwert betroffen ist) von der unabhängigen Variablen ab.
    - Falls $R^2 = 1$, dann liegen alle Punkte auf der Regressionsgerade





Beispiel: Unternehmensgewinne und Umsatz (Fortsetzung)

Für die Regression aus dem obigen Beispiel soll das Bestimmtheitsmaß berechnet werden. Die geschätzte Regressionsbeziehung lautete: y = -50,05 + 25,52x. Dazu sollen zunächst die Schätzwerte $\hat{y}$ und die Residuen $\hat{u}$ berechnet werden.

Für sämtliche Beobachtungen gelten die Werte in der Tabelle:

i	x_i	y_i	$\widehat{y_i}$	$\widehat{u_i}$
1	4,2	56,1	57,13	-1,03
2	3,97	38,9	51,26	-12,36
3	3,92	46	49,99	-3,99
4	4,02	54,7	52,54	2,16
5	4,29	67,9	59,43	8,47
6	4,74	76,5	70,92	5,58
7	5,02	87,7	78,06	9,64
8	5,32	82,5	85,72	-3,22
9	5,45	65,4	89,04	-23,64
10	4,44	60,8	63,26	-2,46
11	4,82	83,7	72,96	10,74
12	5,37	97,1	86,99	10,11

Und daraus ergibt sich

$$R^2 = \frac{2308,76}{3471,67} = 0,67.$$

4. Das Rechnen mit Wahrscheinlichkeiten

In nahezu allen Anwendungsgebieten der Wirtschaftswissenschaften werden häufig Vorgänge beobachtet oder Versuche durchgeführt, die zufallsabhängig sind. Beispiele sind etwa Staus im Straßenverkehr, Warteschlangen vor Bedienschaltern, das Zustandekommen der Aktienkurse, die Durchführung einer Meinungsumfrage sowie die die Qualitätskontrolle einer laufenden Produktion. Von besonderer Bedeutung für die Statistik ist das Experiment der zufälligen Ziehung einer Stichprobe aus einer Grundgesamtheit. Da das Verständnis der Wahrscheinlichkeitsrechnung anhand realer Entscheidungssituationen aus dem Wirtschaftsleben jedoch oft schwer fällt, werden die Regeln der Wahrscheinlichkeitsrechnung häufig an einfacheren Zufalls Vorgängen aus dem Bereich der Glücksspiele, wie etwa Würfel- oder Kartenspiele erläutert.

4.1 Zufallsvorgänge und deren Beschreibung

4.1.1 Zufallsvorgang

Ein Zufalls vorgang führt zu einem von mehreren sich gegenseitig ausschließenden Ergebnissen. Es ist vor der Durchführung ungewiss, welches Ergebnis eintreten wird.

Beispiel: Würfeln mit einem Würfel

Es ist naheliegend, dass die möglichen die 6 verschiedenen Zahlen 1, 2, 3, 4, 5, 6 sind.

Die Menge dieser möglichen Versuchsergebnisse wird im Folgenden immer mit dem griechischen Buchstabe Ω n bezeichnet.

4.1.2 Elementarereignis

Die einzelnen - nicht weiter zerlegbaren - möglichen Ergebnisse eines Zufallsvorgangs, die sich gegenseitig ausschließen, werden als Elementarereignisse ω bezeichnet.

4.1.3 Ergebnisraum

Die Menge Ω aller Elementarereignisse eines Zufallsvorgangs nennen wir Ergebnisraum.

Beispiel: Münzwurf

Bezeichnen wir das Elementarereignis, dass die Münze nach dem Wurf „Kopf“ zeigt mit K, und dass sie „Zahl“ zeigt mit Z, ergibt sich der Ergebnisraum

$$\Omega = \{K,Z\}.$$

Beispiel: Zweimaliger Münzwurf

Hier sind die Elementarereignisse geordnete Paare wobei das erste Element das Ergebnis des

ersten Wurfes und das zweite Element das Ergebnis des zweiten Wurfes darstellt. Der Ergebnisraum ist damit $\Omega = \{(K, K), (K, Z), (Z, K), (Z, Z)\}$ bzw. kürzer $\Omega = \{KK, KZ, ZK, ZZ\}$.

Beispiel: Zweimaliger Wurf eines Würfels

Ein Würfel wird zweimal hintereinander geworfen. Jedes Ergebnis *a)* dieses Zufallsexperimentes besteht aus einem Zahlenpaar wobei *i* und *j* jeweils eine der 6 Augenzahlen sind. Bei $i = 1, \ldots, 6$ und $j = 1, \ldots, 6$ ergeben sich $6x6 = 36$ Elementarereignisse und der folgende Ergebnisraum:

$$\begin{aligned}\Omega = \{&(1,1), (1,2), (1,3), (1,4), (1,5), (1,6),\\ &(2,1), (2,2), (2,3), (2,4), (2,5), (2,6),\\ &(3,1), (3,2), (3,3), (3,4), (3,5), (3,6),\\ &(4,1), (4,2), (4,3), (4,4), (4,5), (4,6),\\ &(5,1), (5,2), (5,3), (5,4), (5,5), (5,6),\\ &(6,1), (6,2), (6,3), (6,4), (6,5), (6,6),\}.\end{aligned}$$

Beispiel: Bedienschalter

Ein Kunde will bei der Post ein Paket versenden. Er interessiert sich für die Wartezeit, die er vor dem Schalter bis zu seiner Bedienung verbringen muss. Seine Wartezeit kann zwischen 0 (falls keine Kunde vor dem Schalter steht) und einer maximalen Zeit *T* (z. B. Schalterschluss) schwanken. Misst er seine Wartezeit in Minuten, so lautet der Ergebnisraum dieses Zufallsexperiments $\Omega = \{0,1, \ldots, T\}$. Lässt man Fragen der Messgenauigkeit außer Acht und betrachtet man die Zeit als stetige Größe, so ergibt sich als Ergebnisraum

$$\Omega = \{x \mid 0 < x < T, x\ reell\}.$$

Neben den einzelnen möglichen Elementarereignissen eines Zufallsexperimentes betrachtet man vor allem gewisse Ereignisse, die aus einer Menge von Elementarereignissen bestehen. Wir bezeichnen solche Ereignisse mit großen lateinischen Buchstaben.

4.1.4 Ereignis

Ein zufälliges Ereignis ist eine Teilmenge von Ω. Man sagt, das Ereignis A tritt ein, wenn das Zufallsexperiment ein Ergebnis ω liefert, das zu A gehört.

Beispiel: Ereignisse

Beim Zufallsexperiment „Würfeln mit einem Würfel“ mit $\Omega = \{1,2,3,4,5,6\}$ sind folgende Ereignisse denkbar:

A_1: „Die geworfene Augenzahl ist gerade“, d. h. $A_1 = \{2,4,6\}$
A_2: „Die 1 wird gewürfelt“, d. h. $A_2 = \{1\}$

Mögliche Ereignisse beim Zufallsexperiment „Zweimaliger Wurf eines Würfels“:

A_1: „Die Summe der Augenzahlen ist mindestens 11“, d.h. $A_1 = \{(5,6), (6,5), (6,6)\}$
A_2: „Es wird zweimal die 6 gewürfelt“, d. h. $A_2 = \{(6,6)\}$
A_3: „Die zuerst gewürfelte Zahl ist eine 1 “,

d. h. A_3 = {(1,1), (1,2), (1,3), (1,4), (1,5), (1,6)}

Im Beispiel „Bedienschalter“interessiert man sich für folgendes Ereignis:

A_1: „Die Wartezeit beträgt zwischen 2 und 4 Minuten“,
d.h. A1 = {x | 2 < x < 4, x reell}
A_2: „Die Wartezeit liegt unter 3 Minuten“, d. h. A_2 = {x | 0 < x < 3, x reell}

4.1.5 Sicheres Ereignis

Ein Ereignis wird als sicheres Ereignis bezeichnet, wenn das Ereignis mit dem Ergebnisraum Ω identisch ist. Dieses tritt es bei jedem Ausführen des Zufallsvorgangs immer ein.

4.1.6 Unmögliches Ereignis

Ein Ereignis tritt sicher nie ein und gilt als unmöglich, wenn es gleich der sogenannten leeren Menge $\emptyset$ ist. Da $\emptyset$ = {} kein Element enthält, tritt dieses Ereignis niemals ein.

4.2 Die Verknüpfung von Ereignissen

Im Zusammenhang mit Ereignissen stehen Fragen wie „Treten zwei bestimmte Ereignisse ein?“ oder „Tritt zumindest eines von mehreren Ereignissen ein?“. Solche Fragen werden durch die Verknüpfung von Ereignissen behandelt.

Ereignisse und Ergebnisraum sowie die Verknüpfung von Ereignissen lassen sich im sog. Venn-Diagramm anschaulich darstellen. Die Ereignisse werden dabei als Flächen dargestellt, die sich überlappen können.

4.2.1 Vereinigung

Die Vereinigung zweier Ereignisse A und *B* ist definiert als die Menge der Elementarereignisse *o),* die entweder zu A allein oder zu *B* allein oder sowohl zu A und zu *B* gemeinsam gehören.

$$A \cup B = \{\omega \mid \omega \in A \text{ oder } \omega \in B\}$$

Das Ereignis A ∪ B tritt somit genau dann ein, wenn A oder *B* allein oder gemeinsam eintreten. Wir sagen kurz „A oder *B“* treten ein.

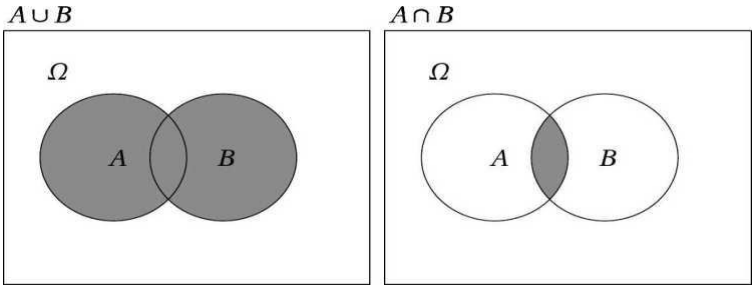
4.2.2 Durchschnitt

Der Durchschnitt zweier Ereignisse *A* und *B* ist definiert als die Menge der Elementarereignisse ω, die sowohl zu *A* als auch zu *B* gehören.

$$A \cap B = \{\omega \mid \omega \in A \text{ und } \omega \in B\}$$

Das Ereignis *A* ∩ *B* tritt somit genau dann ein, wenn Ereignis *A* und Ereignis *B* gemeinsam eintreten.

Wir sagen kurz „*A* und *B*“ treten ein.



4.2.3 Differenz

Die Differenz zwei Ereignisse ist die Menge aller Elementarereignisse ω, die zu A gehören aber nicht zu B.

$$A \setminus B = \{\omega \mid \omega \in A \text{ und } \omega \notin B\}.$$

Das Ereignis $A \setminus B$ tritt somit genau dann ein, wenn zwar *A*, aber nicht *B* eintritt.

4.2.4 Komplement

Ein Ereignis, dass genau dann eintritt, wenn sich ein ω ergibt, das nicht zu dem Ereignis *A* gehört, heißt Komplement von *A*.

$$\bar{A} = \Omega \setminus A = \{\omega \mid \omega \in \Omega \text{ und } \omega \notin A\}.$$

Wir sagen kurz, das Ereignis nicht *A* tritt ein.

4.2.5 Disjunkte Ereignisse

Zwei Ereignisse heißen disjunkt oder unvereinbar, wenn gilt

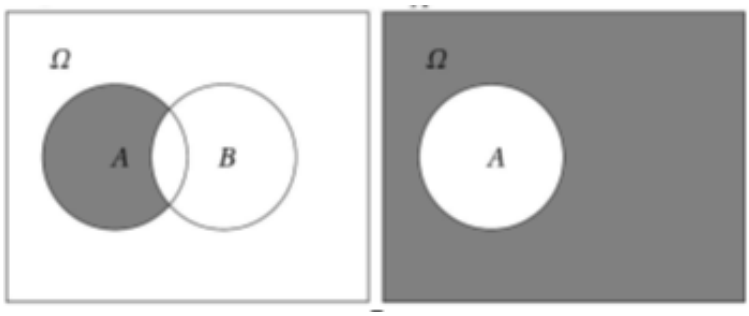
$$A \cap B = \emptyset.$$

Bemerkung:

Die Vereinigung und der Durchschnitt von abzählbar unendlich vielen Ereignissen Ai, A_2, A_3, ... werden analog wie oben definiert:

$$\bigcap_{i=1}^{\infty} A_i = A1 \cup A2 \cup A3 \cup \ldots.$$

$$\bigcup_{i=1}^{\infty} A_i = A1 \cup A2 \cup A3 \cup \ldots.$$



$A \setminus B$ $\quad\quad$ $\bar{A}$

Beispiel: Zweimaliger Wurf eines Würfels

Wir betrachten beim zweimaligen Wurf eines Würfels die folgenden Ereignisse:
A_1: „Die Summe der Augenzahlen ist mindestens 11“
A_2: „Es wird zweimal die 6 gewürfelt“
A_3: „Die zuerst gewürfelte Zahl ist eine 1“

A_2 und A_3 sind unvereinbare Ereignisse, d.h. es gilt $A_2 \cap A_3 = \emptyset$.

Weiter gilt:
$A_1 \cap A_2 = \{(6,6)\}$
$A_1 \cup A_2 \cup A_3 = \{(1,1), (1,2), (1,3), (1,4), (1,5), (1,6), (5,6), (6,5), (6,6)\}$
$A_1 \setminus A_2 = \{(5,6), (6,5)\}$.

.

4.3 Die Axiome von Kolmogoroff

Das Ergebnis eines Zufallsexperimentes ist nicht vorhersehbar. Es ist höchstens möglich, den Ereignissen gewisse Wahrscheinlichkeiten zuzuordnen. Eine Wahrscheinlichkeit ist also nichts anderes als ein Maß zur Quantifizierung des Grades der Sicherheit oder Unsicherheit des Eintretens eines bestimmten Ereignisses im Rahmen eines Zufallsexperimentes.
Jede Funktion *P*, die einem Ereignis *A* eine Wahrscheinlichkeit *P(A)* zuordnet, wird als Wahrscheinlichkeitsfunktion und P(A) als Wahrscheinlichkeit von *A* bezeichnet, wenn sie die drei folgenden, von A. N. Kolmogoroff 1933 formulierten Axiome erfüllt:

4.3.1 Axiome der Wahrscheinlichkeitsrechnung

Axiom 1: Jedem Ereignis *A* wird eine nicht-negative reelle Zahl P(A),
d.h. $P(A) \geq 0$.
Axiom 2: Die Wahrscheinlichkeit des sicheren Ereignisses Ω ist gleich eins:
$P(\Omega) = 1$.
Axiom 3: Additivität: Für abzählbar unendlich viele, paarweise disjunkte
Ereignisse Ai, A_2, A_3, ... gilt stets: $P(A_1 \cup A_2 \cup A_3 \cup ...) = P(A_1) + P(A_2) + P(A_3) + ...$

Das Axiom 3 in der obigen Gestalt ist nur für unendliche Ergebnismengen von Bedeutung, aber auch für endlich viele Ereignisse, d.h.:

$P(A_1 \cup A_2 \cup A_3 \cup .... \cup A_n) = P(A_1) + P(A_2) + ... + P(A_n)$, falls $A_1 \cap A_2 = \emptyset$ für alle $i\ j$ mit $i, j = 1,2,3,..., n$.

Für zwei nicht disjunkte Ereignisse gilt:

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

4.3.2 Laplace-Experiment

Ein Laplace-Experiment ist ein Experiment, in dem die Elementarereignisse dieselbe Wahrscheinlichkeit haben.

Falls Ω endlich ist, dann gilt für jedes beliebiges Ereignis:

$$P(A) = \frac{|A|}{|\Omega|} = \frac{Anzahl\ der\ f\ddot{u}r\ A\ g\ddot{u}nstigen\ Elementarereignisse}{Anzahl\ der\ m\ddot{o}glichen\ gleich\ wahrscheinlichen\ Elementarereignisse}$$

Beispiel: Würfeln mit einem Würfel

- Die Wahrscheinlichkeit beim Würfeln mit einem Laplace-Würfel eine gerade Augenzahl zu erhalten, ergibt sich wie folgt:

$$P(\text{„gerade Augenzahl"}) = P(\{2,4,6\}) = \frac{|\{2,4,6\}|}{|\Omega|} = \frac{3}{6} = 0{,}5.$$

- Die Wahrscheinlichkeit beim Würfeln mit zwei Laplace-Würfeln eine Augensumme größer 10 zu erhalten ist:

$$P(\text{„Augensumme größer 10"}) = P(\{(5,6),(6,5),(6,6)\}) = \frac{|\{(5,6),(6,5),(6,6)\}|}{|\Omega|} = \frac{3}{36} = \frac{1}{12}.$$

4.4 Zufallsauswahl und Kombinatorik

Die Kombinatorik ist ein Teilgebiet der Mathematik, das sich mit Fragen des Abzählens beschäftigt. Gezählt werden die möglichen unterschiedlichen Anordnungen von Elementen oder die möglichen Auswahlen von Elementen aus einer Menge. Letzteres entspricht auch den möglichen Stichproben aus einer Grundgesamtheit, die in der Statistik eine wichtige Rolle spielen. Im Rahmen der Wahrscheinlichkeitsrechnung bedient man sich der Kombinatorik, um Verteilungen herzuleiten, die die Behandlung von Stichproben ermöglichen. Historisch war die Wahrscheinlichkeitsrechnung eng mit dem Glücksspiel verknüpft. Hier geben Ergebnisse der Kombinatorik zusammen mit dem Wahrscheinlichkeitsbegriff von Laplace Aufschluss über Chancen. So kann die Frage beantwortet werden, wie wahrscheinlich es ist, einen „Sechser“ im Lotto zu haben, oder ein bestimmtes Blatt bei einem Kartenspiel zu erhalten.

4.4.1 Zufallsauswahl und Urnenmodell

In den obigen Abschnitten haben wir gesehen, dass es in bestimmten Fällen möglich ist, Wahrscheinlichkeiten durch einfaches Abzählen zu ermitteln, wobei dazu jeweils der Quotient aus der

Anzahl der für ein Ereignis günstigen Elementarereignisse und der Anzahl aller möglichen Elementarereignisse gebildet wurde. Das Urnenmodell ist ein gutes Beispiel, um zu zeigen, dass die Anzahl von möglichen Ereignissen davon abhängt, in welcher Weise die Ziehung von Kugeln aus einer Urne vorgenommen wird.
Die dort angestellten Überlegungen lassen sich verallgemeinern. Unter einer zufälligen Auswahl verstehen wir das zufällige Ziehen von *k* Objekten (Stichprobe) aus einer endlichen Menge von *n* Objekten (Grundgesamtheit). Zur Veranschaulichung dieser Situation wird das sog. Urnenmodell herangezogen,

4.4.2 Modell mit Zurücklegen

Bei einer Ziehung mit Zurücklegen aus einer Grundgesamtheit vom Umfang *n* ist die Anzahl der möglichen Stichproben vom Umfang *k* gleich

$$n * n * \ldots * n = n^k.$$

4.4.3 Permutationen

Gegeben sei eine Menge von *n* Elementen. Jede Zusammenstellung dieser *n* Elemente in einer beliebigen Reihenfolge heißt Permutation dieser *n* Elemente.
In diesem Abschnitt wird untersucht, wie viele unterschiedliche Permutationen es bei *n* Elementen gibt. Dabei wird unterschieden, ob sich sämtliche Elemente unterscheiden lassen oder nicht. Zunächst soll angenommen werden, dass sich sämtliche Elemente unterscheiden.

Beispiel:

Aus der Menge {a, b, c} lassen sich 6 Permutationen bilden, nämlich

$$\{abc\}, \{bac\}, \{cab\}, \{acb\}, \{bca\}, \{cba\}$$

Die Anzahl aller Permutation von n Elementen beträgt

$$n! = n * (n - 1) * (n - 2) * \ldots * 1.$$

4.4.4 Kombinationen

Eine Kombination k-ter Ordnung ist eine Zusammenstellung von k Elementen aus einer Gesamtmenge von n Elementen.

Um die Anzahl der möglichen Kom binationen anzugeben, müssen wir unterscheiden, ob wir Kombinationen, die sich nur in der Reihenfolge unterscheiden, als verschieden ansehen oder nicht.

Beispiel: Kombinationen mit Berücksichtigung der Anordnung

Aus der Menge {a, b, c} lassen sich 6 Kombinationen 2-ter Ordnung bilden, wenn wir die Reihenfolge unterscheiden, nämlich

{ab}, {ba}, {ac}, {ca}, {bc}, {cb}.

Die im Beispiel ermittelte Zahl kann wieder *verallgemeinert werden: Es gibt genau n Möglichkeiten, den 1. Platz der Kombination zu besetzen, danach n - 1 Mögli*chkeiten für den 2. Platz usw. und schließlich $n - k + 1$ Möglichkeiten für den k-ten und letzten Platz. Somit ergeben sich

$$n * (n - 1) * (n - 2) * \dots * (n - k + 1) = \frac{n!}{(n-k)!}$$

Möglichkeiten (Modell ohne Zurücklegen mit Berücksichtigung der Reihenfolge).

Ohne die Berücksictigung der Reihenfolge bekommen wir weniger Mögichkeiten, nämich

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

4.5 Bedingte Wahrscheinlichkeiten und Unabhängigkeit von Ereignissen

Bei der Anwendung der Wahrscheinlichkeitstheorie auf praktische Probleme müssen häufig bestimmte Zusatzinformationen berücksichtigt werden. Bezeichne A das Ereignis, dass ein zufällig aus der Grundgesamtheit aller Haushalte in der Welt ausgewählter Haushalt ein jährliches Einkommen von über 40 000 € hat, und B das Ereignis, dass es sich dabei um einen europäischen Haushalt handelt, so wird die Wahrscheinlichkeit von A wohl eine andere sein als die Wahrscheinlichkeit von A unter der Bedingung B, d.h. wenn man sich nur auf die verkleinerte Grundgesamtheit der europäischen Haushalte bezieht.

Die Wahrscheinlichkeit für das Eintreten eines Ereignisses A unter der Bedingung, dass das Ereignis B eingetreten ist (oder gleichzeitig mit A eintritt), heißt „bedingte Wahrscheinlichkeit von A unter der Bedingung B“, und man schreibt $P(A|B)$.

4.5.1 Bedingte Wahrscheinlichkeit

Sei $P(B) > 0$, so ist

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

die bedingte Wahrscheinlichkeit von A unter der Bedingung, dass B eingetreten ist.

Beispiel: Bedingte Wahrscheinlichkeit

Beim Zufallsexperiment „Würfeln mit einem Würfel“ betrachten wir die beiden Ereignisse A: „6“ und B: „gerade Zahl“ = {2,4,6}. Die Wahrscheinlichkeit eine 6 zu werfen mit der Zusatzinformation, dass die gewürfelte Augenzahl eine gerade Zahl ist, berechnen wir intuitiv nach Laplace mit

$$P(A|B) = \frac{|A \cap B|}{|B|} = \frac{1}{3}.$$

Dieses Ergebnis würden wir auch mit der obigen Formel erhalten

$$P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{\frac{1}{6}}{\frac{3}{6}} = \frac{1}{3}.$$

Beispiel: Laptop und Tasche

Ein Kaufhaus hat einen Laptop und die dazugehörige Tasche im Angebot. Es wurde die Erfahrung gemacht, dass von 100 Interessenten 40 den Laptop (Ereignis A) kaufen 30 die Tasche kaufen (Ereignis B) und 20 sowohl den Laptop (Ereignis A) als auch die Tasche (Ereignis B). Es werden folglich die Wahrscheinlichkeiten

$$P(A) = 0{,}4\,, P(B) = 0{,}3 \text{ und } P(A \cap B) = 0{,}2$$

angenommen. Ein Verkäufer fragt sich nun, wie groß die Wahrscheinlichkeit ist, dass ein Käufer eines Laptops auch eine Tasche erwirbt. Die entsprechende Wahrscheinlichkeit ist durch

$$P(B|A) = \frac{P(A \cap B)}{P(A)} = \frac{0{,}2}{0{,}4} = 0{,}5.$$

gegeben. Folglich kaufen 50 % derer, die einen Laptop kaufen, auch eine Tasche.

Bemerkung:

Die bedingten Wahrscheinlichkeiten *P(A\B)* gehorchen ebenfalls den Axiomen von Kolmogo- roff:

1. $P(A|B) > 0$ für jedes Ereignis A,
2. $P(\Omega\,|B) = 1$,
3. $P(A_1 \cup A_2|B) = P(A_1|B) + P(A2|B)$, falls $A_1 \cap A_2 = \emptyset$.

Wie man sich leicht überzeugt. Somit sind alle bisher entwickelten Regeln der Wahrscheinlich- keitsrechnung auch für bedingte Wahrscheinlichkeiten gültig.

Aus der Definition der bedingten Wahrscheinlichkeit ergibt sich durch Umformung unmittelbar der wichtige Multiplikationssatz der Wahrscheinlichkeitsrechnung.

4.5.2 Multiplikationssatz für unabhängige Ereignisse

A und B sind genau dann unabhängig, wenn die Wahrscheinlichkeit, dass A und B eintreten, gleich dem Produkt der einzelnen Wahrscheinlichkeiten von A und B ist.

$$P(A \cap B) = P(A) * P(B)$$

Man kann den Multiplikationssatz auch auf mehr als zwei Ereignisse ausdehnen: Die Ereignisse $A_1, A_2, \ldots, A_n$ heißen unabhängig, wenn für jede Auswahl $A_{i1}, A_{i2}, \ldots, A_{in}$ von Ereignissen (mit $1 < m \leq n$) gilt:

$$P(A_{i1} \cap A_{i2} \cap \ldots \cap A_{in}) = P(A_{i1}) * P(A_{i2}) * \ldots * P(A_{im}).$$

4.6 Totale Wahrscheinlichkeit

Unter der totalen Wahrscheinlichkeit versteht man nichts anderes als eine einfache Wahrscheinlichkeit. Das Wort „total“ will den Gegensatz zur bedingten Wahrscheinlichkeit ausdrücken, die sich ja nur auf eine Teilmenge von Ω bezieht. In gewissen Fällen ist es möglich, die Information über das bedingte Eintreten eines Ereignisses zu nutzen, um die Wahrscheinlichkeit dieses Ereignisses insgesamt zu ermitteln.
Wir gehen von einer **disjunkten Zerlegung** des Ergebnisraumes Ω aus. Dabei spricht man von einer disjunkten Zerlegung, wenn sich Ω als Vereinigung von Teilmengen von Ω schreiben lässt, die aber paarweise disjunkt sein müssen.

Mit der disjunkten Zerlung von Ω in $A_1, A_2, \dots, A_n$. Damit können wir die Wahrscheinlichkeit für das Eintreten von B anhand der Rechenregeln für Wahrscheinlichkeiten berechnen als

$$P(B) = P(B|A_1)P(A_1) + P(B|A2)P(A_2) + \dots + P(B|A_n)P(A_n).$$

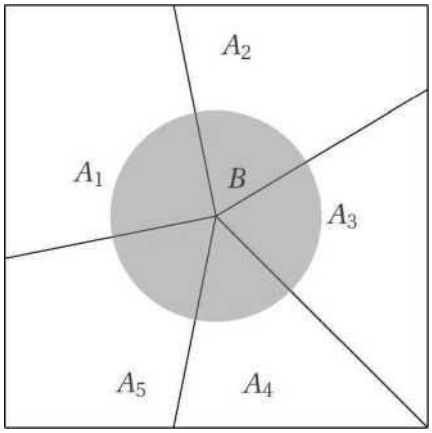


Abb. 11: Darstellung einer Menge A über ihre disjunkte Zerlegung

5. Diskrete Zufallsvariable

Eine diskrete Zufallsvariable ist eine Zufallsvariable, die nur endlich viele oder abzählbar viele Werte annehmen kann, zum Besipiel der Wurf mit einem Würfel oder die Anzahl der Kreditausfälle einer Bank in einem Jahr.

5.1 Verteilung einer diskreten Zufallsvariable

Die Werte einer diskreten Zufallsvariable werden mit festen Wahrscheinlichkeiten angenommen. Diese entsprechen den Wahrscheinlichkeiten der zugrunde liegenden Elementarereignisse. In der Regel werden für Zufallsvariablen die zugrunde liegenden Elementarereignisse gar nicht mehr betrachtet, sondern es werden direkt die Wahrscheinlichkeiten für die Zufallsvariable angegeben.

5.1.1 Wahrscheinlichkeitsfunktion

Ordnet man jedem möglichen Ergebnis einer diskreten Zufallsvariable eine Wahrscheinlichkeit zu, so können daraus auch die Wahrscheinlichkeiten für beliebige Ereignisse bestimmt werden. Somit ist die Zufallsvariable vollständig beschrieben. Die Zuordnung von Wahrscheinlichkeiten zu den möglichen Werten, die angenommen werden können, wird Wahrscheinlichkeitsfunktion genannt.

Wahrscheinlichkeitsfunktion

Sei *X* eine diskrete Zufallsvariable. Die Funktion

$$f(x) = \begin{cases} P(X = x_i) = p_i, & \text{für } x_i \in \{x_1, x_2, \dots, x_k, \dots\} \\ 0, & sonst \end{cases}$$

heißt Wahrscheinlichkeitsfunktion von *X*. Es gilt:

$$\sum_{i=1}^{n} f(x_i) = 1$$

Beispiel Wahrscheinlichkeitsfunktion beim Münzwurf

- Wir ordnen einem Münzwurf einer Zufallsvariable *X* zu. *X* bezeichne die Anzahl von „Kopf“ bei diesem Versuch. Dann kann *X* zwei Werte annehmen, nämlich

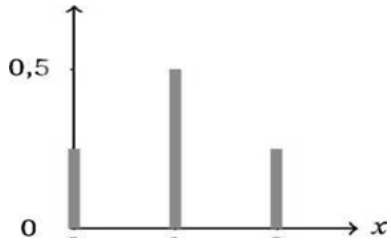
  X(„Zahl“) = 0 und X(„Kopf“) = 1

  mit den Wahrscheinlichkeiten

  P(X = 0) = P(„Zahl“) = 1/2 und *P(X*= l) = P(„Kopf“) = 1/2.

- Es wird zweimal eine Münze geworfen. *X* soll der Anzahl entsprechen, in denen „Zahl“ geworfen wird. Dann nimmt *X* die Werte x = 0,1,2 an. Die Wahrscheinlichkeitsfunktion von *X* lautet:

$$f(x) = \begin{cases} \frac{1}{4}, & f\ddot{u}r\ x = 0 \\ \frac{1}{2}, & f\ddot{u}r\ x = 1 \\ \frac{1}{4}, & f\ddot{u}r\ x = 2 \\ 0, & sonst \end{cases}$$



Diskrete Gleichverteilung

Eine diskrete Zufallsvariable X mit den Realisierungen $x_1, x_2, \dots, x_N$ für die gilt

$$P(X = x_i) = \frac{1}{N} \quad i = 1, \dots, N$$

nennt man gleichverteilt.

Bernoulli-Verteilung

Eine weitere elementare Verteilung ist die Bernoulli-Verteilung. Ihr liegt ein sogenanntes Bernoulli-Experiment zugrunde. Bei diesem Experiment gibt es nur zwei mögliche Ausgänge (Ergebnisse) A und $\bar{A}$ (z. B. Kopf/Zahl, Junge/Mädchen, fehlerhaft/fehlerfrei, Erfolg/Misserfolg), die mit den Wahrscheinlichkeiten

$$P(A) = p \quad und \quad P(\bar{A}) = 1 - p$$

eintreten.

Man bezeichnet p häufig als Erfolgswahrscheinlichkeit und schreibt

$$X \sim B(1, p).$$

Geometrische Verteilung

Wird ein Bernoulli-Experiment, mit $P(A) = p$, öfter unabhängig wiederholt, so entspricht dies einer Reihe von Bernoulli-Experimenten. Darauf baut eine weitere elementare Verteilung, die geometrische Verteilung. Es wird dabei das Experiment so oft unabhängig wiederholt, bis zum ersten Mal A eintritt. Als Zufallsvariable definieren wir:

$$X = \text{„}Anzahl\ der\ Versuche, bis\ zum\ ersten\ Mal\ A\ eintritt\text{“}.$$

Die Zahl der notwendigen Versuche, bis das Ereignis A eintritt, kann sehr groß werden, hierfür gibt es keine obere Grenze. Daher kann X mit positiver Wahrscheinlichkeit jeden der Werte $x = 1, 2, 3, 4, \dots$ annehmen.

$$P(X = k) = p * (1 - p)^{k-1}$$

Die obige Formel lässt sich interpretieren: Falls der erste Erfolg erst nach k Versuchen erreicht wird, impliziert dies erst $k - 1$ Misserfolge und dann einen Erfolg.

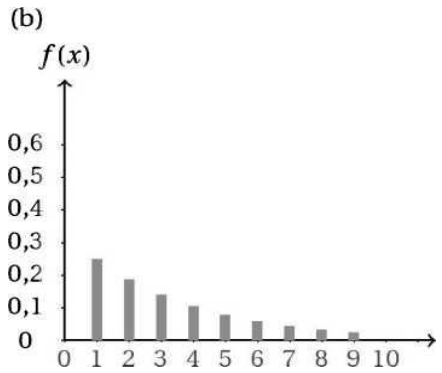


Abb. 12: Beispiel einer geometrischen Verteilung mit dem Paramter 0,25.

Beispiel: Geometrische Verteilung

Der Produktionsleiter einer Firma hat ermittelt, dass die Lieferanten die vereinbarten Fristen im Mittel bei 85% der Bestellungen einhalten. Die Firma hat mit einem neuen Lieferanten laufende Zusendungen von Halbfertigerzeugnissen für die Herstellung eines Produktes vereinbart. Nachdem der Lieferant drei Mal fristgerecht geliefert hat, ist er bei der vierten Zusendung in Verzug geraten. Der Produktionsleiter möchte nun wissen, mit welcher Wahrscheinlichkeit ein solches Verhalten zu erwarten ist.

Die Zufallsvariable X ist hier die Anzahl der fristgerechten Lieferungen (Ereignis $\bar{A}$), sodass $P(\bar{A}) = 1 - p = 0{,}85$ ist. Gesucht ist die Wahrscheinlichkeit, dass nach drei fristgerechten erstmals keine fristgerechte Lieferung (Ereignis A) erfolgt. Dies ist die Wahrscheinlichkeit dafür, dass die geometrisch verteilte Zufallsvariable X den Wert 4 annimmt.

$$P(X = 4) = f(4) = (1-p)^{4-1} \bullet p^1 = 0{,}85^3 * 0{,}15^1 = 0{,}092.$$

Beispiel:

Ein Produkt wird auf dem Fließband gefertigt. Dabei kommt es mit einer Wahrscheinlichkeit von 0,03 zu einem fehlerhaften Produkt. Die Wahrscheinlichkeit dafür, dass unter drei infolge gefertigter Produkte mindestens eines nicht funktionsfähig ist, kann mithilfe der geometrischen Verteilung bestimmt werden.

$$\begin{aligned} P(X \leq 3) &= P(X = 1) + P(X = 2) + P(X = 3) \\ &= 0{,}03 * 0{,}97^0 + 0{,}03 * 0.97^1 + 0{,}03 * 0{,}97^2 \\ &= 0{,}087. \end{aligned}$$

5.1.2 Verteilungsfunktion

Neben der Wahrscheinlichkeitsfunktion kann eine diskrete Zufallsvariable auch mithilfe der Verteilungsfunktion eindeutig und vollständig beschrieben werden. Die Verteilungsfunktion gibt die kumulierten Wahrscheinlichkeiten $P(X < x)$ einer Zufallsvariable X an. In der Anwendung sind oft Wahrscheinlichkeiten für zusammenhängende Zahlenbereiche gefragt. Solche Intervalle lassen sich mithilfe der Verteilungsfunktion leicht bestimmen.

Verteilungsfunktion einer diskreten Zufallsvariable

Es sei X eine diskrete Zufallsvariable und x eine reelle Zahl. Die Funktion

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

heißt Verteilungsfunktion von X.

Die Verteilungsfunktion hat die folgenden Eigenschaften:

- Der kleinste Wert von $F(x)$ ist null, der größte Wert eins, d. h. es gilt $0 \leq F(x) \leq 1$ für alle x.
- $F(x)$ ist eine Treppenfunktion, die an den Stellen x_i, den Realisierungen der Zufallsvariable X, um den Wert $f(xi)$ nach oben springt. Zwischen den Realisierungen verläuft sie konstant, d.h. parallel zur x-Achse.
- Für $x < x_1$, wobei x_1 die kleinste Realisierung sei, ist $F(x) = 0$ und für endliche diskrete
- Zufallsvariable mit größtem Wert x_N ist $F(x) = 1$ für $x \geq x_N$.

Beispiel: Verteilungsfunktion beim fairen Würfeln

$$F(x) = \begin{cases} 0 & x < 1 \\ \frac{1}{6} & 1 \leq x < 2 \\ \frac{1}{3} & 2 \leq x < 3 \\ \frac{1}{2} & 3 \leq x < 4 \\ \frac{2}{3} & 4 \leq x < 5 \\ \frac{5}{6} & 5 \leq x < 6 \\ 1 & x \geq 6 \end{cases}$$

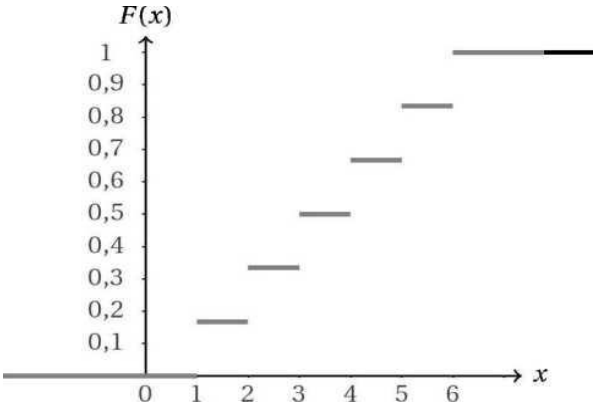


Abb. 13: Verteilungsfunktion beim fairen Würfel

Beispiel: Verteilungsfunktion einer geometrischen Verteilung

Ist X ~ *G(p)* verteilt, so besitzt *X* die Verteilungsfunktion

$$F(x) = \begin{cases} 0 & \text{für } x < 1 \\ 1 - (1 - p)^x & \text{für } x = 1,2,3,\ldots \end{cases}$$

Berechnung von Wahrscheinlichkeiten mithilfe der Verteilungsfunktion

Für beliebige reelle Werte *a* und *b* gilt:

- $P(X \leq a) = F(a)$ (gemäß Definition)
- $P(X < a) = F(a) - P(X = a)$
- $P(X > a) = 1 - F(a)$
- $P(X \geq a) = 1 - F(a) + P(X = a)$
- $P(a \leq X \leq b) = F(b) - F(a) + P(X = a)$
- $P(a < X \leq b) = F(b) - F(a)$
- $P(a < X < b) = F(b) - F(a) - P(X = b)$
- $P(a \leq X < b) = F(b) - F(a) + P(X = a) - P(X = b)$

5.2 Unabhängigkeit von diskreten Zufallsvariablen

Oft lassen sich reale Vorgänge durch einen Zufallsvorgang beschreiben, der sich aus einzelnen Zufallsvariablen zusammensetzt. Aus Sicht der Wahrscheinlichkeitsrechnung interessiert man sich für die gemeinsame Verteilung mehrerer oder einer Vielzahl von Zufallsvariablen. Nimmt man an, dass die einzelnen Zufallsvariablen unabhängig voneinander sind, sich also gegenseitig nicht beeinflussen, so lässt sich die gemeinsame Verteilung einfach bestimmen.

Unabhängigkeit von diskreten Zufallsvariablen

n diskrete Zufallsvariable $X_1, X_2, \dots, X_n$ sind unabhängig, wenn

$$P(X_1 = x_1, X_2 = x_2, \dots, X_n = x_n) = P(X_1 = x_1) * P(X_2 = x_2) * \dots * P(X_n = x_n)$$

für beliebige Werte $x_1, x_2, \dots, x_n$ aus der Menge der jeweiligen Elementarereignisse gilt.

Es gilt dann insbesondere für zwei auch für beliebige Zufallsvariablen

$$P(X_{i1} = x_{i1}, X_{i2} = x_{i2}) = P(X_{i1} = x_{i1}) * P(X_{i2} = x_{i2})$$

Beispiel: Verträge

In einem Fitnessstudio laufen drei Verträge aus. Erfahrungsgemäß sei bekannt, dass ein Vertrag in 60 % der Fälle nicht verlängert wird. Den drei Verträgen wird jeweils eine Bernoulli-Variable X_1, X_2, X_3 zugeordnet. Im Falle der Verlängerung nimmt X_i den Wert 1 an, sonst den Wert 0. Wie groß ist die Wahrscheinlichkeit, dass alle drei Verträge verlängert werden? Unter Annahme der Unabhängigkeit gilt:

$$P(X_1 = 1, X_2 = 1, X_3 = 1) = P(X_1 = 1) * P(X_2 = 1) * P(X_3 = 1) = 0{,}4 * 0{,}4 * 0{,}4 = 0{,}064$$

5.3 Parameter von diskreten Zufallsvariablen

Analog zu den Lage- und Streuungsparametern in der deskriptiven Statistik, die die Ausprägungen eines Merkmals zusammengefasst haben, lassen sich auch Verteilungen mittels Parameter zusammenfassen. In diesem Abschnitt werden der Erwartungswert, der dem arithmetischen Mittel in der deskriptiven Statistik entspricht, und die Varianz, das Analogon zur empirischen Varianz, behandelt. Die Bestimmung beider Parameter ist der Berechnung, wie wir sie im Rahmen der deskriptiven Statistik kennengelernt haben, sehr ähnlich. An die Stelle der relativen Häufigkeiten treten in der Wahrscheinlichkeitsrechnung die Wahrscheinlichkeiten. Die Interpretation der Parameter entspricht dem Vorgehen in der deskriptiven Statistik. So charakterisiert der Erwartungswert das Zentrum einer Verteilung und die Varianz das Ausmaß der Streuung um den Erwartungswert.

5.3.1 Erwartungswert

Der Erwartungswert $E(X)$ einer Zufallsvariable X fasst die Verteilungsfunktion bzw. die Wahrscheinlichkeitsfunktion in einer Zahl, die die Lage oder den Schwerpunkt der Verteilung beschreibt, zusammen. Der Erwartungswert kann auch der erwartete Wert der Verteilung genannt werden. Er entspricht dem Durchschnitt aller Werte, die angenommen werden können, gewichtet mit den jeweiligen Wahrscheinlichkeiten.

Erwartungswert bei diskreten Zufallsvariablen

Ist X eine diskrete Zufallsvariable mit den Realisierungen $x_1, \dots, x_k, \dots$ und den zugehörigen Wahrscheinlichkeiten $f(x_i)$ so wird der Erwartungswert von X definiert als

$$E(X) = \sum_{i \geq 1} x_i * f(x_i).$$

Bemerkung:
Der Erwartungswert ist ebenso wie das arithmetische Mittel als Lageparameter der Verteilung anzusehen. Er ist ein Maß für die zentrale Tendenz der Verteilung. Man kann sich zur Veranschaulichung vorstellen, dass der Erwartungswert dem Mittelwert der Ergebnisse entspricht, die auftreten, wenn das Zufallsexperiment, das X beschreibt, immer und immer wieder ausgeführt wird.

Rechenregeln für den Erwartungswert

Sind a und b beliebige reelle Zahlen und sind X und Y zwei diskrete Zufallsvariablen, so gilt:

- **Linearität des Erwartungswertes**
  - $E(a) = a$
  - $E(b * X) = b * E(X)$
  - $E(b*X + a) = b*E(X) + a$
- **Additivität des Erwartungswertes**
  - $E(X+Y) = E(X) + E(Y)$.
- **Produktregel**
  - Sind X und Y *unabhängig*, dann gilt: $E(X*Y) = E(X)*E(Y)$.

5.3.2 Varianz

Neben dem Erwartungswert gehört die Varianz zu den zentralen Parametern, die eine diskrete Zufallsvariable charakterisieren. Die Varianz drückt die Stärke der Streuung der Zufallsvariablen um den Erwartungswert aus. Die Interpretation erfolgt analog zur Varianz in der deskriptiven Statistik. Auch berechnet sich die Varianz ähnlich, nämlich indem die relativen Häufigkeiten durch die Wahrscheinlichkeiten ersetzt werden.

Varianz und Standardabweichung einer diskreten Zufallsvariable

Ist X eine diskrete Zufallsvariable mit den Realisierungen $x_1, \dots, x_k, \dots$ und den zugehörigen Wahrscheinlichkeiten $f(x_i)$, so wird die Varianz von X definiert als

$$V(X) = \sum_{i \geq 1} (x_i - E(X))^2 * f(x_i).$$

Die Wurzel aus der Varianz

$$\sigma = \sqrt{Var(X)}$$

wird als Standardabweichung von X bezeichnet.

Bemerkung:
Die Varianz einer Zufallsvariable ist der empirischen Varianz s^2 von statistischen Variablen nachgebildet und ebenso wie diese als ein Streuungsparameter der Verteilung anzusehen. Deshalb hat sie auch Eigenschaften, die denen von s^2 entsprechen.

Verschiebungssatz der Varianz

Die Varianz einer Zufallsvariable lässt sich alternativ als

$$Var(X) = E(X2) - [E(X)]^2$$

schreiben.

Rechenregeln für die Varianz

Sind a und b beliebige reelle Zahlen und sind X und Y zwei diskrete Zufallsvariablen, so gilt:

- **lineare Transformationen**
  - $\mathrm{Var}(X + a) = \mathrm{Var}(X)$
  - $\mathrm{Var}(b*X) = b^2*\mathrm{Var}(X)$
  - $\mathrm{Var}(b*X + a) = b^2*\mathrm{Var}(X)$
- **Additivität der Varianz bei Unabhängigkeit**
  - Sind X und Y unabhängige Zufallsvariablen, so gilt: $Var(X + Y) = Var(X) + Var(Y)$.

5.4 Spezielle diskrete Verteilunge

5.4.1 Die Binomialverteilung

Die Binomialverteilung dient der Ermittlung von Wahrscheinlichkeiten für die Häufigkeit des Eintretens bestimmter Ereignisse bei speziellen Experimenten. Beispiele sind die Anzahl an Garantiefällen bei IT-Geräten, die Anzahl defekter Stücke bei einer Produktion oder die Anzahl von Kreditausfällen bei Banken.
Ausgangspunkt einen Binomialverteilung ist ein Bernoulli-Experiment. Bei einem Bernoulli Experiment gibt es nur zwei mögliche Ausgänge (Ergebnisse) A und $\bar{A}$ (z. B. Kopf/Zahl, Junge/Mädchen, fehlerhaft/fehlerfrei, Erfolg/Misserfolg). Für die Wahrscheinlichkeiten dieser Ereignisse gilt:

$$P(A) = p \quad und \quad P(\bar{A}) = 1 - p \; mit \; 0 \leq p \leq 1.$$

Man bezeichnet *p* häufig als Erfolgswahrscheinlichkeit.

Das Bernoulli-Experiment wird nun *n*-mal unter gleichen Bedingungen wiederholt. Damit sind die Wiederholungen voneinander unabhängig. Unabhängigkeit bedeutet in diesem Zusammenhang, dass die Erfolgswahrscheinlichkeit eines Experimentes nicht durch die anderen Experimente beeinflusst wird. Die Anzahl *X* des Eintretens des Ereignisses *A* bei diesen *n* unabhängig voneinander durchgeführten Bernoulli-Experimenten ist eine Zufallsvariable:

X = Anzahl der Erfolge, *x* = 0,1,2,3, ..., *n*.

X heißt binomialverteilt. Man schreibt

$$X \sim B(n, p).$$

Die Wahrscheinlichkeitsfunktion von *X* lautet:

$$f(x) = \begin{cases} \binom{n}{x} p^x (1-p)^{n-x} & für\ x = 0,1,2,\dots,n \\ 0, & sonst \end{cases}$$

Die Binomialverteilung gibt also die Wahrscheinlichkeiten an, dass bei einer Summe von n Ereignissen eine bestimmte Anzahl *x* an Treffern realisiert werden.

Mit $n = 1$ sind auch die Bernoulli-Verteilungen Mitglieder dieser Familie.

Verteilungsfunktion einer Binomialverteilung

Ist $X \sim B(n, p)$ verteilt, so besitzt *X* die Verteilungsfunktion:

$$F(x) = \begin{cases} \sum_{k=0}^{x} \binom{n}{k} p^k (1-p)^{n-k}, & für\ x \geq 0 \\ 0, & für\ x < 0 \end{cases}$$

Erwartungswert und Varianz der Binomialverteilung

einer $X \sim B(n, p)$-verteilten Zufallsvariable X

$$E(X) = np \ und \ V(X) = np(1-p).$$

Additivität zweier Binomialverteilungen mit gleichem Parameter p

Sind X und Y zwei binomialverteilte Zufallsvariablen mit gleichem Parameter p, d.h. $X \sim B(n_X, p)$ und $Y \sim B(n_Y, p)$ so gilt:

$$X + Y \sim B(n_X + n_Y, p).$$

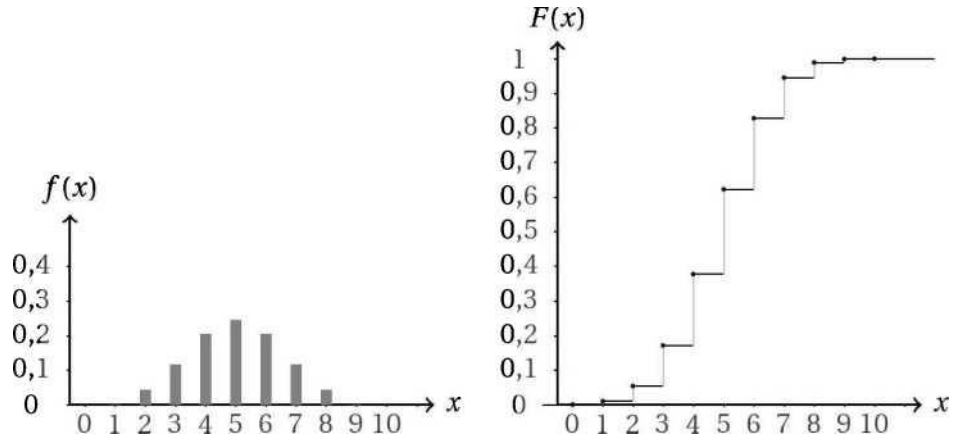


Abb. 14: Wahrscheinlichkeitsfunktion und Verteilungsfunktion von $X \sim B(10; 0,5)$

Beispiel:

Ein Werk für Solaranlangen hat eine Ausschussquote von 20 % bei den gefertigten Solarmodulen. Wie wahrscheinlich ist es, dass unter 10 Solarmodulen

a) alle 10 Module voll funktionsfähig sind?

b) höchstens 3 Module nicht in vollem Maße funktionsfähig sind?

Die Anzahl der defekten Stücke soll durch eine Zufallsvariable $X \sim B(10;0,2)$ abgebildet werden.

Für a) berechnet man:

$$P(X = 0) = \binom{10}{0} 0{,}2^x (1-0{,}2)^{10-0} = 0{,}107.$$

Für b) ergibt sich:

$$P(X \leq 3) = F(3) = P(X = 0) + P(X = 1) + P(X = 2) + P(X = 3)$$
$$= 0{,}107 + 0{,}268 + 0{,}302 + 0{,}201 = 0{,}879.$$

5.4.2 Die Poisson-Verteilung

Durch die Binomialverteilung wird die Anzahl der eingetretenen Erfolgsereignisse *A* bei einer festen Anzahl von Bernoulli-Experimenten beschrieben. Ist die Anzahl der Versuche sehr groß oder nach oben nicht beschränkt und die Wahrscheinlichkeit für den Eintritt des Erfolgsereignisse *A* sehr klein, kann die Poisson-Verteilung anstelle der Binomialverteilung verwendet werden. Dies ergibt sich aus der Tatsache, dass sich eine B(n, *p)*-Binomialverteilung für großes *n* und kleines *p* der Poisson-Verteilung annähert.
Die Poisson-Verteilung eignet sich daher ebenfalls zur Modellierung von Zählvorgängen. Dabei werden bestimmte Ereignisse gezählt, die innerhalb eines festen, vorgegebenen Zeitintervalls eintreten können. Beispiele sind die die Anzahl der Erkrankungen an einer seltenen Krankheit in einem Monat, die Schadensfälle in einer Versicherung in einer Periode oder die Anfragen an einen Webserver in einem Zeitfenster. Wichtig bei der Anwendung der Poisson-Verteilung ist, dass das Auftreten der einzelnen Ereignisse voneinander unabhängig ist. Auch liegt die Annahme zugrunde, dass sich die Wahrscheinlichkeit für das Eintreten eines Ereignisses gleichmäßig über das Zeitintervall verteilt und die Ereignisse nacheinander und nicht gleichzeitig auftreten.
Die Poisson-Verteilung wird auch als Verteilung für seltene Ereignisse bezeichnet. Sie ist wie folgt definiert:
Die Wahrscheinlichkeitsfunktion Poisson-Verteilung mit dem Parameter λ lautet:

Poisson-Wahrscheinlichkeitsfunktion

$$f(x) = \begin{cases} \frac{\lambda^x}{x!} * \exp(-\lambda) & \text{für } x = 0,1,2,\dots,n \\ 0, & sonst. \end{cases}$$

Man schreibt kurz:

$$X \sim P(\lambda).$$

Verteilungsfunktion einer Poisson-Verteilung

Ist $X \sim P(\lambda)$ verteilt, so besitzt X die Verteilungsfunktion

$$F(x) = \begin{cases} \exp(-\lambda) * \sum_{k=0}^{x} \frac{\lambda^k}{k!} & , \quad \text{für } x \geq 0 \\ 0, & \text{für } x < 0. \end{cases}$$

Erwartungswert und Varianz der Poisson-Verteilung

Ist $X \sim P(\lambda)$ verteilt, dann gilt:

$$E(X) = V(X) = \lambda.$$

Additionssatz der Poi sson-Verteilung

Sind $X \sim P(\lambda_1)$ und $Y \sim P(\lambda_2)$ unabhängige Zufallsvariablen, dann gilt: $X + y \sim P(\lambda_1 + \lambda_2)$

.

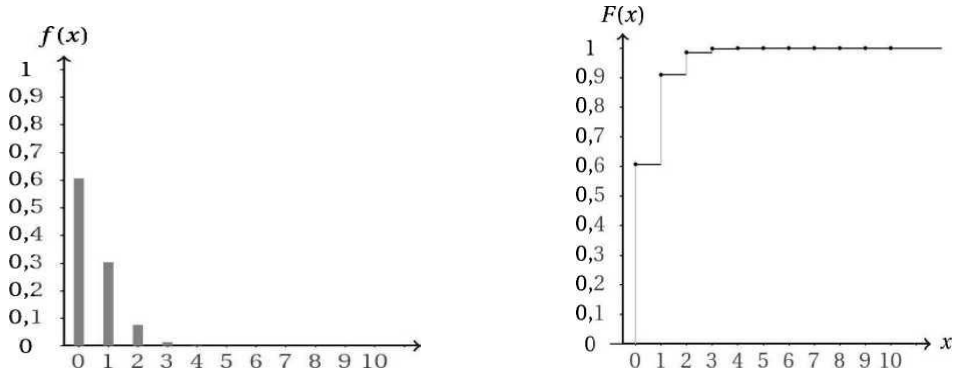


Abb. 15: Wahrscheinlichkeits- und Verteilungsfunktion der $P(0,5)$-Verteilung.

Die Zufallsvariable *X* der Poisson-Verteilung ist definiert als Zahl der Erfolge bei „sehr vielen" Bernoulli-Experimenten mit „sehr kleiner" Erfolgswahrscheinlichkeit. Die Poisson-Verteilung ist somit ein Grenzfall der Binomialverteilung.

Beispiel

Bei einer Hotline weiß man aus Erfahrung, dass freitags zwischen 14 und 15 Uhr im Durchschnitt 7 Kunden den Dienst in Anspruch nehmen. Die Wahrscheinlichkeit dafür, dass die Anzahl der Kunden, die die Hotline anrufen, 9 beträgt, ist

$$P(X = 9) = \frac{7^9}{9!} * \exp(-7) = 0{,}1014.$$

5.4.3 Die hypergeometrische Verteilung

Die hypergeometrische Verteilung gibt über die Häufigkeit eines Merkmals in einer Stichprobe Auskunft. Weisen in einer Grundgesamtheit vom Umfang *N*, *M* Elemente ein bestimmtes Merkmal auf, so gibt die hypergeometrische Verteilung die Wahrscheinlichkeit dafür an, dass sich in einer Stichprobe vom Umfang *n*, *k* Elemente mit dem entsprechenden Merkmal befinden. So kann mithilfe der Verteilung zum Beispiel abgeschätzt werden, mit welcher Wahrscheinlichkeit eine Stichprobe die Grundgesamtheit realistisch widerspiegelt. Darüber hinaus hat die hypergeometrische Verteilung weitere Anwendungen zum Beispiel in der Qualitätskontrolle.

Hypergeometrische Verteilung

In einer Population befinden sich *N* Einheiten, von denen *M* das gewünschte Merkmal aufweisen. Werden von dieser Population n Elemente ohne Zurücklegen gezogen und interessiert man sich für die Anzahl *X*der mit Merkmal *M* gezogenen Einheiten, so gibt die hypergeometrische Verteilung die Wahrscheinlichkeit dafür an.

Kurzschreibweise: $X \sim H(N, M, n)$

$$P(X = x) = \frac{\binom{M}{x}\binom{N-M}{n-x}}{\binom{N}{n}}$$

Erwartungswert und Varianz der hypergeometrischen Verteilung

$$E(X) = n * \frac{M}{N} \quad und \quad Var(X) = n * \frac{M}{N} * \left(1 - \frac{M}{N}\right) * \left(\frac{N-n}{N-1}\right) .$$

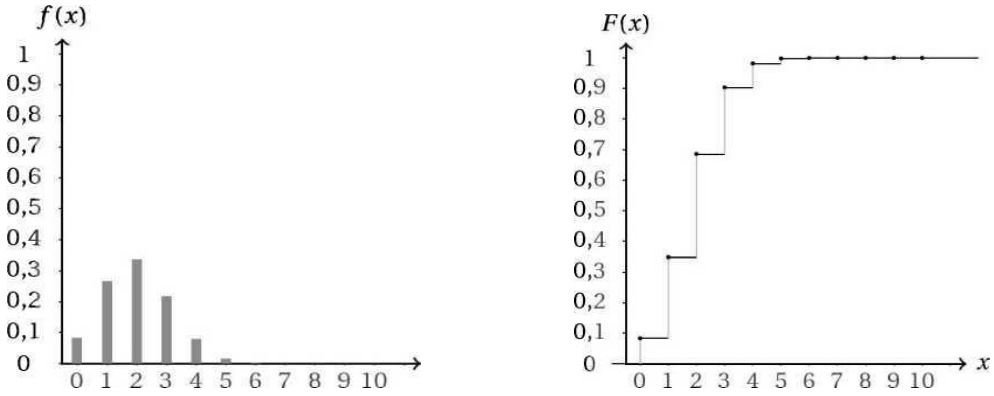


Abb. 16: Wahrscheinlichkeits- und Verteilungsfunktion einer $H(\mathbf{50}, \mathbf{10}, \mathbf{10})$-Verteilung

Beispiel:

In einer Grundgesamtheit aus 100 Elementen, weist die Hälfte ein bestimmtes Merkmal auf. Für eine Stichprobe vom Umfang 10 sollen zwei Fragen beantwortet werden.

a) Wie wahrscheinlich ist es, genau einen Anteil von 50 % der Elemente mit dem Merkmal zu ziehen?

b) Wie wahrscheinlich ist es, einen Anteil zwischen 40 % und 60 % der Elemente mit dem Merkmal zu ziehen?

Zu a)

$$P(X = 5) = \frac{\binom{50}{5}\binom{50}{5}}{\binom{100}{10}}$$

Zu b)

$$P(X = 4) + P(X = 5) + P(X = 6) = \frac{\binom{50}{4}\binom{50}{6}}{\binom{100}{10}} + \frac{\binom{50}{5}\binom{50}{5}}{\binom{100}{10}} + \frac{\binom{50}{6}\binom{50}{4}}{\binom{100}{10}} = 0{,}682$$

6. Stetige Zufallsvariable

In diesem Kapitel wird der Begriff der Zufallsvariable auf Zufallsgrößen, die kontinuierliche Werte annehmen können, erweitert. Im Gegensatz zu den diskreten Zufallsvariablen, die nur bestimmte Werte annehmen, nehmen stetige Zufallsvariable beliebige Werte aus einem Intervall oder den reellen Zahlen an. Im Rahmen der deskriptiven Statistik wurden stetige Merkmale, wie in der Betriebswirtschaft der Umsatz, der Gewinn, Ausgaben oder Preise darstellen, behandelt. Mittels stetiger Zufallsvariablen können die Verteilungen solcher stetigen Merkmale beschrieben werden. Der Umgang mit stetigen Zufallsvariablen gestaltet sich oft ähnlich wie bei den diskreten Zufallsvariablen. So gelten zum Beispiel sämtliche Rechenregeln für Erwartungswert und Varianz auch hier. Der wesentliche Unterschied zu den diskreten Zufallsvariablen ist, dass die sogenannte Dichtefunktion an die Stelle der Wahrscheinlichkeitsfunktion tritt. Der Hintergrund dafür ist der, dass hinter kontinuierlichen Werten so viele verschiedene Werte stehen, dass jeder einzelne Wert eine Wahrscheinlichkeit von null hat und somit die Wahrscheinlichkeitsfunktion nicht mehr anwendbar ist.

6.1 Definition und Verteilung

Eine wichtige Eigenschaft stetiger Zufallsvariablen ist, dass sich die Wahrscheinlichkeit über kontinuierliche Bereiche verteilt. Dabei können verschiedene Zahlenbereiche größere Wahrscheinlichkeiten annehmen. Dies kann durch die sogenannte Dichtefunktion beschrieben werden. Intuitiv kann die Höhe der Dichte als die Höhe der Wahrscheinlichkeit für die Bereiche interpretiert werden.

6.1.1 Dichtefunktion

Eine stetige Zufallsvariable *X* kann durch eine reelle Funktion *f(x)* mit folgenden Eigenschaften beschrieben werden:

- $f(x) \geq 0$ (Nichtnegativität)
- $\int_{-\infty}^{\infty} f(x)dx = 1$ (Normiertheit)

Die Fläche unter der Funktion ist eins.

$f(x)$ heißt Dichtefunktion oder Dichte von *X*.

Beispiel:

Ein Student stellt sich ohne auf die Uhr zu schauen an eine Bushaltestelle. Der Bus, mit dem er fahren möchte, kommt alle 10 Minuten. Die Wartezeit wird sich dann gleichmäßig auf 10 Minuten aufteilen. Dies kann durch die folgende Dichtefunktion beschrieben werden:

$$f(x) = \begin{cases} 0 & \text{für } x < 0 \\ 0{,}1 & \text{für } x \in [0{,}10) \\ 0 & \text{für } x \geq 10 \end{cases}$$

Die angegebene Funktion ist stets > 0. Zudem prüft man leicht nach, dass die Fläche unter der Funktion eins ergibt.

Beispiel 7.2

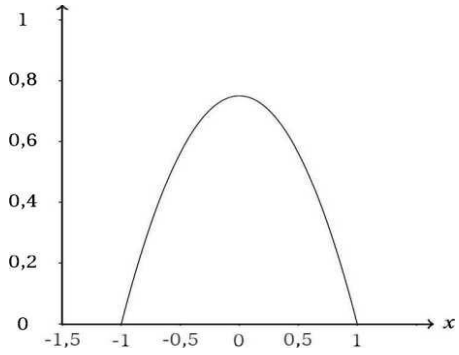
Eine Maschine fertigt Werkstücke mit einer bestimmten Präzision. Die Abweichungen von der Sollgröße sind nach oben und unten auf 1 mm begrenzt. Geringe Abweichungen kommen häufiger vor. Möglicherweise lassen sich die Abweichungen durch die folgende Funktion beschreiben:

$$f(x) = \begin{cases} 0 & f\ddot{u}r\ x < -1 \\ \frac{3}{4}(1-x)^2 & f\ddot{u}r\ x \in [-1,1] \\ 0 & f\ddot{u}r\ x > 1 \end{cases}$$

Die angegebene Funktion ist stets > 0. Für die Fläche berechnet sich:

$$\int_{-\infty}^{\infty} f(x)dx = \int_{-1}^{1} \frac{3}{4}(1-x^2)dx = \frac{3}{4}\left[x - \frac{x^3}{3}\right]_{-1}^{1} \quad \underline{= 1}$$

Somit handelt es sich um eine Dichte. Die Dichte ist im Folgenden abgebildet. Man sieht, dass sie für x-Werte näher an der Null höhere Werte annimmt. Dies entspricht den höheren Wahrscheinlichkeiten von geringen Abweichungen vom Sollwert.



Durch eine Dichte ist stets auch eine stetige Zufallsvariable gegeben. Die Wahrscheinlichkeit, dass die Zufallsvariable einen Wert in einem bestimmten Bereich annimmt, entspricht der Fläche des Bereichs unter der Dichte.

Stetige Zufallsvariable

Zu einer Dichte kann stets eine stetige Zufallsvariable angegeben werden. Die Wahrscheinlichkeit für ein Intervall ist durch

$$P(a \leq X \leq b) = \int_a^b f(x)dx$$

gegeben.

Insbesondere sind dann die Wahrscheinlichkeiten für einen einzelnen Wert *a* gleich null, da

$$\int_a^a f(x)dx = 0$$

gilt.

Beispiel:

Es soll berechnet werden, wie wahrscheinlich es im obigen Beispiel ist, dass der Student höchstens 5 Minuten auf den Bus wartet.

$$P(-\infty < X \leq 5) = \int_{-\infty}^{5} f(x)dx = \int_0^5 0{,}1\ dx = 0{,}5.$$

Definition einer stetigen Zufallsvariable

Eine Zufallsvariable *X* ist stetig, wenn sich ihre Wahrscheinlichkeiten durch eine Dichte angeben lassen:

$$P(a \leq X \leq b) = \int_a^b f(x)dx$$

Die im ersten Beispiel (bzw. obigem Beispiel) verwendete Verteilung nennt man stetige Gleichverteilung. Sie hat die Eigenschaft, dass sich die Wahrscheinlichkeit gleichmäßig auf einen Bereich verteilt. Somit ist die Wahrscheinlichkeit für alle Bereiche gleicher Länge innerhalb des Intervalls der Gleichverteilung identisch.

Stetige Gleichverteilung

Ist *X* eine Zufallsvariable, die Werte aus einem festen Intervall *[a; b]* annimmt, und auf dem Intervall eine gleichmäßige Dichte besitzt, dann ist *X* stetig gleich verteilt. Die Dichte ist durch

$$f(x) = \begin{cases} 0 & \text{für } x < a \\ 1/(b-a) & \text{für } x \in [a,b] \\ 0 & \text{für } x > b \end{cases}$$

gegeben. Als Kurzschreibweise für eine stetig gleichverteilte Zufallsvariable *X* auf *[a, b]* verwendet man: $X \sim U(a,b)$.

6.1.2 Verteilungsfunktion

Neben der Dichtefunktion, kann eine stetige Zufallsvariable auch über ihre Verteilungsfunktion eindeutig und vollständig beschrieben werden. Die Definition der Verteilungsfunktion ist mit der der diskreten Zufallsvariablen identisch.

Verteilungsfunktion stetiger Zufallsvariablen

Die Verteilungsfunktion einer Zufallsvariablen X ist definiert durch

$$F(x) = P(X \leq x) = P(-\infty \leq X \leq x).$$

Die Verteilungsfunktion kann durch die Dichtefunktion berechnet werden. Es gilt:

$$F(x) = P(X \leq x) = P(-\infty \leq X \leq x) = \int_{-\infty}^{x} f(u)du$$

Daraus ergibt sich die Eigenschaft, dass die Ableitung der Verteilungsfunktion (an Stellen, an denen sie differenzierbar ist,) mit der Dichte übereinstimmt:

6.2 Unabhängigkeit von stetigen Zufallsvariablen

Stetige Zufallsvariablen heißen wie diskrete Zufallsvariablen unabhängig, wenn die Unabhängigkeit für beliebige Ereignisse der einzelnen Zufallsvariablen gilt. Dies folgt aber bereits, wenn die Unabhängigkeit für beliebige Intervalle der Form $(-\infty, x]$ erfüllt ist.

Unabhängigkeit von stetigen Zufallsvariablen

Zwei stetige Zufallsvariablen X_1, X_2 heißen unabhängig, wenn für beliebige reele Zahlen x_1 und x_2

$$P(X_1 \leq x_1 \; und \; X_2 \leq x_2) = P(X_1 \leq x_1) * P(X_2 \leq x_2)$$

gilt.

Sind zwei Zufallsvariablen unabhängig, so folgt für beliebige Ereignisse $A_1 \; und \; A_2$

$$P(X_1 \in A_1 \; und \; X_2 \in A_2) = P(X_1 \in A_1) * P(X_2 \in A_2)$$

6.3 Parameter von stetigen Zufallsvariablen

In diesem Abschnitt wird auf wichtige Parameter stetiger Zufallsvariablen eingegangen. Dies sind der Erwartungswert, die Varianz. Die Interpretation des Erwartungswerts und der Varianz entspricht diskreten Zufallsvariablen. Der Erwartungswert drückt den Schwerpunkt einer Verteilung aus, während die Varianz die Streuung in Form der erwarteten quadratischen Abweichung vom Mittelwert angibt. Die Berechnung der Parameter trägt der Tatsache Rechnung, dass es sich bei stetigen Zufallsvariablen um kontinuierliche Werte handelt, die angenommen werden können. Die möglichen Werte werden hier durch die Dichte gewichtet und in Form eines Integrals, das an die Stelle der Summe tritt, zusammengefasst.

6.3.1 Erwartungswert stetiger Zufallsvariablen

Für eine stetige Zufallsvariable X mit der Dichtefunktion $f(x)$ ist der Erwartungswert von X durch

$$E(X) = \int_{-\infty}^{+\infty} xf(x)dx$$

gegeben.

Bei diesem Integral durchläuft *x* alle Werte, die die Zufallsvariable annehmen kann, *x* wird zudem mit den Werten, die die Dichte annimmt, $f(x)$, multipliziert, was einer Gewichtung entspricht.

Erwartungswert der stetigen Gleichverteilung

Eine Zufallsvariable, die stetig gleichverteilt auf *[a, b]* ist, besitzt den Erwartungswert

$$E(X) = \int_{-\infty}^{+\infty} xf(x)dx = \int_{a}^{b} x/(b-a) = \frac{a+b}{2}$$

Rechenregeln für den Erwartungswert

Sind *a* und *b* beliebige Konstanten und *X* und *Y* beliebige Zufallsvariablen, so gilt:

- $E(a) = a$
- $E(b * X) = b * E(X)$
- $E(X + Y) = E(X) + E(Y)$

6.3.2 Varianz stetiger Zufallsvariablen

Die Varianz misst, wie im Falle diskreter Zufallsvariablen, die erwartete quadratische Abweichung vom Erwartungswert. Benutzt man den Erwartungswert zur Definition der Varianz, so schreibt sich die Varianz wie im Falle der diskreten Zufallsvariablen.

Varianz bei stetigen Zufallsvariablen

Die Varianz einer Zufallsvariablen *X* ist definiert als:

$$Var(X) = E[X - E(X)]^2.$$

Für eine stetige Zufallsvariable berechnet sie sich wie folgt:

$$Var(X) = \int_{-\infty}^{+\infty} (x - E(X))^2 f(x)dx$$

Der Verschiebungssatz der Varianz gilt auch für stetige Zufallsvariablen:

$$Var(X) = E(X^2) - [E(X)]^2.$$

Rechenregeln für die Varianz

Sind *a* und *b* beliebige Konstanten und *X* und *Y* unabhängige Zufallsvariablen, so gilt:

- $Var(X + a) = Var(X)$
- $Var(b * X) = b^2 * Var(X)$
- $Var(b * X + a) = b^2 * Var(X)$
- $Var(X + Y) = Var(X) + Var(Y)$

6.4 Die Normalverteilung

Die Normalverteilung ist die wohl wichtigste stetige Verteilung. Viele Größen in der realen Welt lassen sich durch eine Normalverteilung beschreiben. In der Betriebswirtschaft sind beispielsweise Messgrößen, die für die Qualitätskontrolle wichtig sind, und Summen aus Einzelgrößen, wie beispielsweise Schadenssummen in Versicherungen, normalverteilt. Die Form der Dichte der Normalverteilung lässt sich durch eine Glocke beschreiben. Sie besitzt die größte Wahrscheinlichkeitsmasse für Werte um den Mittelpunkt der Verteilung, der durch den Erwartungswert gegeben ist. Je weiter Werte vom Erwartungswert entfernt sind, desto unwahrscheinlicher treten sie auf. Die Normalverteilung ist symmetrisch um den Erwartungswert. Die Bedeutung der Normalverteilung ergibt sich auch aus dem später behandelten Zentralen Grenzwertsatz, der besagt, dass die Summe zufälliger Einflüsse (unter gewissen Bedingungen) normal verteilt ist. So kommt die Normal Verteilung für Größen infrage, die aus vielen zufälligen Einflüssen zusammengesetzt sind. Auch lassen sich aus diesem Grunde andere Verteilungen durch die Normalverteilung annähern.

Normalverteilung

Eine Zufallsvariable ist normalverteilt mit den Parametern μ und σ^2, wenn ihre Dichte die Gestalt

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

besitzt.
Ist eine Zufallsvariable *X* normalverteilt mit μ und σ^2, so wird das durch $X \sim N(\mu;\sigma^2)$ abgekürzt.

Die Normalverteilung wird auch Gaußverteilung genannt.
Die Normalverteilung mit den Parametern $\mu = 0$ und $\sigma^2 = 1$ wird auch Standardnormalverteilung genannt.

Die Verteilungsfunktion der Normalverteilung ist nach der Definition der Verteilungsfunktion durch

$$F(x) = \int_{-\infty}^{x} \frac{1}{\sigma\sqrt{2\pi}} * \exp\left(-\frac{(u-\mu)^2}{2\sigma^2}\right) du$$

gegeben. Die Verteilungsfunktion der Standardnormalverteilung an der Stelle *x* wird meist mit $\Phi(x)$ bezeichnet.

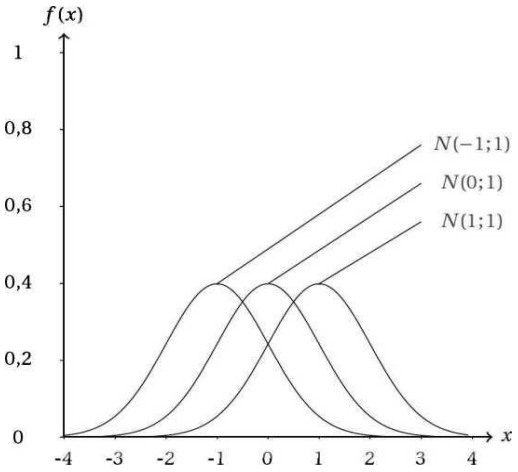


Abb. 17: Dichten der Normalverteilung mit $\mu - 1, 0, 1$ und $\sigma^2 = 1$

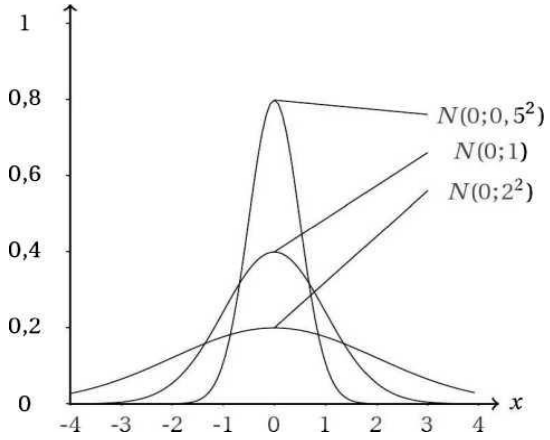


Abb. 18: Dichten der Normalverteilung mit $\mu = 0$ und σ^2 = 0.5, 1 ,2

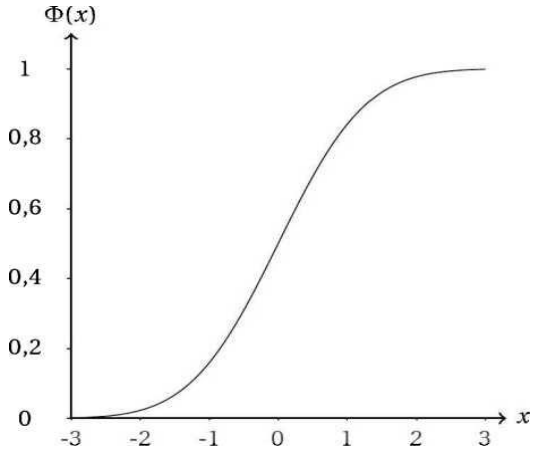


Abb. 19: Verteilungsfunktion der Standardnormalverteilung ($\mu = 0$ und $\sigma^2 = 1$)

Erwartungswert und Varianz der Normalverteilung

Der Erwartungswert und die Varianz der Normalverteilung $N(\mu;\sigma^2)$ entsprechen den Parametern. Es gilt

$$E(X) = \mu \text{ und } Var(X) = \sigma^2.$$

Es folgen nun wichtige Eigenschaften der Normalverteilung.

Symnetrie der Normalverteilung:

Die Normalverteilung ist symmetrisch um seinen Mittelwert.
Für die Standardnormalvertelung gilt:

$$\Phi(-x) = 1 - \Phi(x)$$

Summen unabhängiger normalverteilter Zufallsvariablen

Die Summe zweier unabhängiger normalverteilter Zufallsvariablen $X_1 \sim N(\mu_1; \sigma_1^2)$, $X_2 \sim N(\mu_2; \sigma_2^2)$. Es gilt:

$$X_1 + X_2 \sim N(\mu_1 + \mu_2; \sigma_1^2 + \sigma_{12}^2),$$

Daraus ergibt sich: Die Summe *n* unabhängiger normalverteilter Zufallsvariablen $X_i \sim N(\mu_i; \sigma_i^2)$ ist normalverteilt und es gilt:

$$\sum_{i=1}^{n} X_i \sim N\left(\sum_{i=1}^{n} \mu_i, \sum_{i=1}^{n} \sigma_i^2\right)$$

Da die Dichte der Normal Verteilung nur numerisch integrierbar ist, sind Wahrscheinlichkeiten für Ereignisse durch die Verteilungsfunktion zu bestimmen. Die Wahrscheinlichkeit für ein beliebiges Intervall kann durch die Verteilungsfunktion ausgedrückt werden. Steht kein Computerprogramm zur Verfügung, so können die Werte der Verteilungsfunktion der Standardnormalverteilung aus einer Tabelle abgelesen werden (siehe die Tabelle im Anhang). Dazu ist eine beliebige Normalverteilung auf die Standardnormalverteilung zurückzuführen. Dies nennt man auch das Standardisieren einer Zufallsvariable.

Standardisieren einer Normalverteilung

Sei $X \sim N(\mu; \sigma^2)$. Dann ist

$$Z = \frac{X - \mu}{\sigma} \sim N(0; 1)$$

Umgekehrt ergibt sich aus einer standardnormalverteilten Zufallsvariable *Z* durch die Transformation $X = \sigma Z + \mu$ eine $N(\mu; \sigma^2)$-verteilte Zufallsvariable *X*.

Das Standardisieren ergibt sich direkt aus den bekannten Rechenregeln für den Erwartungswert und die Varianz, wenn man berücksichtigt, dass lineare Transformationen wieder normalverteilt sind.
Damit kann die Verteilungsfunktion einer Zufallsvariable *X* mit einer beliebigen Normalverteilung $N(\mu; \sigma^2)$ durch Φ bestimmt werden.

Rechenregeln für normalverteilte Zufallsvariablen Für eine Zufallsvariable $X \sim N(\mu; \sigma^2)$ gilt:

- $P(X \leq b) = F(b) = \Phi\left(\frac{b-\mu}{\sigma}\right)$
- $P(X > b) = 1 - \Phi\left(\frac{b-\mu}{\sigma}\right)$
- $P(a \leq X \leq b) = \Phi\left(\frac{b-\mu}{\sigma}\right) - \Phi\left(\frac{a-\mu}{\sigma}\right)$
- $P(-a \leq X \leq a) = 2 * \Phi\left(\frac{b-\mu}{\sigma}\right) - 1$

Beispiel:

Es soll die Wahrscheinlichkeit bestimmt werden, mit der der Anleger 5% oder mehr Verlust mit seinem ersten Investment erleidet. Für die Rendite, in diesem Beispiel mit X bezeichnet, gilt: $X \sim N(5,5^2)$.

Gesucht ist die Wahrscheinlichkeit:

$$P(X \leq -5) = \Phi\left(\frac{-5-5}{5}\right) = \Phi(-2)$$

Der Quantile-Tabelle der Standardmormalverteilung (siehe Anhang) entnimmt man den Wert $\Phi(2) = 0{,}9772$

Somit gilt:

$$P(X \leq -5) = \Phi\left(\frac{-5-5}{5}\right) = \Phi(-2) = 1 - \Phi(2) = 0{,}0228.$$

Auch lassen sich durch die Standardisierung die Quantile beliebiger Normalverteilungen auf die Quantile der Standardnormalverteilung zurückführen.

6.5 Die Exponentialverteilung

Die Exponentialverteilung gehört zu den sogenannten Lebensdauerverteilungen, mit denen Wartezeiten, d.h. die Dauer bis ein bestimmtes Ereignis eintritt, modelliert werden. Ereignisse können die Lebensdauer von Produkten, die Zeit bis zur nächsten Schadensmeldung in einer Versicherung, die Bearbeitungszeit in einer Kundenhotline oder auch die Zeit zwischen zwei Serveranfragen sein. Da Wartezeiten stets positiv sind, entspricht der Wertebereich von Lebensdauerverteilungen den positiven reellen Zahlen einschließlich der Null. Die Dichte und die Wahrscheinlichkeiten nehmen ab einem bestimmten Zeitpunkt für längere Wartezeiten ab. Die Exponentialverteilung stellt auch die stetige Verallgemeinerung der geometrischen Verteilung dar, bei der ein Experiment bis zum Eintreten eines Erfolgs wiederholt wird.

Exponentialverteilung

Eine stetige Zufallsvariable mit Werten auf $[0, \infty)$ ist exponentialverteilt mit Parameter λ, wenn die Dichte die Gestalt

$$f(x) = \begin{cases} \lambda * \exp(-\lambda x) & x \geq 0 \\ 0, & x < 0 \end{cases}$$

besitzt.

Ist eine Zufallsvariable X exponentialverteilt mit Parameter λ, so schreibt man auch kurz

$$X \sim Exp(\lambda)$$

Nach der Definition lautet die Verteilungsfunktion der Exponentialverteilung:

$$F(x) = \begin{cases} 1 - \exp(-\lambda x) & x \geq 0 \\ 0, & x < 0 \end{cases}$$

Erwartungswert und Varianz der Exponentialverteilung

Wenn $X \sim Exp, dan\ gelten$:

$$E(X) = \frac{1}{\lambda} \quad und \quad Var(X) = \frac{1}{\lambda^2}.$$

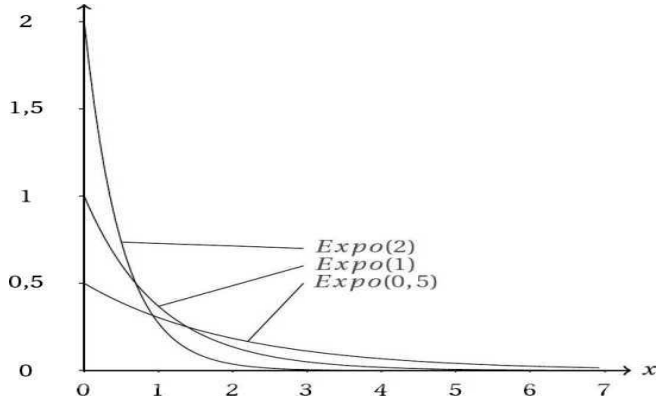


Abb. 20: Dichte der Exponentialverteilung für $\lambda = 0,5;\ 1;\ 2$

Beispiel

Die erwartete Lebensdauer eines bestimmten Fabrikats von Glühbirnen beträgt 2 000 Stunden. Wie wahrscheinlich ist es, dass eine Glühbirne nach 5 000 Stunden noch brennt?

Für den Erwartungswert der Lebenszeit der Glühbirne, *X* ist E(X) = 2000 bekannt.

Folglich kann $\frac{1}{\lambda} = 2000$ also $\lambda = \frac{1}{2000}$ angenommen werden.

Mithilfe der Verteilungsfunktion kann dann die gesucht Wahrscheinlichkeit ermittelt werden.

$$P(X > 5000) = 1 - F(5000) = 1 - \left(1 - \exp\left(-\frac{1}{2000} * 5000\right)\right) = \exp(-2{,}5) = 0{,}082.$$

6.6 Sätze der Wahrscheinlichkeitsrechnung

Im folgenden Abschnitt werden wichtige Sätze der Wahrscheinlichkeitsrechnung behandelt. Es sind das Gesetz der großen Zahlen und der Zentrale Grenzwertsatz. Das Gesetz der großen Zahlen besagt, dass sich das arithmetische Mittel mit steigender Anzahl an Versuchen immer genauer dem Erwartungswert annähert. Diese Tatsache bildet die Rechtfertigung für empirische Untersuchungen. Sie besagen, dass man sich durch eine zunehmende Anzahl an Beobachtungen dem Erwartungswert oder der Wahrscheinlichkeit einer Verteilung, die hinter einem empirischen Phänomen steckt, beliebig nähern kann. Der zentrale Grenzwertsatz besagt, dass die Summe oder der Durchschnitt unabhängiger, identisch verteilter Zufallsvariablen normalverteilt ist. Im Rahmen der induktiven Statistik werden oftmals Stichprobenmittel betrachtet. Nach dem Zentralen Grenzwertsatz nähert sich das Stichprobenmittel der Normalverteilung an. Somit können Wahrscheinlichkeitsaussagen formuliert werden, auch wenn die Verteilung der Einzelgrößen unbekannt ist. Auch begründet der Zentrale Grenzwertsatz, warum Merkmale die aus einzelnen zufälligen Größen zusammengesetzt sind, oft einer Normalverteilung folgen.

6.6.1 Gesetz der großen Zahlen

Das Gesetz der großen Zahlen sagt aus, dass sich das arithmetische Mittel einer wachsenden Anzahl von unabhängigen Zufallsvariablen mit Erwartungswert μ und Varianz σ^2 immer stärker dem Erwartungswert μ annähert. Die Wahrscheinlichkeit, dass das Mittel beliebig wenig vom Erwartungswert abweicht, strebt gegen eins.

Gesetz der großen Zahlen

Seien $X_1, \ldots, X_n$

- unabhängig und identisch verteilte Zufallsvariablen mit
- $E(X_i) = \mu$ und
- $Var(X_i) = \sigma^2$, und sei
- $\overline{X_n} = \sum_{i=1}^{n} X_i / n$ das arithmetische Mittel.

Dann gilt:

$$\lim_{n\to\infty} P(|\overline{X_n} - \mu| < c) = 1 \quad \text{für beliebig kleines } c > 0.$$

In Worten: Die Wahrscheinlichkeit, dafür, dass $\overline{X_n}$ beliebig wenig (c) von μ abweicht, geht für $n \to \infty$ gegen eins. Man sagt auch $\overline{X_n}$ konvergiert nach Wahrscheinlichkeit gegen p.

6.6.2 Zentraler Grenzwertsatz

Der Zentrale Grenzwertsatz gehört zu den wichtigsten Aussagen der Wahrscheinlichkeitstheorie. In einfachen Worten besagt er, dass die Summe oder der Durchschnitt von unabhängigen, identisch verteilten Zufallsvariablen sich der Normalverteilung beliebig annähert, wenn die Anzahl der Zufallsvariablen steigt. In der Formulierung des Satzes standardisiert man die Summe der Zufallsvariablen meist noch. Die standardisierte Summe nähert sich dann der

Standardnormalverteilung. Die Standardisierung ergibt sich direkt aus den Rechenregeln für Erwartungswerte und Varianzen.

Standardisierte Summe

Seien $X_i, i = 1, \dots, n$ unabhängige, identisch verteilte Zufallsvariablen mit $E(X_i) = \mu$ und $Var\ (Xi) = \sigma^2$. Dann besitzt die Zufallsvariable $\sum_{i=1}^{n} X_i$ den Erwartungswert $E(\sum_{i=1}^{n} X_i) = n * \mu$ und die Varianz $Var(\sum_{i=1}^{n} X_i) = n * \sigma^2$, so dass

$$Y_n = \frac{\sum_{i=1}^{n} X_i - n * \mu}{\sqrt{n\sigma^2}}$$

standarddisiert ist, also $E(Y_n) = 0$ and $Var(Y_n) = 1$. Y_n nennt man standardisierte Summe der $X_1, \dots, X_n$.

Zentraler Grenzwertsatz

Die Verteilungsfunktion von Y_n an jeder Stelle konvergiert gegen die Verteilungsfunktion der Standardnormalverteilung wenn $n \to \infty$.

$$\lim_{n\to\infty} P(Y_n \leq y) = \Phi(y) \quad \text{für alle } y.$$

Der Zentrale Grenzwertsatz begründet, dass viele Größen in den Anwendungen, die sich aus der Überlagerung einzelner zufälliger Effekte zusammensetzen, normal verteilt sind.

Durchschnitte von Zufallsvariablen tauchen im Rahmen von Stichproben, wie sie in der induktiven Statistik verwendet werden, häufig auf. Hier liefert der Zentrale Grenzwertsatz die Rechtfertigung für die Verwendung der Normalverteilung.

6.7 Prüfverteilungen

Im Rahmen der induktiven Statistik werden aus Stichproben, die unabhängige Zufallsvariablen darstellen, Statistiken berechnet, um Parameter zu schätzen oder Aussagen zu prüfen. Um Wahrscheinlichkeitsaussagen treffen zu können, ist es notwendig, die Verteilung dieser Statistiken zu kennen. Im Folgenden werden die später für die induktive Statistik wichtigen Verteilungen behandelt. Dies sind die χ^2-Verteilung, die t-Verteilung und die F-Verteilung. Für diese Verteilungen sind die Verteilungsfunktionen und Quantile tabelliert. Üblicherweise arbeitet man nicht mit den Dichten der Verteilungen, sodass auf ihre Angabe verzichtet wird.

6.7.1 χ^2-Verteilung

Seien X_n n unabhängige und identisch $N(0,1)$-verteilte Zufallsvariablen. Dann ist die Summe der Quadrate

$$Z = X_1^2 + \dots + X_n^2$$

χ^2-verteilt mit n Freiheitsgraden: $Z \sim \chi^2(n)$.

Es gilt

$$E(Z) = n \quad \text{und} \quad Var(Z) = 2n.$$

Die χ^2-Verteilung nimmt nur positive Werte an. Die Gestalt ihrer Dichte ist im Folgenden für unterschiedliche Freiheitsgrade angegeben.

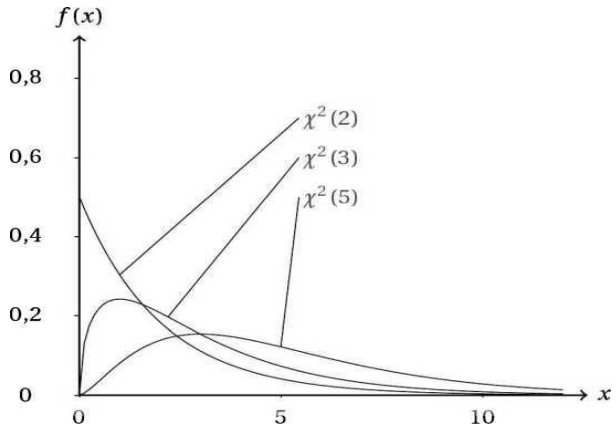


Abb. 21: Dichten der χ^2-Verteilung für verschiedene Freiheitsgrade

6.7.2 t-Verteilung

Die t-Verteilung wird später benötigt, um Aussagen über Erwartungswerte zu treffen, wenn sowohl der Erwartungswert als auch die Varianz unbekannt sind

Seien X und Y unabhängige Zufallsvariablen mit $X \sim N(0,1)$ und $Y \sim \chi^2(n)$. Dann ist der Quotient

$$T = \frac{X}{\sqrt{Y/n}}$$

t-verteilt mit n Freiheitsgraden.

$$E(T) = 0 \; f\ddot{u}r \; n \geq 2 \quad \text{und} \quad Var(T) = \frac{n}{n-2} \; f\ddot{u}r \; n \geq 3.$$

Die t-Verteilung ist von der Gestalt ähnlich der Normalverteilung, streut allerdings stärker. Die t-Verteilung mit zunehmenden Freiheitsgraden nähert die Standardnormalverteilung an.

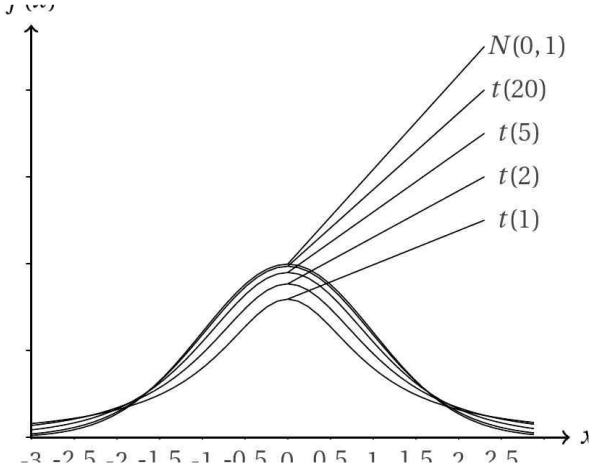


Abb. 22: Dichten der t-Verteilung für verschiedene Freiheitsgrade

6.7.3 F-Verteilung

Seien X und Y unabhängige Zufallsvariablen mit $X \sim \chi^2(m)$ und $Y \sim \chi^2(n)$. Dann ist der Quotient

$$Z = \frac{X/m}{Y/n}$$

F-verteilt mit m und n Freiheitsgraden. Es gilt:

$$E(Z) = \frac{n}{n-2} \; f\ddot{u}r\, n \geq 3 \quad \text{und} \quad Var(Z) = \frac{2n^2(n+m-2)}{m(n-4)(n-2)^2} \; f\ddot{u}r\, n \geq 5.$$

7. Zweidimensionale Zufallsvariablen

In den vorherigen Kapiteln wurden bereits mehrdimensionale unabhängige Zufallsvariablen betrachtet. Mehrdimensionale unabhängige Zufallsvariablen werden später im Rahmen der induktiven Statistik benötigt. Dieses Kapitel stellt eine Ergänzung dar. Es werden zweidimensionale Zufallsvariablen beschrieben, die für viele praxisrelevante Fragestellungen von Bedeutung sind. In diesem Zusammenhang tauchen ähnliche Eigenschaften auf, wie sie im Rahmen von zwei Merkmalen in der deskriptiven Statistik beschrieben wurden. So werden die Kovarianz und Korrelation definiert, die ähnlich wie in der deskriptiven Statistik interpretiert werden können.

Im Gegensatz zu den eindimensionalen Zufallsvariablen werden nun Zufallsvorgänge betrachtet, deren Ergebnis sich durch zwei reelle Zahlen beschreiben lässt. Beispiele sind:

- Umsatz und Gewinn eines Unternehmens
- Tagestemperatur an zwei aufeinander folgenden Tagen
- Werbekosten und Umsatzsteigerung

In allen Beispielen liegt nahe, dass die einzelnen Größen nicht vollständig unabhängig voneinander sind. Im Folgenden wird gezeigt, wie zwei- bzw. mehrdimensionale Zufallsvariablen im Allgemeinen beschrieben werden können.

7.1 Diskrete zweidimensionale Zufallsvariablen

Im Folgenden werden zwei diskrete Zufallsvariablen *X* und *Y* gleichzeitig betrachtet. Die gemeinsame Verteilung, die die Wahrscheinlichkeiten bestimmt, mit denen die verschiedenen möglichen Ergebnisse auftreten, kann durch die gemeinsame Wahrscheinlichkeitsfunktion beschrieben werden.

7.1.1 Gemeinsame Wahrscheinlichkeitsfunktion

Gegeben sei *X* mit den möglichen Ausprägungen $x_1, \ldots, x_I$ und *Y* mit den Ausprägungen $y_1, \ldots, y_J$. Die gemeinsame Wahrscheinlichkeitsfunktion ist dann definiert als

$$P(X = x_i, Y = y_j) = p_{ij}, \qquad i = 1, \ldots, I \; und \; j = 1, \ldots, J$$

Diese lässt sich in Form einer Kontingenztafel darstellen:

	y_1	y_j	y_J
x_1	P_{11}	P_{1j}	P_{1J}
x_i	P_{i1}	Pij	P_{iJ}
x_I	P_{I1}	P_{Jj}	P_{IJ}

7.1.2 Bedingte Wahrscheinlichkeitsfunktion

Die bedingte Wahrscheinlichkeitsfunktion von X gegeben, dass Y den Wert y_j annimmt, ist durch

$$P(X = x_i | Y = y_j) = \frac{P(X = x_i, Y = y_j)}{P(Y = y_j)}$$

für alle möglichen Ausprägungen $x_i$von X. Ist $P(Y = y_j) = 0$, so gilt $P(X = x_i, Y = y_j) = 0$.

Alternativ zur gemeinsamen Wahrscheinlichkeitsfunktion kann eine zweidimensionale Zufallsvariable durch die gemeinsae Verteilungsfunktion beschrieben werden.

7.1.3 Gemeinsame Verteilungsfunktion

Die gemeinsame Verteilungsfunktion ist definiert durch:

$$F(x, y) = P(X \leq x, Y \leq y) = \sum_{x_i \leq x} \sum_{y_j \leq y} P(X = x_i, Y = y_j).$$

7.2 Stetige zweidimensionale Zufallsvariablen

7.2.1 Gemeinsame Dichtefunktion

Eine zweidimensionale Zufallsvariable X, Y heißt stetig, falls es eine nichtnegative reelle Funktion $f_{XY}(x, y) \geq 0$ gibt, sodass

$$P(a \leq X \leq b, c \leq Y \leq d) = \int_a^b \int_c^d f_{XY}(x, y) dy dx$$

erfüllt ist.

Ähnlich wie bei den zweidimensionalen diskreten Zufallsvariablen lassen sich Randdichten bestimmen, die den Dichten der einzelnen stetigen Zufallsvariablen entsprechen, wenn der Wert der anderen Zufallsvariable unberücksichtigt bleibt.

7.2.2 Randdichten

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

und

$$f_Y(y) = \int_{-\infty}^{+\infty} f_{XY}(x, y) dx$$

Die Verteilung von zweidimensionalen stetigen Zufallsvariablen kann neben der gemeinsamen Dichte durch die gemeinsame Verteilungsfunktion beschrieben werden.

7.2.3 Gemeinsame Verteilungsfunktion

Wie bei zweidimensionalen diskreten Zufallsvariablen entspricht die Verteilungsfunktion den Wahrscheinlichkeiten, dass Werte $X \leq x$ und $Y \leq y$ auftreten. Sie kann durch die gemeinsame Dichte bestimmt werden:

$$F(x,y) = P(-\infty \leq X \leq x, -\infty \leq Y \leq y) = \int_{-\infty}^{x}\int_{-\infty}^{y} f_{XY}(u,v)dvdu$$

7.3 Eigenschaften zweidimensionaler Zufallsvariablen

7.3.1 Unabhängigkeit

Die Unabhängigkeit zweier Zufallsvariablen wurde bereits in den Abschnitten über diskrete und stetige Zufallsvariablen behandelt. Dort wurde bemerkt, dass zwei diskrete Zufallsvariablen genau dann unabhängig sind, wenn

$$P(X = x, Y = y) = P(X = x) * P(Y = y)$$

gilt.

Die Wahrscheinlichkeitsfunktionen *P(X = x)* und *P(Y =* y) entsprechen auch den Randverteilungen. Die gemeinsame Wahrscheinlichkeitsfunktion lässt sich also als das Produkt der Randverteilungen darstellen.

Über stetige unabhängige Zufallsvariablen wurde bemerkt, dass

$$P(X < x, Y < y) = P(X < x) * P(Y < y)$$

für beliebige x und y gilt. Für stetige Zufallsvariablen lässt sich die Unabhängigkeit auch über die gemeinsame Dichtefunktion formulieren.

Unabhängige stetige Zufallsvariablen

Zwei Zufallsvariablen X und Y sind voneinander unabhängig, wenn die gemeinsame Dichtefunktion gleich dem Produkt ihrer Randdichten ist, also

$$f_{XY}(x,y) = f_X(x) * f_Y(y) \; für \; alle \; x,y$$

gilt.

7.3.2 Kovarianz

Ähnlich wie in der deskriptiven Statistik ist bei der Betrachtung zweier Zufallsvariablen die Frage nach einer Abhängigkeit zweier Zufallsvariablen von Interesse. So kann es sein, dass bestimmte Wertepaare mit höherer Wahrscheinlichkeit auftreten als andere. Analog zur deskriptiven Statistik ist die Kovarianz auch für zweidimensionale Zufallsvariablen definiert. Anstelle der relativen Häufigkeiten der Wertepaare tritt hier die gemeinsame Wahrscheinlichkeitsfunktion bzw. die gemeinsame Dichte. Die

Interpretation der Kovarianz ist vergleichbar zur deskriptiven Statistik. Es wird die Stärke des linearen Zusammenhangs bestimmt. Je höher der Betrag der Kovarianz, desto stärker der lineare Zusammenhang. Ein positiver Wert der Kovarianz spricht für einen positiven linearen Zusammenhang, während ein negativer einen negativen andeutet. Die Kovarianz ist wie folgt definiert.

7.3.3 Kovarianz

Die Kovarianz von zwei Zufallsvariablen *X* und *Y* ist definiert als

$$Cov(X,Y) = E\left[\left(X - E(X)\right) * (Y - E(Y))\right]$$

Nach dem Verschiebungssatz, der in der Umformung der oben angegebenen Darstellung beruht, gelangt man auf die alternative Darstellung

$$Cov(X,Y) = E(X * Y) - E(X) * E(Y).$$

Der Erwartungswert eines Produkts diskreter Zufallsvariablen ist durch

$$E(X * Y) = \sum_i \sum_j x_i y_j P(X = x_i, Y = y_j)$$

gegeben.

Der Erwartungswert des Produkts zweier stetiger Zufallsvariablen ist durch

$$E(X * Y) = \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} xy f_{XY}(x,y) dy dx$$

gegeben.

Eigenschaften der Kovarianz

- Die Kovarianz ist symmetrisch bezüglich der beiden Zufallsvariablen. Es gilt: *Cov(X, Y) = Cov(Y,X).*
- Werden die Zufallsvariable *X* und *Y* linear transformiert, $X^* = a + bX,\ Y^* = c + dY$,

so gilt

$$Cov(X^*, Y^*) = b * d * Cov(X,Y).$$

Im Rahmen der Zufallsvariablen wurde die Rechenregel behandelt, die besagt, dass die Varianz der Summe zweier Zufallsvariablen im Falle von unabhängigen Zufallsvariablen gleich der Summe der Varianzen der beiden Zufallsvariablen ist. Diese Rechenregel lässt sich mithilfe der Kovarianz auf beliebige Zufallsvariablen erweitern.

Varianz der Summe zweier Zufallsvariablen

Sind *X* und *Y* Zufallsvariablen (nicht notwendig unabhängig), so gilt

$$Var(X + Y) = Var(X) + Var(Y) + 2Cov(X,Y).$$

7.3.4 Korrelationskoeffizient

Analog zur deskriptiven Statistik kann der Korrelationskoeffizient für Zufallsvariablen definiert werden, indem die Kovarianz durch das Produkt der Standardabweichungen geteilt wird.

Korrelationskoeffizient

Der Korrelationskoeffizient von X und Y ist definiert durch

$$\rho(X,Y) = \frac{cov(X,Y)}{\sqrt{Var(X) * Var(Y)}}$$

Es gilt

$$-1 \leq \rho(X,Y) \leq 1.$$

Die Interpretation ist vergleichbar zum Korrelationskoeffizienten der deskriptiven Statistik. Im Folgenden werden nochmals die wichtigsten Eigenschaften aufgeführt.

Interpretation des Korrelationskoeffizienten

Es gelten die folgenden Interpretationen möglicher Werte:

- $\rho(X,Y) > 0$: Es liegt ein positiver linearer Zusammenhang vor.
- $\rho(X,Y) < 0$: Es liegt ein negativer linearer Zusammenhang vor.
- $\rho(X,Y) = 0$: Die Zufallsvariablen X und Y sind linear unabhängig oder unkorreliert.
- $\rho(X,Y) = 1$: Die Zufallsvariablen X und Y sind vollständig linear abhängig. Es können nur Werte angenommen werden, die auf einer Geraden mit positiver Steigung liegen.
- $\rho(X,Y) = -1$: Die Zufallsvariablen X und Y sind vollständig linear abhängig. Es können nur Werte angenommen werden, die auf einer Geraden mit negativer Steigung liegen.

Die Korrelation misst lediglich die lineare Abhängigkeit. Sind zwei Zufallsvariablen unkorreliert, so folgt deswegen nicht, dass sie unabhängig sind.

Korrelation und Unabhängigkeit

Sind X und Y unabhängig, dann folgt daraus, dass sie auch unkorreliert sind, d. h.

$$Unabhängigkeit \quad => \quad Cov(X,Y) = \rho(X,Y) = 0.$$

Die Umkehrung ist im Allgemeinen falsch!

Unabhängigkeit und Korrelation bei der bivariaten Normalverteilung

Sind zwei Zufallsvariablen X und Y bivariat normalverteilt, so folgt auch im Falle der Unkorreliertheit auch die Unabhängigkeit der beiden Zufallsvariablen.

8. Punktschätzung von Parametern

Im Rahmen dieses Kapitels wird das Schätzen wichtiger Charakteristiken einer Grundgesamtheit behandelt. Die Grundgesamtheit wird dabei durch eine Verteilung beschrieben, deren Parameter unbekannt und zu bestimmen sind. Um genau zu sein, entsprechen die zufälligen Ziehungen einer Stichprobe dem Zufallsvorgang. Soll zum Beispiel der unbekannte Anteil *p* in der Grundgesamtheit geschätzt werden, der ein bestimmtes Merkmal aufweist, so tritt bei einer zufälligen Wahl einer Untersuchungseinheit das Merkmal mit der Wahrscheinlichkeit *p* auf. Die zufällige Ziehung kann dann durch eine Bernoulli-Variable mit unbekanntem Parameter *p* beschrieben werden. In diesem Kapitel sollen Funktionen der zufälligen Stichprobe besprochen werden, die konkrete Schätzwerte für die interessierenden Parameter liefern. Parameter können zum Beispiel Mittelwerte oder Anteilswerte in der Grundgesamtheit repräsentieren.

8.1 Der Begriff der Punktschätzung

Mithilfe einer Punktschätzung möchte man einen konkreten Wert für einen unbekannten Parameter bestimmen, der die Verteilung festlegt. Im obigen Beispiel möchte man einen Schätzwert für den Anteil bestimmen.

Beispiel:

> Der Anteil der erwerbstätigen Personen in Deutschland kann durch eine unbekannte Wahrscheinlichkeit *p* repräsentiert werden. Eine zufällig ausgewählte Person wird mit Wahrscheinlichkeit *p* erwerbstätig sein. Anhand einer zufälligen Stichprobe soll ein Wert für *p* bestimmt werden.

Beispiel:

> Die Präzision einer Füllanlage ist bekannt, die Streuung der Füllmenge kann mit $a^2 = 1$ angenommen werden. Allerdings besteht aktuell Unklarheit bezüglich der durchschnittlichen Füllmenge. Beschrieben werden kann die Füllmenge durch eine Normalverteilung mit unbekanntem *p* und bekannten $a^2 = 1$. Durch die Beobachtung mehrerer Füllmengen soll eine Schätzung für den unbekannten Parameter *p* gewonnen werden.

Manchmal sind mehrere Parameter für eine Verteilung zu bestimmen.

8.2 Stichprobe und Schätzer

Unter einer Stichprobe verstehen wir allgemein bei endlicher Grundgesamtheit eine zufällige Auswahl von *n* Elementen aus den *N* Elementen der Grundgesamtheit. Bei einem Zufallsexperiment erhält man die Stichprobe durch *n*-fache Wiederholung des Experiments. Falls alle X_i unabhängig und identisch verteilt sind, bezeichnen wir $X = (X_1, \ldots, X_n)$ als unabhängig, identisch verteilte Stichprobe. Die Schreibweise $X = (X_1, \ldots, X_n)$ bezeichnet die Stichprobe (als Zufallsgröße), die X_i sind Zufallsvariablen. Nach Durchführung der Stichprobenziehung, d.h. nach Realisierung der Zufallsvariablen X_i in einem zufälligen Versuch, erhält man die konkrete Stichprobe $x = (x_1, \ldots, x_n)$ mit den realisierten Werten x_i der Zufallsvariablen X_i.

Ein Schätzer oder eine Schätzfunktinist eine Funktion der Zufallsvariablen $(X_1, \ldots, X_n)$.

Beispiel:

Im obigen Beispiel ist der Anteil an erwerbstätigen Personen in der Stichprobe eine naheliegende Schätzung für den entsprechenden Anteil in der Grundgesamtheit. Beschreibt man die Verteilung der Ziehungen durch eine Bernoulli-Verteilung und betrachtet man die Stichprobe als unabhängige Ziehungen aus dieser Verteilung, so lässt sich die Schätzfunktion wie folgt formulieren.

$$\hat{p} = T(X) = \frac{1}{n}\sum_{i=1}^{n} X_i$$

Wobei

$$X_i = \begin{cases} 1 & erwerbstätig \\ 0 & nicht\ erwerbstätig \end{cases} \quad i = 1, \ldots, n.$$

8.2.1 Eigenschaften von guten Schätzfunktionen

Zu den wichtigen Eigenschaften von Schätzfunktionen zählen die Erwartungstreue und die Streuung der Schätzfunktion.

Erwartungstreue

Ein erwartungstreuer Schätzer trifft den unbekannten Parameter im Schnitt genau. Eine Schätzung $T(X)$ eines Parameters θ heißt erwartungstreu (oder unverzerrt, unverfälscht, unbiased), wenn gilt:

$$E(T(X)) = \theta.$$

Eine Interpretation der Erwartungstreue ergibt sich, wenn hypothetisch angenommen wird, dass die Schätzung oft wiederholt würde. Mit steigender Anzahl der Wiederholungen würde sich der durchschnittliche Schätzwert dem wahren Wert immer genauer annähern. Die Schätzfunktion unter- oder überschätzt den wahren Parameter also nicht systematisch.

Bias

Die Differenz zwischen dem Erwartungswert $E(T(X))$ der Schätzung und dem zu schätzenden Parameter θ wird als Bias (Verzerrung) bezeichnet:

$$bias\big(T(X)\big) = E\big(T(X)\big) - \theta.$$

Für eine erwartungstreue Schätzung T(X) von θ gilt: bias(T(X)) = 0.

Je kleiner der Bias desto besser der Schätzer ist

In der Praxis kann es durchaus vorkommen, dass verzerrte Schätzer verwendet werden, die dann allerdings in der Regel für größere Stichprobenumfänge gute Schätzeigenschaften besitzen. Auch kann es vorkommen, dass eine geringe Verzerrung in Kauf genommen wird, weil die Streuung des gelieferten Schätzwerts geringer ist als bei einem alternativen unverzerrten Schätzer. Die Streuung einer Schätzfunktion ist eine andere wichtige Eigenschaft die im Anschluss an ein Beispiel für einen verzerrten Schätzer besprochen wird.

Mean Square Error (MSE)

Die weitere, wichtige Eigenschaft einer Schätzfunktion ist deren Streuung. Das gängige Maß entspricht der erwarteten quadratischen Abweichung der Schätzung vom wahren Parameter. Meist wird der englische Ausdruck „Mean Square Error“ (MSE) für die erwartete quadratische Abweichung verwendet. Der Mean Square Error spiegelt die Genauigkeit der Schätzwerte. Streuen die möglichen Werte, die man erhält, stark um den unbekannten Parameter oder liegen sie nahe zum unbekannten Wert, dem das Interesse gilt? Je kleiner der MSE ist, desto besser ist der Schätzer.

$$MSE(T(X)) = E([T(X) - \theta]^2).$$

8.3 Spezielle Schätzfunktionen

In den vor angegangenen Beispielen tauchten für gängige Schätzprobleme bereits die naheliegenden Schätzfunktionen auf. In diesem Abschnitt sollen nochmals die Schätzer für die gängigen Fragestellungen zusammengefasst werden und deren Eigenschaften besprochen werden.

8.3.1 Schätzen von Anteilswerten

Um Anteilswerte in einer Grundgesamtheit zu schätzen, geht man davon aus, dass sich der Anteil in der Wahrscheinlichkeit, eine Untersuchungseinheit mit einer gewissen Eigenschaft zu ziehen, spiegelt. Die Ziehung kann auch in Form einer Bernoulli-Variable formuliert werden. Bezüglich der Stichprobe geht man davon aus, dass sie der Realisierung unabhängiger Bernoulli-Variablen entspricht. Der effiziente Schätzer für den Anteilswert entspricht dem Anteil in der Stichprobe oder dem arithmetischen Mittel der Bernoulli-Variablen.

Schätzung von Anteilswerten

Sei $X = (X_1, \ldots, X_n)$ eine unabhängige Stichprobe von $B(1, p)$-verteilten Bernoulli-Variablen, dann ist der Schätzer der effiziente Schätzer für die Wahrscheinlichkeit bzw. den Anteil p in der Grundgesamtheit.

Beispiel

In einer Umfrage, die eine Zufallsstichprobe darstellt, gaben 86 von 200 Personen an, dass ihnen ein bestimmter Markenname bekannt ist. Der Anteil

$$\hat{p} = \frac{86}{200} = 0{,}43$$

schätzt den Anteil in der Grundgesamtheit optimal in dem Sinne, dass kein anderer Schätzer, der unverzerrt ist, eine kleinere oder gleich große Varianz besitzt.

8.3.2 Schätzen von Mittelwerten

Kann ein stetiges Merkmal in der Grundgesamtheit (genauer: die Ziehung aus der Grundgesamtheit) durch eine Normalverteilung $N(\mu, \sigma^2)$ beschrieben werden, dann ist das arithmetische Mittel der Stichprobe der effiziente Schätzer.

Sei $X = (X_1, \ldots, X_n)$ eine unabhängige Stichprobe von $N(\mu, \sigma^2)$-verteilten Ziehungen, dann ist der Schätzer

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

der effiziente Schätzer für den Erwartungswert *p*.

Ist die Stichprobe X unabhängig identisch verteilt, dann ist p erwartungstreu und konsistent.

8.3.3 Schätzen der Varianz

Wird ein stetiger Merkmal in der Grundgesamtheit durch eine Normalverteilung $N(\mu, \sigma^2)$ beschrieben, dann kann ist

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

ein Schätzer der Varianz σ^2.

Stellt die Stichprobe unabhängige identisch verteilte Ziehungen aus einer beliebigen Verteilung mit μ und σ^2 dar, so ist der obige Schätzer erwartungstreu und konsistent.

Beispiel

Ein Anleger möchte die Chance und das Risiko seines Portfolios abschätzen. Dies kann durch die Stärke der Streuung der Renditen ausgedrückt werden. Er nimmt an, dass die Tagesrenditen seines Portfolios annähernd unabhängig normal verteilt sind. Es ergaben sich in den letzten zehn Tagen die folgenden Renditen in %.

0,1 0,2 -0,1 -0,2 0 0,1 -0,1 0 0,2 -0,2

Es errechnet sich ein Durchschnitt von 0%. Die Stichprobenvarianz ist $S^2 = 0{,}022$.

Sie stellt eine erwartungstreue Schätzung der Varianz der Rendite im Allgemeinen dar.
Im Falle normalverteilter Renditen ist sie effizient.

9. Intervallschätzung

9.1 Bedeutung und Definition des Konfidenzintervalls

Besonders bei kleinen Stichprobenumfängen können Punktschätzungen sehr ungenau sein. Der Abstand zwischen der Schätzung und dem wahren Parameter ist unter Umständen groß. Leichter interpretierbar sind Konfidenzintervalle. Ein Konfidenzintervall ist ein Bereich, der den unbekannten Parameter mit einer gewissen Sicherheit beinhaltet. Bei Konfidenzintervallen wird vorab eine gewünschte Sicherheit vorgegeben und dann dazu ein Bereich konstruiert, der den wahren Parameter mit der gegebenen Sicherheit enthält. Der Zufallsvorgang beim Schätzen besteht in dem Ziehen der Stichprobe. Vor dem Ziehen der Stichprobe entspricht die Sicherheit der Wahrscheinlichkeit, dass das zu der Stichprobe errechnete Intervall den wahren Parameter enthält.

Ziel ist ein Parameter θ zuschätzen. Sei $X = (X_1, \dots, X_N)$ eine Stichprobe, wobei die Verteilung von X_i vom Paramter θabhängt.

Ein Konfidenzintervall ist ein Intervall $[I_u(X); I_o(X)]$, das den zu schätzenden Parameter θ mit einer vorgebenen Wahrscheinlichkeit enthält. Mathematisch kann dies folgendermaßen frmuliert werden:

$$P(\theta \in [I_u(X); I_o(X)]) \geq 1 - \alpha$$

wobei

- α bezeichnet die Fehlerwahscheinlichkeit, $1 - \alpha$ ist das Konfidenzniveau,
- Falls die beiden Grenzen endliche Zahlen sind, dann spricht man von einem zweiseitigen Konfidenzintervall,
- Falls eine der beiden Grenzen von vorne festgelegt ist und nicht zu schätzen ist, dann spricht man von einem einseitigen Konfidenzintervall,

9.2 Konfidenzintervalle für den Erwartungswert

In diesem Abschnitt sollen Konfidenzintervalle für Erwartungswerte beschrieben werden. Dabei gehen wir davon aus, dass ein stetiges Merkmal in der Grundgesamtheit durch eine Normalverteilung beschrieben werden kann. Der Erwartungswert entspricht dann dem Mittel des Merkmals in der Grundgesamtheit. Die Stichprobe $X = (X_1, \dots, X_n)$ entspricht n unabhängigen Ziehungen aus einer Normalverteilung $N(\mu;\ \sigma^2)$. Gesucht ist ein Konfidenzintervall für μ. Als erstes soll der Fall behandelt werden, bei dem die Varianz σ^2 bekannt ist. Dieser Fall ist in der Anwendung oft unrealistisch, da im Falle, dass der Erwartungswert gesucht ist, die Varianz meist unbekannt ist. Die Konstruktion eines Konfidenzintervalls ist hier aber anschaulicher erklärbar. Im Nachfolgenden wird dann der realistischere Fall betrachtet, in dem auch die Varianz unbekannt ist.

9.2.1 Konfidenzintervall für μ bei bekanntem σ^2

Durch die Verwendung der Tatsache, der arithmetische Mittelwert $\overline{X_n}$ selbst Normalverteilt ist und bei Anwendung der Standardisierung erhält man di folgenden Ergenisse:

- $[I_u(X); I_o(X)] = \left[\bar{X} - \frac{\sigma}{\sqrt{n}} z_{1-\frac{\alpha}{2}};\ \bar{X} + \frac{\sigma}{\sqrt{n}} z_{1-\frac{\alpha}{2}}\right]$ ist ein zweiseitigens Konfidenzintervall zum Konfidenzniveau $1 - \alpha$;
- $[I_u(X); I_o(X)] = [\bar{X} - \frac{\sigma}{\sqrt{n}} z_{1-\alpha};\ +\infty)$ ist ein einseitigens Konfidenzintervall (nach oben gerichtet) zum Konfidenzniveau $1 - \alpha$;
- $[I_u(X); I_o(X)] = (-\infty;\ \bar{X} + \frac{\sigma}{\sqrt{n}} z_{1-\alpha}]$ ist ein einseitigens Konfidenzintervall (nach unten gerichtet) zum Konfidenzniveau $1 - \alpha$.

wobei z_u den u-Quantile der Normalverteilung bezeichnet. Dieser kann in den Quantiltabellen ablesen werden (siehe die Quantiletablellen am Ende des Skripts).

9.2.2 Konfidenzintervall für μ bei unbekanntem σ^2

In der Praxis wird in der Regel, wenn der Erwartungswert μ einer unabhängigen Stichprobe aus einer Normalverteilung $N(\mu;\ \sigma^2)$ unbekannt ist, auch die Varianz σ^2 unbekannt sein. In diesem für die Anwendung wichtigem Fall wird die Varianz durch die Stichprobenvarianz S^2 ersetzt. Daher bekommen wir die folgenden Ergebnisse:

- $[I_u(X); I_o(X)] = \left[\bar{X} - \frac{S}{\sqrt{n}} t(n-1)_{1-\frac{\alpha}{2}};\ \bar{X} + \frac{S}{\sqrt{n}} t(n-1)_{1-\frac{\alpha}{2}}\right]$ ist ein zweiseitigens Konfidenzintervall zum Konfidenzniveau $1 - \alpha$;
- $[I_u(X); I_o(X)] = [\bar{X} - \frac{S}{\sqrt{n}} t(n-1)_{1-\alpha};\ +\infty)$ ist ein einseitigens Konfidenzintervall (nach oben gerichtet) zum Konfidenzniveau $1 - \alpha$;
- $[I_u(X); I_o(X)] = (-\infty;\ \bar{X} + \frac{S}{\sqrt{n}} t(n-1)_{1-\alpha}\ ]$ ist ein einseitigens Konfidenzintervall (nach unten gerichtet) zum Konfidenzniveau $1 - \alpha$.

wobei S^2 die Stichprobevarianz ist und $t(n-1)_u$ den u-Quantile der t-Verteilung mit $n-1$ Freiheitsgraden bezeichnet. Dieser kann in den Quantiltabellen ablesen werden (siehe die Quantiletablellen am Ende des Skripts).

9.3 Konfidenzintervall für die Varianz

In manchen Fällen interessiert man sich für die Streuung eines Merkmals. Ein Beispiel ist die Präzision einer Fertigungsanlage. Je geringer die Varianz ist, desto präziser sind die gefertigten Produkte. Im Folgenden wird angenommen, dass eine Stichprobe von n unabhängigen Ziehungen aus einer Normalverteilung, $N(\mu;\ \sigma^2)$, die die Ziehungen aus der Grundgesamtheit charakterisiert, vorliegt. Es wird angenommen, dass μ und σ^2 unbekannt sind.

Für die Varianz erhält man das folgende Konfidenzintervall:

$$[I_u(X); I_o(X)] = \left[\frac{(n-1)S^2}{c(n-1)_{1-\alpha/2}}; \frac{(n-1)S^2}{c(n-1)_{\alpha/2}}\right]$$

wobei $\boldsymbol{c(n-1)_u}$ den u-Quantile der χ^2-Verteilung mit $(n-1)$ Freiheitsgraden bezeichnet. Dieser kann in den Quantiltabellen ablesen werden (siehe die Quantiletablellen am Ende des Skripts).

9.4 Konfidenzintervalle für eine Wahrscheinlichkeit

Ein für die Anwendung wichtiger Fall sind Konfidenzintervalle für Anteilswerte in einer Grundgesamtheit. Anteilswerte der Grundgesamtheit werden durch die Wahrscheinlichkeit, ein Element mit der konkreten Eigenschaft in einer Stichprobe zu ziehen, modelliert. Dies kann auch durch eine Bernoulli-Variable $B(1,p)$ ausgedrückt werden. Die Wahrscheinlichkeit p entspricht dann dem Anteilswert der Grundgesamtheit. Um ein Konfidenzintervall zu formulieren, geht man von einer Stichprobe mit Umfang n, $X_1, \dots, X_n$ von unabhängigen, Bernoulli-verteilten Zufallsvariablen aus.

Durch die Verwendung des Zentralgrenzwertsatzes erhalten wir den folgenden approximativen Konfidenzintervall

$$[I_u(X); I_o(X)] = \left[\bar{p} - \sqrt{\frac{\bar{p}(1-\bar{p})}{n}} z_{1-\frac{\alpha}{2}};\ \bar{p} + \sqrt{\frac{\bar{p}(1-\bar{p})}{n}} z_{1-\frac{\alpha}{2}}\right]$$

wobei z_u den u-Quantile der Normalverteilung bezeichnet. Dieser kann in den Quantiltabellen ablesen werden (siehe die Quantiletablellen am Ende des Skripts).

Damit die Approxmation angemessen ist sollen die folgenden Faustregel erfüllt sein sollte: $n\bar{p} \geq 5$ als auch $n(1-\bar{p}) \geq 5$ erfüllt sein.

10. Statistischer Test

Unter einem statistischen Test versteht man Regeln, mit deren Hilfe eine Aussage über eine Grundgesamtheit überprüft werden kann. Mittels einer Stichprobe soll Aufschluss über die Grundgesamtheit gewonnen und die Aussage überprüft werden. Dabei stellt die Stichprobe unabhängige Ziehungen aus einer Verteilung dar, die die Grundgesamtheit beschreibt. Statistische Aussagen über die Grundgesamtheit beziehen sich beispielsweise auf den Anteil von Objekten mit einer Eigenschaft oder den Mittelwert eines Merkmals. Ein wichtiges Anwendungsgebiet in der Betriebswirtschaft ist die Qualitätskontrolle. Beispielsweise möchte man durch eine Stichprobe überprüfen, ob die Produktion einen gewissen Anteil an Ausschuss überschreitet, oder ob eine Füllmenge bei abgepackten Waren eingehalten wird. Für Pharmaunternehmen kann es wichtig sein, anhand einer Stichprobe zu belegen, dass ein neues Medikament wirksamer als ein herkömmliches ist. In der Marktforschung möchte man zum Beispiel wissen, ob eine Markteinführung aussichtsreich ist, was an dem Anteil der Marktteilnehmer, die sich für ein neues Produkt entscheiden würden, gemessen werden kann. Solche Aussagen werden dann im Rahmen des Testens in Verteilungen übersetzt, die zutreffen, wenn die Aussage richtig ist.

10.1 Der Binomial-Test und Gaußtest

10.1.1 Binomial-Test

Anhand des Binomial-Tests möchte man eine Aussage über einen Anteil an Objekten, die eine Eigenschaft besitzen, überprüfen. Die Funktionsweise des Tests soll mittels eines Beispiels beschrieben werden.

Beispiel 11

Ein Getränkehersteller überlegt sich, das Rezept für einen Softdrink abzuändern. Da sich der Hersteller unsicher ist, ob die Mehrheit die neue Rezeptur der alten vorzieht, möchte er vor der Markteinführung Sicherheit bekommen, dass die neue Rezeptur erfolgreich sein wird. Um dies herauszufinden, will er einen statistischen Test durchführen. Er formuliert die folgende Aussage:

H_0: *Mindestens so vielen Kunden schmeckt die alte Rezeptur genauso gut wie die neue.*

Diese Aussage wird in der Statistik auch als Nullhypothese bezeichnet. Wenn eine Stichprobe gegen diese Aussage spricht, gilt sie als widerlegt. Dann ist die gegenteilige Aussage,

H_1: *Mehr Kunden schmeckt die neue Rezeptur besser,*

die als Alternative bezeichnet wird, bewiesen.

Ausgangspunkt eines statistischen Tests ist es, eine Aussage und die dazugehörige Gegenaussage über die Grundgesamtheit zu formulieren, die dann mittels einer Stichprobe überprüft werden. In diesem Zusammenhang ist zu überlegen, welche Verteilungen für die Stichprobe infrage kommen und wie sich die Aussagen in Aussagen über die möglichen Verteilungen übersetzen lassen.

Beispiel (Fortsetzung):

Der Getränkehersteller aus obigem Beispiel möchte eine Stichprobe von zehn seiner Kunden ziehen. Jeder Kunde soll die gleiche Chance haben, bei jeder der zehn Ziehungen gezogen zu werden. Die Teilnehmer seiner Studie lässt er beide Rezepturen vergleichen. Er notiert

$$X_i = \begin{cases} 1, & wenn\ der\ Teilnehmer\ die\ neue\ Rezeptur\ besser\ findet, \\ 0, & wenn\ der\ Teilnehmer\ die\ alte\ Rezeptur\ gleich\ gut\ findet \end{cases}$$

.

Da es sich um eine Zufallsstichprobe handelt, nimmt der Hersteller an, dass die $X_i, i = 1, \ldots, 10$, unabhängige bernoulliverteilte Zufallsvariablen, $Xi \sim B(1, p)$ sind. Wenn die Aussage, dass mindestens so viele Kunden die alte Rezeptur genauso gut finden, richtig ist, sollte die Wahrscheinlichkeit, dass ein zufällig gewählter Kunde die neue Rezeptur besser findet, höchstens gleich 0,5 sein. Die Nullhypothese lässt sich also folgendermaßen formulieren:

$$H_0 : p < 0{,}5.$$

Die Alternative lautet entsprechend:

$$H_1 : p > 0{,}5.$$

Im Rahmen des Beispiels sind nun die ersten beiden Schritte bei der Lösung des Testproblems gemacht. Diese Schritte bestehen darin, ein Zufallsexperiment zu beschreiben sowie eine Aussage und Gegenaussage über den interessierenden Parameter zu treffen. Das Zufallsexperiment stellt hier eine Stichprobe dar, die unabhängigen Ziehungen einer Bernoulli-Variable entspricht. Die Aussage betrifft den Parameter *p.*

Formulierung der Modellannahmen

Ein Test wird anhand eines Zufallsexperiments entschieden. Dieses Zufallsexperiment ist zuvor zu formulieren. Die möglichen Verteilungen, die dem Zufalls experiment zugrunde liegen, sind festzulegen.

Die Formulierung in der Beschreibung im Beispiel schränkt die möglichen Verteilungen, die dem Zufalls experiment zugrunde liegen, näher ein. Es handelt sich hier um Bernoulli- Variablen, bei denen der Parameter *p* unbekannt ist. In der Praxis ist es, wie in unserem Beispiel, meist so, dass die möglichen Verteilungen einer konkreten Verteilung entsprechen und durch einen oder mehrere Parameter festgelegt werden können.

Im Rahmen des Beispiels soll nun überlegt werden, welche Versuchsausgänge gegen die Nullhypothese sprechen. Dazu ist es hilfreich, die einzelnen Beobachtungen in einer Größe, der sogenannten Prüfgröße, zusammenzufassen.

Beispiel (Fortsetzung)

Anstatt der im Beispiel beobachteten zehn Bernoulli-Variablen, die den einzelnen befragten Personen entsprechen, ist es sinnvoll, die Summe der Bernoulli-Variablen zu betrachten. Es sei

$$T = \sum_{i=1}^{10} X_i$$

Dann gilt $T \sim B(10, p)$. T entspricht der Anzahl der befragten Kunden, die die neue Rezeptur besser finden. So sprechen kleine Werte von T für die Nullhypothese, während große Werte gegen sie sprechen.

Um zu entscheiden, ob eine Nullhypothese für wahr gehalten wird oder abgelehnt wird, wird im Allgemeinen eine Prüfgröße betrachtet.

Prüfgröße

Im obigen Beispiel, T ist die Prüfgröße. Im Allgemeinen die Prüßgröße ist eine Funktion der Stchprobe und anhand seines Wertes wird entschieden, ob die Nullhypothese abgelehnt wird oder nicht.

Annahme- und Ablehnungsbereich

Die möglichen Werte, die eine Prüfgröße annehmen kann, werden in zwei Bereiche eingeteilt. Dies sind:

- Annahmebereich:
  Fällt die der Prüfgröße in diesen Bereich, so entscheidet man sich für die Nullhypothese.
- Ablehnungsbereich oder kritischer Bereich (Bezeichnung: K):
  Fällt die Realisierung der Prüfgröße in diesen Bereich, so entscheidet man sich für die Alternative.

Für den Fall, dass die Nullhypothese wahr ist, sollte die Wahrscheinlichkeit, dass der Test die Nullhypothese fälschlicherweise widerlegt, höchstens gleich dem Signifikanzniveau sein.

Signifikanzniveau

Eine Forderung an einen statistischen Test besteht darin, dass die Nullhypothese nur dann widerlegt werden soll, wenn der Ausgang des Zufallsexperiments in besonders starkem Maße gegen die Hypothese spricht. Dies spiegelt sich im Signifikanzniveau.

Ein statistischer Test heißt Test zum Signifikanzniveau α, wenn

$$P\ (\text{Ho wird abgelehnt obwohl } \mathrm{H}_0 \text{ trifft zu}) \leq\ \alpha$$

gilt.

Beispiel (Fortsetzung)

Wir geben uns ein Signifikanzniveau von 5 % vor und wollen eine Entscheidungsregel ableiten. Der Test soll die Nullhypothese ablehnen, wenn für die Prüfgröße T sehr hohe Werte realisiert werden. T nimmt Werte zwischen null und zehn an. Nun soll der Ablehnungsbereich K bestimmt werden. Aus der Definition des Signifikanzniveaus

$$P\ (T \in K\ |p \leq 0{,}5) \leq 0{,}05$$

und der Tatsache, dass T Binomial-verteilt ist, kann dann der Ablehnungsbreich Kbestimmt

werden und nämlich

$$K = \{9,10\}.$$

Dies bedeutet, dass der Annahmebereich lautet: $\{0,1, \dots 7,8\}$.

Interpretation eines Testergebnisses

Entscheidet man sich für die Nullhypothese, so sagt man auch, man behält die Nullhypothese bei. In diesem Fall kann die Alternative nicht bewiesen werden. Trotzdem ist nicht gesichert, dass die Nullhypothese wahr ist.
Lehnt man die Nullhypothese ab, so gilt die Alternative als statistisch (zum Niveau α) bewiesen.

10.1.2 Gauß-test

Der einfachste Test, der eine Aussage über den Erwartungswert überprüft, stellt der Gaußtest dar. Beim Gaußtest geht man davon aus, dass unabhängige Realisierungen einer normalverteilten Zufallsvariable vorliegen. Diese können unabhängige Ziehungen aus einer Grundgesamtheit repräsentieren. Es wird davon ausgegangen, dass der Erwartungswert der Zufallsvariablen unbekannt, die Varianz allerdings bekannt ist. In der Praxis ist oftmals die Varianz auch nicht bekannt, wenn der Erwartungswert unbekannt ist. Diesen allgemeineren Fall behandelt der t-Test, der später behandelt wird. Der Gaußtest prüft schließlich eine Aussage über den Erwartungswert anhand der realisierten Zufallsvariablen. Die gängigen Aussagen, die man überprüfen möchte, sind die folgenden. Zum einen kann getestet werden, ob der unbekannte Erwartungswert größer als ein vorgegebener Wert ist. Umgekehrt kann man prüfen, ob er kleiner als ein vorgegebener Wert ist. Zum anderen kann überprüft werden, ob er sich von einem vorgegebenen Wert unterscheidet. In den ersten beiden Fällen spricht man von einseitigen Tests, im zweiten Fall von einem zweiseitigen Test.

Formulierung der Modellannahmen

Wir nehmen an, dass wir N Beobachtungen $X_1, \dots, X_N$, die normal-verteilt sind mit den Parametern μ und σ^2, wobei σ^2 bekannt ist.

Null- und Alternative-Hypothesen:

μ_0 stellt einen konkreten Wert dar, gegen den man den unbekannten Parameter μ prüfen möchte.

Fall	Null Hypothese	Alternativ-Hypothese	Testproblem
(a)	$\mu \geq \mu_0$	$\mu < \mu_0$	einseitig
(b)	$\mu \leq \mu_0$	$\mu > \mu_0$	einseitig
(c)	$\mu = \mu_0$	$\mu \neq \mu_0$	zweiseitig

Prüfgröße

$$T = \frac{\overline{X_N} - \mu_0}{\sigma/\sqrt{N}}$$

wobei $\overline{X_n} = (X_1 + \cdots + X_N)/N$ den arithmetischen Mittelwert der Beobachtungen bezeichnet. Es kann nachgewiesen werden, dass $\boldsymbol{T}$ under der Hypothese $\boldsymbol{H_0}$Standardnormal ist.

Ablehnungsbereich (zum Signifinkanzniveau $\boldsymbol{\alpha}$)

Fall	Null Hypothese	Alternativ-Hypothese	Ablehnungsbereich
(a)	$\mu \geq \mu_0$	$\mu < \mu_0$	$K = (-\infty, -z_{1-\alpha})$
(b)	$\mu \leq \mu_0$	$\mu > \mu_0$	$K = (z_{1-\alpha}, +\infty)$
(c)	$\mu = \mu_0$	$\mu \neq \mu_0$	$K = (-\infty, -z_{1-\alpha}) \cup (z_{1-\alpha}, +\infty)$

wobei z_u den u-Quantile der Normalverteilung bezeichnet. Dieser kann in den Quantiltabellen ablesen werden (siehe die Quantiletablellen am Ende des Skripts).

Beispiel 11.5

Eine Coffeeshop-Kette bezieht regelmäßig Kaffee von einem Großanbieter. Der Kaffee ist in 1-kg-Päckchen abgepackt. Der Anbieter gibt für die Füllmenge eine Schwankung in Form einer Standardabweichung von 0,01 an. Da eine Menge von über 1 kg im eigenen Interesse liegt, soll lediglich überprüft werden, ob man dem Anbieter eine Unter- schreitung der Menge nachweisen kann. Folglich werden die Hypothesen

$$H_0 : \mu \geq 1\ kg\ und\ H_1 : \mu < 1\ kg$$

formuliert, die zu einem Signifikanzniveau von $\alpha = 0{,}05$ überprüft werden sollen. Für 20 zufällig ausgewählten Päckchen wird ein durchschnittliches Gewicht von $0{,}99$ kg nachgewogen. Man geht davon aus, dass die Füllmenge der 20 Päckchen unabhängig normalverteilt mit einer Standardabweichung von 0,01 ist.

$$T = \frac{\overline{X_N} - \mu_0}{\sigma/\sqrt{n}} = \frac{0{,}99 - 1}{0{,}01/\sqrt{20}} = -4{,}47$$

Zum Signifikanzniveau $\alpha = 0{,}05$ ist der Ablehnungsbeich $K = (-\infty\ ;\ -1{,}64)$.

Der Prüfwert liegt innerhalb des Ablehnungsbeichs. Somit ist zu einem Signifikanzniveau von 5% bewiesen, dass die Kaffeepäckchen im Schnitt weniger als 1 kg wiegen.

10.2 11.2 Fehlentscheidungen

Die Aufgabe eines statistischen Tests ist es, zwischen einer Hypothese, der Nullhypothese, und der Gegenhypothese, der Alternative, zu entscheiden. Die getroffene Entscheidung ist aber nicht unbedingt

richtig. Wie bereits deutlich wurde, möchte man im Falle, dass die Nullhypothese richtig ist, die Wahrscheinlichkeit für eine Fehlentscheidung durch das Signifikanzniveau begrenzen. Diese Art von Fehlentscheidung wird auch Fehler 1. Art genannt. Davon unterscheidet man den Fehler 2. Art Fehler, der darin besteht, dass der Test die Nullhypothese beibehält, obwohl sie falsch ist. Insgesamt können vier Ausgänge eines Tests unterschieden werden, je nachdem ob die Nullhypothese richtig oder falsch ist und welche Entscheidung getroffen wird.

Die folgende Tabelle fasst die Ausgänge eines statistischen Tests zusammen:

	Ho ist richtig	Ho ist falsch
H_o wird beibehalten	richtige Entscheidung	*Fehler 2. Art*
H_o wird abgelehnt	*Fehler I.Art*	richtige Entscheidung

Für den Fehler 1. Art gibt man sich eine Schranke in Form einer Wahrscheinlichkeit, die nicht überschritten wird, vor. Die Wahrscheinlichkeit für einen Fehler 2. Art kann unter Umständen sehr hoch sein, weswegen man im Falle, dass die Nullhypothese beibehalten wird, nicht davon ausgehen kann, dass sie richtig ist. Die genaue Wahrscheinlichkeit ist davon abhängig, welchen genauen Wert der unbekannte Parameter besitzt. Es ist wünschenswert, dass der Fehler 2. Art gering ist, da er ja die Chance bestimmt, dass eine richtige Alternative erkannt wird. Betrachtet man die Prüfvorschriften einzelner Tests, so bemerkt man, dass die Fehler 2. Art für einen gegebenen Parameter mit steigendem Stichprobenumfang abnehmen. Somit ist die Wahrscheinlichkeit, mit der eine richtige Alternative bewiesen werden kann, auch vom Stichprobenumfang abhängig.

10.3 Spezielle Testverfahren

In diesem Kapitel sollen Tests zu gängigen Fragestellungen beschrieben werden. Die wohl am häufigsten gebrauchten Tests sind t-Tests, die zunächst erläutert werden. Mit t-Tests kann die Lage von Verteilungen getestet werden. Dabei wird vorausgesetzt, dass die Prüfgröße normalverteilt ist. Ist dies fraglich, so kann auf nichtparametrische Tests ausgewichen werden, von denen der Vorzeichentest vorgestellt wird. Die einfaktorielle Varianzanalyse gestattet den Vergleich von mehreren Erwartungswerten. Weiter wird darauf eingegangen, wie Anteilswerte geprüft werden können. Zudem wird der j^2 -Anpassungstest, mit dem man testen kann, ob eine empirische Verteilung von einer theoretisch angenommenen Verteilung abweicht, und der x^2 - Abhängigkeitstest, mit dessen Hilfe die Abhängigkeit von zwei Merkmalen überprüft werden kann, besprochen

10.3.1 t-Tests (Lagetests)

Die gängigen statistischen Tests zur Überprüfung der Lage von Verteilungen sind t-Tests. Beim einfachen t-Test wird der Erwartungswert einer Normalverteilung gegen einen bestimmten Wert getestet. In der Anwendung wird angenommen, dass die Ereignise aus der Grundgesamtheit durch eine Normalverteilung mit unbekanntem Erwartungswert und unbekannter Varianz beschrieben werden können.

10.3.1.1 Einfacher t-Test

Der Unterschied zwischen dem einfachen t-Test und dem vorher beschrieben Gauß-Test, ist dass beim einfachen t-Test ist die Varianz unbekannt. Dies führt zu einer Änderung der Prüfgröße und der Verteilung anhand derer der Ablehnungsbereich bestimmt wird.

Formulierung der Modellannahmen

Wir nehmen an, dass wir N Beobachtungen $X_1, \dots, X_N$, die normal-verteilt sind mit den Parametern μ und σ^2, wobei σ^2 **unbekannt** ist.

Null- und Alternative-Hypothesen:

μ_0 stellt einen konkreten Wert dar, gegen den man den unbekannten Parameter μ prüfen möchte.

Fall	Null Hypothese	Alternativ-Hypothese	Testproblem
(a)	$\mu \geq \mu_0$	$\mu < \mu_0$	einseitig
(b)	$\mu \leq \mu_0$	$\mu > \mu_0$	einseitig
(c)	$\mu = \mu_0$	$\mu \neq \mu_0$	zweiseitig

Prüfgröße

$$T = \frac{\overline{X_N} - \mu_0}{S/\sqrt{N}}$$

wobei $\overline{X_n} = (X_1 + \cdots + X_N)/N$ den arithmetischen Mittelwert der Beobachtungen bezeichnet und S^2 die Stichprobevarianz ist.

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

Es kann nachgewiesen werden, dass die Prüfgröße (under der Hypothese H_0) t-verteilt ist mit $N-1$ Freiheitsgraden. Daher erginbt sich der folgende Ablehnungsbereich:

Ablehnungsbereich (zum Signifinkanzniveau α)

Fall	Null Hypothese	Alternativ-Hypothese	Ablehnungsbereich
(a)	$\mu \geq \mu_0$	$\mu < \mu_0$	$K = (-\infty, -t(N-1)_{1-\alpha})$
(b)	$\mu \leq \mu_0$	$\mu > \mu_0$	$K = (t(N-1)_{1-\alpha}, +\infty)$
(c)	$\mu = \mu_0$	$\mu \neq \mu_0$	$(-\infty, -t(N-1)_{1-\alpha}) \cup (t(N-1)_{1-\alpha}, +\infty)$

wobei $t(N-1)_u$ den u-Quantile der t-Verteilung mit $N-1$ Freiheitsgraden bezeichnet. Dieser kann in den Quantiltabellen ablesen werden (siehe die Quantiletablellen am Ende des Skripts).

10.3.1.2 Doppelter t-Test

Der doppelte t-Test wird verwendet, wenn getestet werden soll, ob sich die Erwartungswerte zweier Merkmale unterscheiden. Man geht davon aus, dass die Varianzen der beiden Zufallsvariablen unbekannt, aber beide gleich sind. Somit gilt:

Formulierung der Modellannahmen

Wir nehmen an, dass wir N_X Beobachtungen $X_1, \dots, X_{N_X}$, die Normal-verteilt sind mit den Parametern μ_X und σ^2 und weiterhin N_Y Beobachtungen $Y_1, \dots, Y_{N_Y}$, die Normal-verteilt sind mit den Parametern μ_Y und σ^2, wobei σ^2 **unbekannt** ist.

Null- und Alternative-Hypothesen:

Fall	Null Hypothese	Alternativ-Hypothese	Testproblem
(a)	$\mu_X \geq \mu_Y$	$\mu_X < \mu_Y$	einseitig
(b)	$\mu_X \leq \mu_Y$	$\mu_X > \mu_Y$	einseitig
(c)	$\mu_X = \mu_Y$	$\mu_X \neq \mu_Y$	zweiseitig

Prüfgröße

$$T = \sqrt{\frac{N_X * N_Y}{N_X + N_Y}} \frac{\overline{X_N} - \overline{Y_N}}{S}$$

wobei $\boldsymbol{S^2}$ die gepoolte Stichprobevarianz ist.

$$S = \frac{(N_X - 1) * S_X^2 + (N_Y - 1) * S_Y^2}{N_X + N_Y - 2}$$

Es kann nachgewiesen werden, dass die Prüfgröße (under der Hypothese H_0) t-verteilt ist mit $N = N_X + N_Y - 2$ Freiheitsgraden. Daher erginbt sich der folgende Ablehnungsbereich:

Ablehnungsbereich (zum Signifinkanzniveau $\boldsymbol{\alpha}$)

Fall	Null Hypothese	Alternativ-Hypothese	Ablehnungsbereich
(a)	$\mu \geq \mu_0$	$\mu < \mu_0$	$K = (-\infty, -t(N-1)_{1-\alpha})$
(b)	$\mu \leq \mu_0$	$\mu > \mu_0$	$K = (t(N-1)_{1-\alpha}, +\infty)$
(c)	$\mu = \mu_0$	$\mu \neq \mu_0$	$(-\infty, -t(N-1)_{1-\alpha}) \cup (t(N-1)_{1-\alpha}, +\infty)$

wobei $t(N-1)_u$ den u-Quantile der t-Verteilung mit $N-1$ Freiheitsgraden bezeichnet. Dieser kann in den Quantiltabellen ablesen werden (siehe die Quantiletablellen am Ende des Skripts).

10.4 Testen von Anteilswerten

In diesem Abschnitt werden Tests für Anteilswerte beschrieben. Beim ersten Test wird der Anteil eines Merkmals in der Grundgesamtheit mit einem vorgegebenen Wert verglichen. Der zweite Test gestattet den Vergleich zweier Anteilswerte. Bei beiden Tests handelt es sich um approximative Tests, die bei großen Stichprobenumfängen angewandt werden sollen.

Im Folgenden soll beschrieben werden, wie der Anteil eines Merkmals in der Grundgesamtheit gegen einen konkreten Wert p_0 getestet werden kann. Der im letzten Kapitel eingangs besprochene Binomial-Test wird in der Praxis meist nur bei sehr kleinen Stichprobenumfängen verwendet. Ist der Stichprobenumfang ausreichend groß, kann für den Stichprobenanteil angenommen werden, dass er nach dem zentralen Grenzwertsatz approximativ normal verteilt ist.

Der Stichprobenanteil kann als der Durchschnitt von unabhängigen Bernoulli-Variablen $B(1,p)$ geschrieben werden, wobei p dem Anteilssatz in der Grundgesamtheit entspricht:

$$X_i = \begin{cases} 1 & bei\ der\ i-ten\ Ziehung\ wird\ das\ Merkmal\ beobachtet \\ 0 & bei\ der\ i-ten\ Ziehung\ wird\ das\ Merkmal\ nicht\ beobachtet \end{cases}$$

Dann ist $X_i \sim B(1,p)$, $i = 1, \dots, N$.

Null- und Alternative-Hypothesen:

p_0 stellt einen konkreten Wert dar, gegen den man den unbekannten Parameter p prüfen möchte.

Fall	Null Hypothese	Alternativ-Hypothese	Testproblem
(a)	$p \geq p_0$	$p < p_0$	einseitig
(b)	$p \leq p_0$	$p > p_0$	einseitig
(c)	$p = p_0$	$p \neq p_0$	zweiseitig

Prüfgröße

$$T = \frac{\bar{p} - p_0}{\sqrt{\frac{\bar{p}(1-\bar{p})}{N}}}$$

wobei $\bar{p} = \overline{X_n} = (X_1 + \cdots + X_N)/N$ den arithmetischen Mittelwert der Beobachtungen bezeichnet .

Ablehnungsbereich (zum Signifinkanzniveau $\boldsymbol{\alpha}$)

Es kann nachgewiesen werden, dass die Prüfgröße (under der Hypothese $\boldsymbol{H_0}$) für große Werte von $\boldsymbol{N}$ approximativ Normalverteilt ist. Daher ergibt der folgende Ablehnungsbereich:

Null	Alternativ-

Fall	Hypothese	Hypothese	Ablehnungsbereich
(a)	$\mu \geq \mu_0$	$\mu < \mu_0$	$K = (-\infty, -z_{1-\alpha})$
(b)	$\mu \leq \mu_0$	$\mu > \mu_0$	$K = (z_{1-\alpha}, +\infty)$
(c)	$\mu = \mu_0$	$\mu \neq \mu_0$	$K = (-\infty, -z_{1-\alpha}) \cup (z_{1-\alpha}, +\infty)$

wobei z_u den u-Quantile der Normalverteilung bezeichnet. Dieser kann in den Quantiltabellen ablesen werden (siehe die Quantiletablellen am Ende des Skripts).

Als Faustregel für den nötigen Stchprobenumfang N in Abhängigkeit von $\bar{p}$ sollte sowohl $N\bar{p} \geq 5$als auch $N(1-\bar{p}) \geq 5$ erfüllt werden.

Beispiel

In einer Wahlumfrage werden 1000 wahlberechtigte Personen befragt, welche Partei sie wählen würden, wenn am nächsten Sonntag Bundestagswahl wäre. Es soll zu einem Signifikanzniveau von 5 % überprüft werden, ob eine bestimmte Partei die 5 % Hürde schaffen würde. Bezeichnet p den Anteil der Wähler dieser Partei, so lauten die Hypothesen

$$H_0 : p > 0{,}05 \quad \text{und} \quad H_1 : p < 0{,}05.$$

Im Falle, dass die Nullhypothese abgelehnt wird, ist statistisch bewiesen, dass der Anteil der Stimmen unter 5 % liegt. Bei der Umfrage antworten 35 Personen, dass sie diese Partei wählen würden. Daraus ergibt sich eine Testgröße von

$$T = \frac{0{,}035 - 0{,}05}{\sqrt{\frac{0{,}035(1 - 0{,}035)}{1000}}} = -2{,}58.$$

Der kritische Bereich zu einem Signifikanzniveau von $\alpha = 0{,}05$ ist durch $K = (-\infty; -1{,}64)$ gegeben.

Da die Testgröße im kritischen Bereich liegt, ist statistisch bewiesen, dass die Partei zur Zeit der Umfrage auf einen Stimmenanteil von unter 5 % kommen würde.

10.5 Unabhängigkeitstest

Der χ^2-Unabhängigkeitstest überprüft, ob zwei kategoriale Merkmale unabhängig voneinander sind. Auch können Ausprägungen diskreter Merkmale als Kategorien aufgefasst werden. Stetige Merkmale können zudem in Klassen eingeteilt werden, sodass man mithilfe des r χ^2-Unabhängigkeitstests allgemein Auskunft über die Unabhängigkeit von Merkmalen erhalten kann.

Formulierung der Modellannahmen

Es liegen N unabhängige Ziehungen der Zufallsvariablen X und Y vor, die die Ausprägungen 1 bis k und 1 bis l annehmen:

Null- und Alternative-Hypothesen:

$$H_0: P(X = i, Y = f) = P(X = i)P(Y = j) \qquad \text{für alle } i = j = 1, \ldots, Z$$

gegen

$$H_1 : P(X = i, Y = j)P[X = i)P[Y = j) \qquad \text{für mindestens eine Kombination } i, j.$$

Prüfgröße

$$T = \sum_{i=1}^{k} \sum_{j=1}^{l} \frac{\left(f_{ij} - \frac{f_i * f_j}{N}\right)^2}{\frac{f_i * f_j}{N}}$$

wobei f_{ij} die gemeinsame Haufigkeit darstellt und f_i und f_j die Randverteilungen sind.

Es kann nachgewiesen werden, dass die Prüfgröße (under der Hypothese H_0) χ^2-verteilt ist mit $A = (k-1)(l-1)$Freiheitsgraden. Daher erginbt sich der folgende Ablehnungsbereich:

Ablehnungsbereich (zum Signifinkanzniveau α)

Der Ablehnungsbereich ist definitert als

$$K = (\boldsymbol{c}\big((k-1)(l-1)\big)_{1-\alpha}, +\infty)$$

wobei $\boldsymbol{c}\big((k-1)(l-1)\big)_{\boldsymbol{u}}$ den u-Quantile der χ^2-Verteilung mit $(k-1)(l-1)$ Freiheitsgraden bezeichnet. Dieser kann in den Quantiltabellen ablesen werden (siehe die Quantiletablellen am Ende des Skripts).

11. Quantile-Tabellen

Quantile der Standardnormalverteilung

Tabelliert sind die Werte der Verteilungsfunktion, Die Zeilen entsprechen der Vorkomma- und der ersten Nachkommastelle, die Spalten der zweiten Nachkommastelle des Werts.
Ablesebeispiel: $\Phi(1{,}75) = 0{,}9599$ Werte kleiner null: $\Phi(-x) = 1 - \Phi(x)$.

Tabelle Verteilungsfunktion der Standardnormalverteilung

	0	0,01	0,02	0,03	0,04	0,05	0,06	0,07	0,08	0,09
0	0,5000	0,5040	0,5080	0,5120	0,5160	0,5199	0,5239	0,5279	0,5319	0,5359
0,1	0,5398	0,5438	0,5478	0,5517	0,5557	0,5596	0,5636	0,5675	0,5714	0,5753
0,2	0,5793	0,5832	0,5871	0,5910	0,5948	0,5987	0,6026	0,6064	0,6103	0,6141
0,3	0,6179	0,6217	0,6255	0,6293	0,6331	0,6368	0,6406	0,6443	0,6480	0,6517
0,4	0,6554	0,6591	0,6628	0,6664	0,6700	0,6736	0,6772	0,6808	0,6844	0,6879
0,5	0,6915	0,6950	0,6985	0,7019	0,7054	0,7088	0,7123	0,7157	0,7190	0,7224
0,6	0,7257	0,7291	0,7324	0,7357	0,7389	0,7422	0,7454	0,7486	0,7517	0,7549
0,7	0,7580	0,7611	0,7642	0,7673	0,7704	0,7734	0,7764	0,7794	0,7823	0,7852
0,8	0,7881	0,7910	0,7939	0,7967	0,7995	0,8023	0,8051	0,8078	0,8106	0,8133
0,9	0,8159	0,8186	0,8212	0,8238	0,8264	0,8289	0,8315	0,8340	0,8365	0,8389
1	0,8413	0,8438	0,8461	0,8485	0,8508	0,8531	0,8554	0,8577	0,8599	0,8621
1,1	0,8643	0,8665	0,8686	0,8708	0,8729	0,8749	0,8770	0,8790	0,8810	0,8830
1,2	0,8849	0,8869	0,8888	0,8907	0,8925	0,8944	0,8962	0,8980	0,8997	0,9015
1,3	0,9032	0,9049	0,9066	0,9082	0,9099	0,9115	0,9131	0,9147	0,9162	0,9177
1,4	0,9192	0,9207	0,9222	0,9236	0,9251	0,9265	0,9279	0,9292	0,9306	0,9319
1,5	0,9332	0,9345	0,9357	0,9370	0,9382	0,9394	0,9406	0,9418	0,9429	0,9441
1,6	0,9452	0,9463	0,9474	0,9484	0,9495	0,9505	0,9515	0,9525	0,9535	0,9545
1,7	0,9554	0,9564	0,9573	0,9582	0,9591	0,9599	0,9608	0,9616	0,9625	0,9633
1,8	0,9641	0,9649	0,9656	0,9664	0,9671	0,9678	0,9686	0,9693	0,9699	0,9706
1,9	0,9713	0,9719	0,9726	0,9732	0,9738	0,9744	0,9750	0,9756	0,9761	0,9767
2	0,9772	0,9778	0,9783	0,9788	0,9793	0,9798	0,9803	0,9808	0,9812	0,9817
2,1	0,9821	0,9826	0,9830	0,9834	0,9838	0,9842	0,9846	0,9850	0,9854	0,9857
2,2	0,9861	0,9864	0,9868	0,9871	0,9875	0,9878	0,9881	0,9884	0,9887	0,9890
2,3	0,9893	0,9896	0,9898	0,9901	0,9904	0,9906	0,9909	0,9911	0,9913	0,9916
2,4	0,9918	0,9920	0,9922	0,9925	0,9927	0,9929	0,9931	0,9932	0,9934	0,9936
2,5	0,9938	0,9940	0,9941	0,9943	0,9945	0,9946	0,9948	0,9949	0,9951	0,9952
2,6	0,9953	0,9955	0,9956	0,9957	0,9959	0,9960	0,9961	0,9962	0,9963	0,9964
2,7	0,9965	0,9966	0,9967	0,9968	0,9969	0,9970	0,9971	0,9972	0,9973	0,9974
2,8	0,9974	0,9975	0,9976	0,9977	0,9977	0,9978	0,9979	0,9979	0,9980	0,9981
2,9	0,9981	0,9982	0,9982	0,9983	0,9984	0,9984	0,9985	0,9985	0,9986	0,9986
3	0,9987	0,9987	0,9987	0,9988	0,9988	0,9989	0,9989	0,9989	0,9990	0,9990
3,1	0,9990	0,9991	0,9991	0,9991	0,9992	0,9992	0,9992	0,9992	0,9993	0,9993
3,2	0,9993	0,9993	0,9994	0,9994	0,9994	0,9994	0,9994	0,9995	0,9995	0,9995
3,3	0,9995	0,9995	0,9995	0,9996	0,9996	0,9996	0,9996	0,9996	0,9996	0,9997
3,4	0,9997	0,9997	0,9997	0,9997	0,9997	0,9997	0,9997	0,9997	0,9997	0,9998
3,5	0,9998	0,9998	0,9998	0,9998	0,9998	0,9998	0,9998	0,9998	0,9998	0,9998
3,6	0,9998	0,9998	0,9999	0,9999	0,9999	0,9999	0,9999	0,9999	0,9999	0,9999
3,7	0,9999	0,9999	0,9999	0,9999	0,9999	0,9999	0,9999	0,9999	0,9999	0,9999
3,8	0,9999	0,9999	0,9999	0,9999	0,9999	0,9999	0,9999	0,9999	0,9999	0,9999
3,9	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000

Quantile der χ^2-Verteilung

Tabelliert sind die Quantile (Spalten) der χ^2-Verteilung. Freiheitsgrade: Zeilen (Bsp.:$c(lO)_{0,95} = 18{,}307$)

df	0,01	0,025	0,05	0,1	0,9	0,95	0,975	0,99
1	0,000	0,001	0,004	0,016	2,706	3,841	5,024	6,635
2	0,020	0,051	0,103	0,211	4,605	5,991	7,378	9,210
3	0,115	0,216	0,352	0,584	6,251	7,815	9,348	11,345
4	0,297	0,484	0,711	1,064	7,779	9,488	11,143	13,277
5	0,554	0,831	1,145	1,610	9,236	11,070	12,833	15,086
6	0,872	1,237	1,635	2,204	10,645	12,592	14,449	16,812
7	1,239	1,690	2,167	2,833	12,017	14,067	16,013	18,475
8	1,646	2,180	2,733	3,490	13,362	15,507	17,535	20,090
9	2,088	2,700	3,325	4,168	14,684	16,919	19,023	21,666
10	2,558	3,247	3,940	4,865	15,987	18,307	20,483	23,209
11	3,053	3,816	4,575	5,578	17,275	19,675	21,920	24,725
12	3,571	4,404	5,226	6,304	18,549	21,026	23,337	26,217
13	4,107	5,009	5,892	7,042	19,812	22,362	24,736	27,688
14	4,660	5,629	6,571	7,790	21,064	23,685	26,119	29,141
15	5,229	6,262	7,261	8,547	22,307	24,996	27,488	30,578
16	5,812	6,908	7,962	9,312	23,542	26,296	28,845	32,000
17	6,408	7,564	8,672	10,085	24,769	27,587	30,191	33,409
18	7,015	8,231	9,390	10,865	25,989	28,869	31,526	34,805
19	7,633	8,907	10,117	11,651	27,204	30,144	32,852	36,191
20	8,260	9,591	10,851	12,443	28,412	31,410	34,170	37,566
21	8,897	10,283	11,591	13,240	29,615	32,671	35,479	38,932
22	9,542	10,982	12,338	14,041	30,813	33,924	36,781	40,289
23	10,196	11,689	13,091	14,848	32,007	35,172	38,076	41,638
24	10,856	12,401	13,848	15,659	33,196	36,415	39,364	42,980
25	11,524	13,120	14,611	16,473	34,382	37,652	40,646	44,314
30	14,953	16,791	18,493	20,599	40,256	43,773	46,979	50,892
40	22,164	24,433	26,509	29,051	51,805	55,758	59,342	63,691
50	29,707	32,357	34,764	37,689	63,167	67,505	71,420	76,154
60	37,485	40,482	43,188	46,459	74,397	79,082	83,298	88,379
70	45,442	48,758	51,739	55,329	85,527	90,531	95,023	100,425
80	53,540	57,153	60,391	64,278	96,578	101,879	106,629	112,329
90	61,754	65,647	69,126	73,291	107,565	113,145	118,136	124,116
100	70,065	74,222	77,929	82,358	118,498	124,342	129,561	135,807

Quantile der t −Verteilung

Tabelliert sind die Quantile (Spalten) der t-Verteilung. Freiheitsgrade: Zeilen (Bsp.: $t(7)_{0,99} = 2{,}998$)

df	0,65	0,7	0,75	0,8	0,85	0,9	0,95	0,975	0,99	0,995
1	0,510	0,727	1,000	1,376	1,963	3,078	6,314	12,706	31,821	63,656
2	0,445	0,617	0,816	1,061	1,386	1,886	2,920	4,303	6,965	9,925
3	0,424	0,584	0,765	0,978	1,250	1,638	2,353	3,182	4,541	5,841
4	0,414	0,569	0,741	0,941	1,190	1,533	2,132	2,776	3,747	4,604
5	0,408	0,559	0,727	0,920	1,156	1,476	2,015	2,571	3,365	4,032
6	0,404	0,553	0,718	0,906	1,134	1,440	1,943	2,447	3,143	3,707
7	0,402	0,549	0,711	0,896	1,119	1,415	1,895	2,365	2,998	3,499
8	0,399	0,546	0,706	0,889	1,108	1,397	1,860	2,306	2,896	3,355
9	0,398	0,543	0,703	0,883	1,100	1,383	1,833	2,262	2,821	3,250
10	0,397	0,542	0,700	0,879	1,093	1,372	1,812	2,228	2,764	3,169
11	0,396	0,540	0,697	0,876	1,088	1,363	1,796	2,201	2,718	3,106
12	0,395	0,539	0,695	0,873	1,083	1,356	1,782	2,179	2,681	3,055
13	0,394	0,538	0,694	0,870	1,079	1,350	1,771	2,160	2,650	3,012
14	0,393	0,537	0,692	0,868	1,076	1,345	1,761	2,145	2,624	2,977
15	0,393	0,536	0,691	0,866	1,074	1,341	1,753	2,131	2,602	2,947
16	0,392	0,535	0,690	0,865	1,071	1,337	1,746	2,120	2,583	2,921
17	0,392	0,534	0,689	0,863	1,069	1,333	1,740	2,110	2,567	2,898
18	0,392	0,534	0,688	0,862	1,067	1,330	1,734	2,101	2,552	2,878
19	0,391	0,533	0,688	0,861	1,066	1,328	1,729	2,093	2,539	2,861
20	0,391	0,533	0,687	0,860	1,064	1,325	1,725	2,086	2,528	2,845
21	0,391	0,532	0,686	0,859	1,063	1,323	1,721	2,080	2,518	2,831
22	0,390	0,532	0,686	0,858	1,061	1,321	1,717	2,074	2,508	2,819
23	0,390	0,532	0,685	0,858	1,060	1,319	1,714	2,069	2,500	2,807
24	0,390	0,531	0,685	0,857	1,059	1,318	1,711	2,064	2,492	2,797
25	0,390	0,531	0,684	0,856	1,058	1,316	1,708	2,060	2,485	2,787
30	0,389	0,530	0,683	0,854	1,055	1,310	1,697	2,042	2,457	2,750
40	0,388	0,529	0,681	0,851	1,050	1,303	1,684	2,021	2,423	2,704
50	0,388	0,528	0,679	0,849	1,047	1,299	1,676	2,009	2,403	2,678
60	0,387	0,527	0,679	0,848	1,045	1,296	1,671	2,000	2,390	2,660
70	0,387	0,527	0,678	0,847	1,044	1,294	1,667	1,994	2,381	2,648
80	0,387	0,526	0,678	0,846	1,043	1,292	1,664	1,990	2,374	2,639
90	0,387	0,526	0,677	0,846	1,042	1,291	1,662	1,987	2,368	2,632
100	0,386	0,526	0,677	0,845	1,042	1,290	1,660	1,984	2,364	2,626

Quantile der F −Verteilung

Tabelliert sind die Quantile der F −Verteilung, $f(dfi, df2)_{1-\alpha}$, Für die $1-\alpha$- Quantile $0{,}9;\ 0{,}95;\ 0{,}975; 0{,}99;\ 0{,}995$ sind jeweils eigene Tabellen abgedruckt. Das Quantil entspricht dem Wert im ersten Kästchen der Tabelle ganz oben. Ablesebeispiel: /(7;4)Q,9 = 3,979

Tabelle: $0{,}9-$ Quantil der F −Verteilung $df_2 = 1, \ldots, 10$

0,9				df_2						
dfl	1	2	3	4	5	6	7	8	9	10
1	39,863	8,526	5,538	4,545	4,060	3,776	3,589	3,458	3,360	3,285
2	49,500	9,000	5,462	4,325	3,780	3,463	3,257	3,113	3,006	2,924
3	53,593	9,162	5,391	4,191	3,619	3,289	3,074	2,924	2,813	2,728
4	55,833	9,243	5,343	4,107	3,520	3,181	2,961	2,806	2,693	2,605
5	57,240	9,293	5,309	4,051	3,453	3,108	2,883	2,726	2,611	2,522
6	58,204	9,326	5,285	4,010	3,405	3,055	2,827	2,668	2,551	2,461
7	58,906	9,349	5,266	3,979	3,368	3,014	2,785	2,624	2,505	2,414
8	59,439	9,367	5,252	3,955	3,339	2,983	2,752	2,589	2,469	2,377
9	59,858	9,381	5,240	3,936	3,316	2,958	2,725	2,561	2,440	2,347
10	60,195	9,392	5,230	3,920	3,297	2,937	2,703	2,538	2,416	2,323
11	60,473	9,401	5,222	3,907	3,282	2,920	2,684	2,519	2,396	2,302
12	60,705	9,408	5,216	3,896	3,268	2,905	2,668	2,502	2,379	2,284
13	60,903	9,415	5,210	3,886	3,257	2,892	2,654	2,488	2,364	2,269
14	61,073	9,420	5,205	3,878	3,247	2,881	2,643	2,475	2,351	2,255
15	61,220	9,425	5,200	3,870	3,238	2,871	2,632	2,464	2,340	2,244
16	61,350	9,429	5,196	3,864	3,230	2,863	2,623	2,455	2,329	2,233
17	61,464	9,433	5,193	3,858	3,223	2,855	2,615	2,446	2,320	2,224
18	61,566	9,436	5,190	3,853	3,217	2,848	2,607	2,438	2,312	2,215
19	61,658	9,439	5,187	3,849	3,212	2,842	2,601	2,431	2,305	2,208
20	61,740	9,441	5,184	3,844	3,207	2,836	2,595	2,425	2,298	2,201
30	62,265	9,458	5,168	3,817	3,174	2,800	2,555	2,383	2,255	2,155
40	62,529	9,466	5,160	3,804	3,157	2,781	2,535	2,361	2,232	2,132
50	62,688	9,471	5,155	3,795	3,147	2,770	2,523	2,348	2,218	2,117
60	62,794	9,475	5,151	3,790	3,140	2,762	2,514	2,339	2,208	2,107
70	62,870	9,477	5,149	3,786	3,135	2,756	2,508	2,333	2,202	2,100
80	62,927	9,479	5,147	3,782	3,132	2,752	2,504	2,328	2,196	2,095
90	62,972	9,480	5,145	3,780	3,129	2,749	2,500	2,324	2,192	2,090
100	63,007	9,481	5,144	3,778	3,126	2,746	2,497	2,321	2,189	2,087
110	63,036	9,482	5,143	3,777	3,124	2,744	2,495	2,318	2,187	2,084
120	63,061	9,483	5,143	3,775	3,123	2,742	2,493	2,316	2,184	2,082
130	63,081	9,484	5,142	3,774	3,121	2,741	2,491	2,314	2,182	2,080
140	63,099	9,484	5,141	3,773	3,120	2,739	2,490	2,313	2,181	2,078
150	63,114	9,485	5,141	3,772	3,119	2,738	2,488	2,312	2,179	2,077

Tabelle: 0,9 − Quantil der F −Verteilung $df_2 = 20, \ldots, 100$

0,9				df_2					
dfi	20	30	40	50	60	70	80	90	100
1	2,975	2,881	2,835	2,809	2,791	2,779	2,769	2,762	2,756
2	2,589	2,489	2,440	2,412	2,393	2,380	2,370	2,363	2,356
3	2,380	2,276	2,226	2,197	2,177	2,164	2,154	2,146	2,139
4	2,249	2,142	2,091	2,061	2,041	2,027	2,016	2,008	2,002
5	2,158	2,049	1,997	1,966	1,946	1,931	1,921	1,912	1,906
6	2,091	1,980	1,927	1,895	1,875	1,860	1,849	1,841	1,834
7	2,040	1,927	1,873	1,840	1,819	1,804	1,793	1,785	1,778
8	1,999	1,884	1,829	1,796	1,775	1,760	1,748	1,739	1,732
9	1,965	1,849	1,793	1,760	1,738	1,723	1,711	1,702	1,695
10	1,937	1,819	1,763	1,729	1,707	1,691	1,680	1,670	1,663
11	1,913	1,794	1,737	1,703	1,680	1,665	1,653	1,643	1,636
12	1,892	1,773	1,715	1,680	1,657	1,641	1,629	1,620	1,612
13	1,875	1,754	1,695	1,660	1,637	1,621	1,609	1,599	1,592
14	1,859	1,737	1,678	1,643	1,619	1,603	1,590	1,581	1,573
15	1,845	1,722	1,662	1,627	1,603	1,587	1,574	1,564	1,557
16	1,833	1,709	1,649	1,613	1,589	1,572	1,559	1,550	1,542
17	1,821	1,697	1,636	1,600	1,576	1,559	1,546	1,536	1,528
18	1,811	1,686	1,625	1,588	1,564	1,547	1,534	1,524	1,516
19	1,802	1,676	1,615	1,578	1,553	1,536	1,523	1,513	1,505
20	1,794	1,667	1,605	1,568	1,543	1,526	1,513	1,503	1,494
30	1,738	1,606	1,541	1,502	1,476	1,457	1,443	1,432	1,423
40	1,708	1,573	1,506	1,465	1,437	1,418	1,403	1,391	1,382
50	1,690	1,552	1,483	1,441	1,413	1,392	1,377	1,365	1,355
60	1,677	1,538	1,467	1,424	1,395	1,374	1,358	1,346	1,336
70	1,667	1,527	1,455	1,412	1,382	1,361	1,344	1,332	1,321
80	1,660	1,519	1,447	1,402	1,372	1,350	1,334	1,321	1,310
90	1,655	1,512	1,439	1,395	1,364	1,342	1,325	1,312	1,301
100	1,650	1,507	1,434	1,388	1,358	1,335	1,318	1,304	1,293
110	1,646	1,503	1,429	1,383	1,352	1,329	1,312	1,298	1,287
120	1,643	1,499	1,425	1,379	1,348	1,325	1,307	1,293	1,282
130	1,641	1,496	1,421	1,375	1,344	1,320	1,303	1,289	1,277
140	1,638	1,493	1,418	1,372	1,340	1,317	1,299	1,285	1,273
150	1,636	1,491	1,416	1,369	1,337	1,314	1,296	1,281	1,270

Tabelle: $0{,}95-$ Quantil der $F-$Verteilung $df_2 = 1, \ldots, 10$

			df_2						
1	2	3	4	5	6	7	8	9	10
161,448	18,513	10,128	7,709	6,608	5,987	5,591	5,318	5,117	4,965
199,500	19,000	9,552	6,944	5,786	5,143	4,737	4,459	4,256	4,103
215,707	19,164	9,277	6,591	5,409	4,757	4,347	4,066	3,863	3,708
224,583	19,247	9,117	6,388	5,192	4,534	4,120	3,838	3,633	3,478
230,162	19,296	9,013	6,256	5,050	4,387	3,972	3,687	3,482	3,326
233,986	19,330	8,941	6,163	4,950	4,284	3,866	3,581	3,374	3,217
236,768	19,353	8,887	6,094	4,876	4,207	3,787	3,500	3,293	3,135
238,883	19,371	8,845	6,041	4,818	4,147	3,726	3,438	3,230	3,072
240,543	19,385	8,812	5,999	4,772	4,099	3,677	3,388	3,179	3,020
241,882	19,396	8,786	5,964	4,735	4,060	3,637	3,347	3,137	2,978
242,983	19,405	8,763	5,936	4,704	4,027	3,603	3,313	3,102	2,943
243,906	19,413	8,745	5,912	4,678	4,000	3,575	3,284	3,073	2,913
244,690	19,419	8,729	5,891	4,655	3,976	3,550	3,259	3,048	2,887
245,364	19,424	8,715	5,873	4,636	3,956	3,529	3,237	3,025	2,865
245,950	19,429	8,703	5,858	4,619	3,938	3,511	3,218	3,006	2,845
246,464	19,433	8,692	5,844	4,604	3,922	3,494	3,202	2,989	2,828
246,918	19,437	8,683	5,832	4,590	3,908	3,480	3,187	2,974	2,812
247,323	19,440	8,675	5,821	4,579	3,896	3,467	3,173	2,960	2,798
247,686	19,443	8,667	5,811	4,568	3,884	3,455	3,161	2,948	2,785
248,013	19,446	8,660	5,803	4,558	3,874	3,445	3,150	2,936	2,774
250,095	19,462	8,617	5,746	4,496	3,808	3,376	3,079	2,864	2,700
251,143	19,471	8,594	5,717	4,464	3,774	3,340	3,043	2,826	2,661
251,774	19,476	8,581	5,699	4,444	3,754	3,319	3,020	2,803	2,637
252,196	19,479	8,572	5,688	4,431	3,740	3,304	3,005	2,787	2,621
252,497	19,481	8,566	5,679	4,422	3,730	3,294	2,994	2,776	2,610
252,724	19,483	8,561	5,673	4,415	3,722	3,286	2,986	2,768	2,601
252,900	19,485	8,557	5,668	4,409	3,716	3,280	2,980	2,761	2,594
253,041	19,486	8,554	5,664	4,405	3,712	3,275	2,975	2,756	2,588
253,157	19,487	8,551	5,661	4,401	3,708	3,271	2,970	2,751	2,584
253,253	19,487	8,549	5,658	4,398	3,705	3,267	2,967	2,748	2,580
253,334	19,488	8,548	5,656	4,396	3,702	3,265	2,964	2,744	2,577
253,404	19,489	8,546	5,654	4,394	3,700	3,262	2,961	2,742	2,574
253,465	19,489	8,545	5,652	4,392	3,698	3,260	2,959	2,739	2,572

Tabelle: 0,95 − Quantil der F −Verteilung $df_2 = 20, \ldots, 100$

0,95	df_2								
dfi	20	30	40	50	60	70	80	90	100
1	4,351	4,171	4,085	4,034	4,001	3,978	3,960	3,947	3,936
2	3,493	3,316	3,232	3,183	3,150	3,128	3,111	3,098	3,087
3	3,098	2,922	2,839	2,790	2,758	2,736	2,719	2,706	2,696
4	2,866	2,690	2,606	2,557	2,525	2,503	2,486	2,473	2,463
5	2,711	2,534	2,449	2,400	2,368	2,346	2,329	2,316	2,305
6	2,599	2,421	2,336	2,286	2,254	2,231	2,214	2,201	2,191
7	2,514	2,334	2,249	2,199	2,167	2,143	2,126	2,113	2,103
8	2,447	2,266	2,180	2,130	2,097	2,074	2,056	2,043	2,032
9	2,393	2,211	2,124	2,073	2,040	2,017	1,999	1,986	1,975
10	2,348	2,165	2,077	2,026	1,993	1,969	1,951	1,938	1,927
11	2,310	2,126	2,038	1,986	1,952	1,928	1,910	1,897	1,886
12	2,278	2,092	2,003	1,952	1,917	1,893	1,875	1,861	1,850
13	2,250	2,063	1,974	1,921	1,887	1,863	1,845	1,830	1,819
14	2,225	2,037	1,948	1,895	1,860	1,836	1,817	1,803	1,792
15	2,203	2,015	1,924	1,871	1,836	1,812	1,793	1,779	1,768
16	2,184	1,995	1,904	1,850	1,815	1,790	1,772	1,757	1,746
17	2,167	1,976	1,885	1,831	1,796	1,771	1,752	1,737	1,726
18	2,151	1,960	1,868	1,814	1,778	1,753	1,734	1,720	1,708
19	2,137	1,945	1,853	1,798	1,763	1,737	1,718	1,703	1,691
20	2,124	1,932	1,839	1,784	1,748	1,722	1,703	1,688	1,676
30	2,039	1,841	1,744	1,687	1,649	1,622	1,602	1,586	1,573
40	1,994	1,792	1,693	1,634	1,594	1,566	1,545	1,528	1,515
50	1,966	1,761	1,660	1,599	1,559	1,530	1,508	1,491	1,477
60	1,946	1,740	1,637	1,576	1,534	1,505	1,482	1,465	1,450
70	1,932	1,724	1,621	1,558	1,516	1,486	1,463	1,445	1,430
80	1,922	1,712	1,608	1,544	1,502	1,471	1,448	1,429	1,415
90	1,913	1,703	1,597	1,534	1,491	1,459	1,436	1,417	1,402
100	1,907	1,695	1,589	1,525	1,481	1,450	1,426	1,407	1,392
110	1,901	1,689	1,582	1,518	1,474	1,442	1,418	1,399	1,383
120	1,896	1,683	1,577	1,511	1,467	1,435	1,411	1,391	1,376
130	1,892	1,679	1,572	1,506	1,462	1,429	1,405	1,385	1,369
140	1,889	1,675	1,567	1,502	1,457	1,424	1,399	1,380	1,364
150	1,886	1,672	1,564	1,498	1,453	1,420	1,395	1,375	1,359

Tabelle: $0{,}975-$ Quantil der $F-$Verteilung $df_2 = 1, \ldots, 10$

			df_2						
1	2	3	4	5	6	7	8	9	10
647,789	38,506	17,443	12,218	10,007	8,813	8,073	7,571	7,209	6,937
799,500	39,000	16,044	10,649	8,434	7,260	6,542	6,059	5,715	5,456
864,163	39,165	15,439	9,979	7,764	6,599	5,890	5,416	5,078	4,826
899,583	39,248	15,101	9,605	7,388	6,227	5,523	5,053	4,718	4,468
921,848	39,298	14,885	9,364	7,146	5,988	5,285	4,817	4,484	4,236
937,111	39,331	14,735	9,197	6,978	5,820	5,119	4,652	4,320	4,072
948,217	39,355	14,624	9,074	6,853	5,695	4,995	4,529	4,197	3,950
956,656	39,373	14,540	8,980	6,757	5,600	4,899	4,433	4,102	3,855
963,285	39,387	14,473	8,905	6,681	5,523	4,823	4,357	4,026	3,779
968,627	39,398	14,419	8,844	6,619	5,461	4,761	4,295	3,964	3,717
973,025	39,407	14,374	8,794	6,568	5,410	4,709	4,243	3,912	3,665
976,708	39,415	14,337	8,751	6,525	5,366	4,666	4,200	3,868	3,621
979,837	39,421	14,304	8,715	6,488	5,329	4,628	4,162	3,831	3,583
982,528	39,427	14,277	8,684	6,456	5,297	4,596	4,130	3,798	3,550
984,867	39,431	14,253	8,657	6,428	5,269	4,568	4,101	3,769	3,522
986,919	39,435	14,232	8,633	6,403	5,244	4,543	4,076	3,744	3,496
988,733	39,439	14,213	8,611	6,381	5,222	4,521	4,054	3,722	3,474
990,349	39,442	14,196	8,592	6,362	5,202	4,501	4,034	3,701	3,453
991,797	39,445	14,181	8,575	6,344	5,184	4,483	4,016	3,683	3,435
993,103	39,448	14,167	8,560	6,329	5,168	4,467	3,999	3,667	3,419
1001,414	39,465	14,081	8,461	6,227	5,065	4,362	3,894	3,560	3,311
1005,598	39,473	14,037	8,411	6,175	5,012	4,309	3,840	3,505	3,255
1008,117	39,478	14,010	8,381	6,144	4,980	4,276	3,807	3,472	3,221
1009,800	39,481	13,992	8,360	6,123	4,959	4,254	3,784	3,449	3,198
1011,004	39,484	13,979	8,346	6,107	4,943	4,239	3,768	3,433	3,182
1011,908	39,485	13,970	8,335	6,096	4,932	4,227	3,756	3,421	3,169
1012,612	39,487	13,962	8,326	6,087	4,923	4,218	3,747	3,411	3,160
1013,175	39,488	13,956	8,319	6,080	4,915	4,210	3,739	3,403	3,152
1013,636	39,489	13,951	8,314	6,074	4,909	4,204	3,733	3,397	3,145
1014,020	39,490	13,947	8,309	6,069	4,904	4,199	3,728	3,392	3,140
1014,346	39,490	13,944	8,305	6,065	4,900	4,195	3,724	3,387	3,135
1014,625	39,491	13,941	8,302	6,062	4,897	4,191	3,720	3,383	3,131
1014,866	39,491	13,938	8,299	6,059	4,893	4,188	3,716	3,380	3,128

Tabelle: 0,975 − Quantil der F −Verteilung $df_2 = 20, \ldots, 100$

0,975				df_2					
äfi	20	30	40	50	60	70	80	90	100
1	5,871	5,568	5,424	5,340	5,286	5,247	5,218	5,196	5,179
2	4,461	4,182	4,051	3,975	3,925	3,890	3,864	3,844	3,828
3	3,859	3,589	3,463	3,390	3,343	3,309	3,284	3,265	3,250
4	3,515	3,250	3,126	3,054	3,008	2,975	2,950	2,932	2,917
5	3,289	3,026	2,904	2,833	2,786	2,754	2,730	2,711	2,696
6	3,128	2,867	2,744	2,674	2,627	2,595	2,571	2,552	2,537
7	3,007	2,746	2,624	2,553	2,507	2,474	2,450	2,432	2,417
8	2,913	2,651	2,529	2,458	2,412	2,379	2,355	2,336	2,321
9	2,837	2,575	2,452	2,381	2,334	2,302	2,277	2,259	2,244
10	2,774	2,511	2,388	2,317	2,270	2,237	2,213	2,194	2,179
11	2,721	2,458	2,334	2,263	2,216	2,183	2,158	2,140	2,124
12	2,676	2,412	2,288	2,216	2,169	2,136	2,111	2,092	2,077
13	2,637	2,372	2,248	2,176	2,129	2,095	2,071	2,051	2,036
14	2,603	2,338	2,213	2,140	2,093	2,059	2,035	2,015	2,000
15	2,573	2,307	2,182	2,109	2,061	2,028	2,003	1,983	1,968
16	2,547	2,280	2,154	2,081	2,033	1,999	1,974	1,955	1,939
17	2,523	2,255	2,129	2,056	2,008	1,974	1,948	1,929	1,913
18	2,501	2,233	2,107	2,033	1,985	1,950	1,925	1,905	1,890
19	2,482	2,213	2,086	2,012	1,964	1,929	1,904	1,884	1,868
20	2,464	2,195	2,068	1,993	1,944	1,910	1,884	1,864	1,849
30	2,349	2,074	1,943	1,866	1,815	1,779	1,752	1,731	1,715
40	2,287	2,009	1,875	1,796	1,744	1,707	1,679	1,657	1,640
50	2,249	1,968	1,832	1,752	1,699	1,660	1,632	1,610	1,592
60	2,223	1,940	1,803	1,721	1,667	1,628	1,599	1,576	1,558
70	2,205	1,920	1,781	1,698	1,643	1,604	1,574	1,551	1,532
80	2,190	1,904	1,764	1,681	1,625	1,585	1,555	1,531	1,512
90	2,179	1,892	1,751	1,667	1,611	1,570	1,540	1,516	1,496
100	2,170	1,882	1,741	1,656	1,599	1,558	1,527	1,503	1,483
110	2,162	1,873	1,732	1,647	1,589	1,548	1,517	1,492	1,472
120	2,156	1,866	1,724	1,639	1,581	1,539	1,508	1,483	1,463
130	2,151	1,861	1,718	1,632	1,574	1,532	1,500	1,475	1,455
140	2,146	1,855	1,712	1,626	1,568	1,526	1,494	1,469	1,448
150	2,142	1,851	1,708	1,621	1,563	1,520	1,488	1,463	1,442

Tabelle: $0{,}99-$ Quantil der $F-$Verteilung $df_2 = 1, \ldots, 10$

			df_2						
1	2	3	4	5	6	7	8	9	10
4052,2	98,50	34,12	21,20	16,26	13,745	12,246	11,259	10,561	10,044
4999,5	99,00	30,82	18,00	13,27	10,925	9,547	8,649	8,022	7,559
5403,4	99,17	29,46	16,69	12,06	9,780	8,451	7,591	6,992	6,552
5624,6	99,25	28,71	15,98	11,39	9,148	7,847	7,006	6,422	5,994
5763,7	99,30	28,24	15,52	10,97	8,746	7,460	6,632	6,057	5,636
5859,0	99,33	27,91	15,21	10,67	8,466	7,191	6,371	5,802	5,386
5928,4	99,36	27,67	14,98	10,46	8,260	6,993	6,178	5,613	5,200
5981,1	99,37	27,49	14,80	10,29	8,102	6,840	6,029	5,467	5,057
6022,5	99,39	27,35	14,66	10,16	7,976	6,719	5,911	5,351	4,942
6055,8	99,40	27,23	14,55	10,05	7,874	6,620	5,814	5,257	4,849
6083,3	99,41	27,13	14,45	9,96	7,790	6,538	5,734	5,178	4,772
6106,3	99,42	27,05	14,37	9,89	7,718	6,469	5,667	5,111	4,706
6125,9	99,42	26,98	14,31	9,83	7,657	6,410	5,609	5,055	4,650
6142,7	99,43	26,92	14,25	9,77	7,605	6,359	5,559	5,005	4,601
6157,3	99,43	26,87	14,20	9,72	7,559	6,314	5,515	4,962	4,558
6170,1	99,44	26,83	14,15	9,68	7,519	6,275	5,477	4,924	4,520
6181,4	99,44	26,79	14,12	9,64	7,483	6,240	5,442	4,890	4,487
6191,5	99,44	26,75	14,08	9,61	7,451	6,209	5,412	4,860	4,457
6200,6	99,45	26,72	14,05	9,58	7,422	6,181	5,384	4,833	4,430
6208,7	99,45	26,69	14,02	9,55	7,396	6,155	5,359	4,808	4,405
6260,6	99,47	26,51	13,84	9,38	7,229	5,992	5,198	4,649	4,247
6286,8	99,47	26,41	13,75	9,29	7,143	5,908	5,116	4,567	4,165
6302,5	99,48	26,35	13,69	9,24	7,091	5,858	5,065	4,517	4,115
6313,0	99,48	26,32	13,65	9,20	7,057	5,824	5,032	4,483	4,082
6320,6	99,49	26,29	13,63	9,18	7,032	5,799	5,007	4,459	4,058
6326,2	99,49	26,27	13,61	9,16	7,013	5,781	4,989	4,441	4,039
6330,6	99,49	26,25	13,59	9,14	6,998	5,766	4,975	4,426	4,025
6334,1	99,49	26,24	13,58	9,13	6,987	5,755	4,963	4,415	4,014
6337,0	99,49	26,23	13,57	9,12	6,977	5,745	4,954	4,406	4,004
6339,4	99,49	26,22	13,56	9,11	6,969	5,737	4,946	4,398	3,996
6341,4	99,49	26,21	13,55	9,11	6,962	5,731	4,939	4,391	3,990
6343,2	99,49	26,21	13,55	9,10	6,956	5,725	4,934	4,385	3,984
6344,7	99,49	26,20	13,54	9,09	6,951	5,720	4,929	4,380	3,979

Tabelle: $0{,}99-$ Quantil der $F-$Verteilung $df_2 = 2, \ldots, 100$

0,99				df_2					
dfl	20	30	40	50	60	70	80	90	100
1	8,096	7,562	7,314	7,171	7,077	7,011	6,963	6,925	6,895
2	5,849	5,390	5,179	5,057	4,977	4,922	4,881	4,849	4,824
3	4,938	4,510	4,313	4,199	4,126	4,074	4,036	4,007	3,984
4	4,431	4,018	3,828	3,720	3,649	3,600	3,563	3,535	3,513
5	4,103	3,699	3,514	3,408	3,339	3,291	3,255	3,228	3,206
6	3,871	3,473	3,291	3,186	3,119	3,071	3,036	3,009	2,988
7	3,699	3,304	3,124	3,020	2,953	2,906	2,871	2,845	2,823
8	3,564	3,173	2,993	2,890	2,823	2,777	2,742	2,715	2,694
9	3,457	3,067	2,888	2,785	2,718	2,672	2,637	2,611	2,590
10	3,368	2,979	2,801	2,698	2,632	2,585	2,551	2,524	2,503
11	3,294	2,906	2,727	2,625	2,559	2,512	2,478	2,451	2,430
12	3,231	2,843	2,665	2,562	2,496	2,450	2,415	2,389	2,368
13	3,177	2,789	2,611	2,508	2,442	2,395	2,361	2,334	2,313
14	3,130	2,742	2,563	2,461	2,394	2,348	2,313	2,286	2,265
15	3,088	2,700	2,522	2,419	2,352	2,306	2,271	2,244	2,223
16	3,051	2,663	2,484	2,382	2,315	2,268	2,233	2,206	2,185
17	3,018	2,630	2,451	2,348	2,281	2,234	2,199	2,172	2,151
18	2,989	2,600	2,421	2,318	2,251	2,204	2,169	2,142	2,120
19	2,962	2,573	2,394	2,290	2,223	2,176	2,141	2,114	2,092
20	2,938	2,549	2,369	2,265	2,198	2,150	2,115	2,088	2,067
30	2,778	2,386	2,203	2,098	2,028	1,980	1,944	1,916	1,893
40	2,695	2,299	2,114	2,007	1,936	1,886	1,849	1,820	1,797
50	2,643	2,245	2,058	1,949	1,877	1,826	1,788	1,759	1,735
60	2,608	2,208	2,019	1,909	1,836	1,785	1,746	1,716	1,692
70	2,582	2,181	1,991	1,880	1,806	1,754	1,714	1,684	1,659
80	2,563	2,160	1,969	1,857	1,783	1,730	1,690	1,659	1,634
90	2,548	2,144	1,952	1,839	1,764	1,711	1,671	1,639	1,614
100	2,535	2,131	1,938	1,825	1,749	1,695	1,655	1,623	1,598
110	2,525	2,120	1,927	1,813	1,737	1,683	1,642	1,610	1,584
120	2,517	2,111	1,917	1,803	1,726	1,672	1,630	1,598	1,572
130	2,510	2,103	1,909	1,794	1,717	1,662	1,621	1,588	1,562
140	2,503	2,096	1,902	1,786	1,710	1,654	1,613	1,580	1,554
150	2,498	2,091	1,896	1,780	1,703	1,647	1,605	1,572	1,546

Tabelle: $0{,}995-$ Quantil der $F-$Verteilung $df_2 = 1, \ldots, 10$

			äf₂						
1	2	3	4	5	6	7	8	9	10
16210,7	198,50	55,55	31,33	22,79	18,635	16,236	14,688	13,614	12,826
19999,5	199,00	49,80	26,28	18,31	14,544	12,404	11,042	10,107	9,427
21614,7	199,17	47,47	24,26	16,53	12,917	10,882	9,596	8,717	8,081
22499,6	199,25	46,20	23,16	15,56	12,028	10,050	8,805	7,956	7,343
23055,8	199,30	45,39	22,46	14,94	11,464	9,522	8,302	7,471	6,872
23437,1	199,33	44,84	21,98	14,51	11,073	9,155	7,952	7,134	6,545
23714,6	199,36	44,43	21,62	14,20	10,786	8,885	7,694	6,885	6,302
23925,4	199,38	44,13	21,35	13,96	10,566	8,678	7,496	6,693	6,116
24091,0	199,39	43,88	21,14	13,77	10,391	8,514	7,339	6,541	5,968
24224,5	199,40	43,69	20,97	13,62	10,250	8,380	7,211	6,417	5,847
24334,4	199,41	43,52	20,82	13,49	10,133	8,270	7,104	6,314	5,746
24426,4	199,42	43,39	20,71	13,38	10,034	8,176	7,015	6,227	5,661
24504,5	199,42	43,27	20,60	13,29	9,950	8,097	6,938	6,153	5,589
24571,8	199,43	43,17	20,52	13,22	9,877	8,028	6,872	6,089	5,526
24630,2	199,43	43,09	20,44	13,15	9,814	7,968	6,814	6,032	5,471
24681,5	199,44	43,01	20,37	13,09	9,758	7,915	6,763	5,983	5,422
24726,8	199,44	42,94	20,31	13,03	9,709	7,868	6,718	5,939	5,379
24767,2	199,44	42,88	20,26	12,99	9,664	7,826	6,678	5,899	5,340
24803,4	199,45	42,83	20,21	12,94	9,625	7,788	6,641	5,864	5,305
24836,0	199,45	42,78	20,17	12,90	9,589	7,754	6,608	5,832	5,274
25043,6	199,47	42,47	19,89	12,66	9,358	7,534	6,396	5,625	5,071
25148,2	199,48	42,31	19,75	12,53	9,241	7,422	6,288	5,519	4,966
25211,1	199,48	42,21	19,67	12,45	9,170	7,354	6,222	5,454	4,902
25253,1	199,48	42,15	19,61	12,40	9,122	7,309	6,177	5,410	4,859
25283,2	199,49	42,10	19,57	12,37	9,088	7,276	6,145	5,379	4,828
25305,8	199,49	42,07	19,54	12,34	9,062	7,251	6,121	5,356	4,805
25323,4	199,49	42,04	19,52	12,32	9,042	7,232	6,103	5,337	4,787
25337,5	199,49	42,02	19,50	12,30	9,026	7,217	6,088	5,322	4,772
25349,0	199,49	42,00	19,48	12,29	9,012	7,204	6,075	5,310	4,760
25358,6	199,49	41,99	19,47	12,27	9,001	7,193	6,065	5,300	4,750
25366,7	199,49	41,98	19,46	12,26	8,992	7,184	6,056	5,292	4,742
25373,7	199,49	41,97	19,45	12,26	8,984	7,177	6,049	5,284	4,734
25379,7	199,49	41,96	19,44	12,25	8,977	7,170	6,042	5,278	4,728

Tabelle: 0,995 − Quantil der F −Verteilung $df_2 = 20, \ldots, 100$

0,995				df_2					
äfi	20	30	40	50	60	70	80	90	100
1	9,944	9,180	8,828	8,626	8,495	8,403	8,335	8,282	8,241
2	6,986	6,355	6,066	5,902	5,795	5,720	5,665	5,623	5,589
3	5,818	5,239	4,976	4,826	4,729	4,661	4,611	4,573	4,542
4	5,174	4,623	4,374	4,232	4,140	4,076	4,029	3,992	3,963
5	4,762	4,228	3,986	3,849	3,760	3,698	3,652	3,617	3,589
6	4,472	3,949	3,713	3,579	3,492	3,431	3,387	3,352	3,325
7	4,257	3,742	3,509	3,376	3,291	3,232	3,188	3,154	3,127
8	4,090	3,580	3,350	3,219	3,134	3,076	3,032	2,999	2,972
9	3,956	3,450	3,222	3,092	3,008	2,950	2,907	2,873	2,847
10	3,847	3,344	3,117	2,988	2,904	2,846	2,803	2,770	2,744
11	3,756	3,255	3,028	2,900	2,817	2,759	2,716	2,683	2,657
12	3,678	3,179	2,953	2,825	2,742	2,684	2,641	2,608	2,583
13	3,611	3,113	2,888	2,760	2,677	2,619	2,577	2,544	2,518
14	3,553	3,056	2,831	2,703	2,620	2,563	2,520	2,487	2,461
15	3,502	3,006	2,781	2,653	2,570	2,513	2,470	2,437	2,411
16	3,457	2,961	2,737	2,609	2,526	2,468	2,425	2,393	2,367
17	3,416	2,921	2,697	2,569	2,486	2,428	2,385	2,353	2,326
18	3,380	2,885	2,661	2,533	2,450	2,392	2,349	2,316	2,290
19	3,347	2,853	2,628	2,500	2,417	2,359	2,316	2,283	2,257
20	3,318	2,823	2,598	2,470	2,387	2,329	2,286	2,253	2,227
30	3,123	2,628	2,401	2,272	2,187	2,128	2,084	2,051	2,024
40	3,022	2,524	2,296	2,164	2,079	2,019	1,974	1,939	1,912
50	2,959	2,459	2,230	2,097	2,010	1,949	1,903	1,868	1,840
60	2,916	2,415	2,184	2,050	1,962	1,900	1,854	1,818	1,790
70	2,885	2,383	2,150	2,015	1,927	1,864	1,817	1,781	1,752
80	2,861	2,358	2,125	1,989	1,900	1,837	1,789	1,752	1,723
90	2,843	2,339	2,105	1,968	1,878	1,815	1,767	1,730	1,700
100	2,828	2,323	2,088	1,951	1,861	1,797	1,748	1,711	1,681
110	2,816	2,311	2,075	1,937	1,846	1,782	1,733	1,695	1,665
120	2,806	2,300	2,064	1,925	1,834	1,769	1,720	1,682	1,652
130	2,797	2,291	2,054	1,915	1,824	1,758	1,709	1,671	1,640
140	2,790	2,283	2,046	1,907	1,815	1,749	1,700	1,661	1,630
150	2,783	2,276	2,038	1,899	1,807	1,741	1,691	1,652	1,621

Literatur

Bamberg G., BaurF., Krapp M.: Statistik. 16. Auflage, Oldenbourg Verlag 2011

Bleymüller J., Gehlert G., Gülicher H.: Statistik für Wirtschaftswissenschaftler. 14. Auflage, Vahlen 2004

Fahrmeir L., Künstler R., Pigeot L, Tutz G.: Statistik. Der Weg zur Datenanalyse. 7. Auflage, Springer Verlag 2010

Fahrmeir L.,Kneib T., LangS.: Regression. Modelle, Methodenund Anwendungen, 2. Auflage, Springer Verlag 2007

FerschlF: Deskriptive Statistik. 3. Auflage, Physika-Verlag 1985

FiszM.: Wahrscheinlichkeitsrechnung und mathematische Statistik. 11. Auflage, Deutscher Verlag der Wissenschaften 1989

Galata, Scheid, Deskriptive und induktive Statistik für Studierenden der BWL, Methoden-Beispiele-Anwndungen, Fachbuchverlag, Leipzig,

Hartung J., ElpeltB., KläsenerK-H: Statistk. Lehr- und Handbuch der angewandten Statistik. 14. Auflage, Oldenbourg Verlag 2009

Johnson N. L., Kotz S.: Discrete distributions. John Wiley 1969

Johnson N. L., Kotz S.: Continuous univariate distributions, Band 1 und 2. John Wiley 1970 *Pflaumer P, Heine B., Hartung J.:* Statistik für Wirtschafts- und Sozialwissenschaften: Deskriptive Statistik. 4. Auflage, Oldenbourg Verlag 2010

Röpcke H., Wessler M.: Wirtschaftsmathematik. Methoden - Beispiele - Anwendungen. Fachbuchverlag Leipzig im Carl Hanser Verlag 2012

Rüger B.: Induktive Statistik. Einführung für Wirtschafts- und Sozialwissenschaftler. 3. Auflage, Oldenbourg Verlag 1995

Rüger B.: Test- und Schätztheorie. Band I: Grundlagen. Oldenbourg Verlag 1999

Rüger B.: Test- und Schätztheorie. Band II: Statistische Tests. Oldenbourg Verlag 2002

Sachs L.: Angewandte Statistik. Anwendung statistischer Methoden. 11. Auflage, Springer Verlag 2002

Schira J.: Statistische Methoden der VWL und BWL. Theorie und Praxis. 3. Auflage, Pearson Studium 2009

SchneeweißH.: Ökonometrie. 4. Auflage, Physika-Verlag 1990

Toutenburg H., Heumann C., *Schomaker M.:* Deskriptive Statistik. Einführung in Methoden und Anwendungen mit R und SPSS. 5. Auflage, Springer Verlag 2009

Toutenburg H., Heumann C., Schomaker M., Wißmann M.: Induktive Statistik. Einführung mit R und SPSS. 4. Auflage, Springer Verlag 2008

Voß W.: Taschenbuch der Statistik. 2. Auflage, Hanser Verlag 2003

Weigand C.: Statistik mit und ohne Zufall. Eine anwendungsorientierte Einführung. 2. Auflage, Physika Verlag