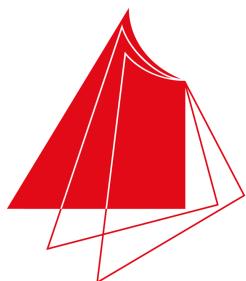


Design For Six Sigma

- Skript Statistik -

Manfred Strohrmann
Stefan Günter



Hochschule Karlsruhe
Technik und Wirtschaft
UNIVERSITY OF APPLIED SCIENCES

Änderungsindex

Datum	Verfasser	Änderungen
09.07.2020	M. Strohrmann	Überarbeitung, Ausgabe für Vorlesung WS 2020/21 und für Design For Six Sigma Online
15.03.2018	M. Strohrmann	Überarbeitung, Synchronisation mit Online-Portal DFSS Online, Ausgabe für Vorlesung SS 2018
15.03.2015	M. Strohrmann	Überarbeitung, Ausgabe für Vorlesung SS 2015
15.03.2014	M. Strohrmann	Überarbeitung, Ausgabe für Vorlesung SS 2014
15.03.2012	M. Strohrmann	Überarbeitung, Ausgabe für Vorlesung SS 2012
15.03.2011	M. Strohrmann	Überarbeitung, Ausgabe für Vorlesung SS 2011
15.03.2010	M. Strohrmann	Überarbeitung, Ausgabe für Vorlesung SS 2010
15.03.2019	M. Strohrmann	Überarbeitung, Ausgabe für Vorlesung SS 2019

Inhaltsverzeichnis

1 Einleitung	1
2 Grundlagen der Wahrscheinlichkeitstheorie	2
2.1 Grundbegriffe und Mengenoperationen	2
2.2 Klassische Wahrscheinlichkeit nach Laplace	6
2.3 Aufbau von Ereignisbäumen	13
2.4 Wahrscheinlichkeitsbegriff der Statistik nach Kolmogoroff	15
2.5 Anwendungsbeispiel: Sensordiagnose im Steuergerät	24
2.6 Literatur	27
3 Beschreibende Statistik univariater Daten	28
3.1 Merkmalstypen	28
3.2 Häufigkeitsverteilungen	30
3.3 Kennwerte einer Stichprobe	38
3.4 Anwendungsbeispiel: Charakterisierung eines Klebeprozesses	56
3.5 Literatur	65
4 Univariate Wahrscheinlichkeitstheorie	66
4.1 Zufallsvariablen und Wahrscheinlichkeitsverteilungen	66
4.2 Erwartungswerte von Verteilungen	73
4.3 Kennwerte von Verteilungen	77
4.4 Funktionen von Zufallsvariablen	85
4.5 Spezielle diskrete Verteilungen	91
4.6 Spezielle stetige Verteilungen	112
4.7 Prüf- oder Testverteilungen	137
4.8 Literatur	145
5 Schätzung von unbekannten Parametern einer Verteilung	146
5.1 Zielsetzung und Problematik der Parameterschätzung	146
5.2 Erwartungstreue der Parameterschätzung	149
5.3 Konfidenzbereiche für die Schätzung von Parametern	152
5.4 Konfidenzbereiche für den Vergleich von Stichproben	162
5.5 Vorhersageintervalle für künftige Stichprobenwerte	172
5.6 Anwendungsbeispiel: Kontaktwiderstand eines Batteriesensors	181
5.7 Literatur	187
6 Motivation mit einem einführenden Beispiel	188
6.1 Motivation mit einem einführenden Beispiel	188
6.2 Praktisches Durchführen von Hypothesentests	191
6.3 Hypothesentest und Konfidenzbereich	202
6.4 Sicherheit bei Hypothesentests	204
6.5 Hypothesentests für die Parameter einer Normalverteilung	212
6.6 Hypothesentests für den Vergleich zweier Normalverteilungen	219
6.7 Anwendungsbeispiel: Diagnose von Feuchtesensoren	231
6.8 Literatur	235
7 Beschreibende Statistik multivariater Daten	236
7.1 Darstellung und Charakterisierung von Datensätzen	236
7.2 Kenngrößen multivariater Stichproben	247
7.3 Anwendungsbeispiel: Schwindung beim Spritzgießen	252
7.4 Literatur	256

8 Multivariate Wahrscheinlichkeitstheorie	257
8.1 Gemeinsame Verteilungs- und Dichtefunktionen	257
8.2 Kenngrößen multivariater Wahrscheinlichkeitsverteilungen	263
8.3 Unabhängige Zufallsvariablen	264
8.4 Funktionen von Zufallsvariablen	267
8.5 Zentraler Grenzwertsatz	271
8.6 Spezielle multivariate Verteilungen	273
8.7 Literatur	283
9 Varianzanalyse	284
9.1 Einfaktorielle Varianzanalysen	285
9.2 Mehrfaktorielle Varianzanalyse	292
9.3 Anwendungsbeispiel: Homogenitätsprüfung eines Luftflusses	301
9.4 Literatur	304
10 Korrelationsanalyse	305
10.1 Korrelationskoeffizient einer Stichprobe	305
10.2 Definition des Korrelationskoeffizienten ρ der Grundgesamtheit	310
10.3 Bewertung des Korrelationskoeffizienten	315
10.4 Bewertung des Korrelationskoeffizienten mehrdimensionaler Stichproben	323
10.5 Korrelation und Kausalzusammenhang	328
10.6 Literatur	329
11 Kapitel11	330
12 Regression zweidimensionaler Datensätze	331
12.1 Lineare Regression	332
12.2 Regression mit Polynomen	360
12.3 Berechnung und Bewertung der Regressionskoeffizienten	360
12.4 Bewertung von Regressionen zweidimensionaler Datensätze	366
12.5 Statistische Bewertung des Bestimmtheitsmaßes	369
12.6 Sonderformen der zweidimensionalen Regression	374
12.7 Literatur	378
13 Regression mehrdimensionaler Datensätze	379
13.1 Bestimmung der Regressionsfunktion	379
13.2 Statistische Bewertung der Regressionsparameter	389
13.3 Korrektheit der Regression und schlecht gestellte Probleme	404
13.4 Korrektheit und Kondition einer Matrix	413
13.5 Regularisierung von Matrizen (21.06.2015)	413
14 Transformation von Zufallsvariablen	414
15 Anhang 2: Grundlagen der linearen Algebra	415
15.1 Matrizenalgebra	415

1 Einleitung

Im Rahmen von Design For Six Sigma werden statistische Methoden eingesetzt. In der Optimize-Phase werden Methoden der Statistik verwendet, um Prozesssicherheiten zu bewerten und Streuungen von Bauelementen und Prozessen zu einer Gesamttoleranz zu überlagern. Die statistischen Methoden zur Optimierung des Produktes in der Entwicklungsphase sind Statistische Simulation und Robust Design sowie statistische Versuchsplanung und statistische Tolerierung. In der Verify-Phase wird die prognostizierten Fertigbarkeit und Zuverlässigkeit bestätigt. Dazu ist neben einem durchdachten Erprobungsplan ein Nachweis der Messfähigkeit von Messeinrichtungen in Labor und Fertigung notwendig. Auf Basis geeigneter Messeinrichtungen erfolgt für die Fertigungsprozesse eine statistische Prozesskontrolle (SPC). Das erforderliche Grundwissen im Bereich der Wahrscheinlichkeitstheorie und Statistik wird in Dokument vermittelt.

Zu Beginn werden die Grundlagen der Wahrscheinlichkeitstheorie behandelt, um einen ersten Einstieg in die Statistik zu ermöglichen. Hierbei werden Mengenoperationen zur Beschreibung statistische Begebenheiten eingeführt und der Begriff der Wahrscheinlichkeit nach Laplace und Kolmogoroff erläutert.

Nach den Grundlagen der Wahrscheinlichkeitstheorie werden dem Leser die Grundlagen vorgestellt, die zur Wahrscheinlichkeitsrechnung mit einer Variablen erforderlich sind. Diese Aufgaben werden als univariate Aufgaben bezeichnet. Ausgehend von der beschreibenden Statistik wird über die univariate Wahrscheinlichkeitstheorie gezeigt, wie univariate Stichproben beurteilt werden können. Der Leser wird dabei in die Lage versetzt, von einer vorliegenden Stichprobe auf die Grundgesamtheit zu schließen und mittels Konfidenzintervallen deren Genauigkeit statistisch zu beschreiben. Außerdem werden die Prinzipien von Hypothesentests und deren Anwendung erläutert.

Im nächsten Schritt wird das Wissen auf multivariate Statistik erweitert. Es wird erklärt, wie mehrdimensionale oder multivariate Stichproben sowohl grafisch als auch mit Kenngrößen dargestellt werden können. Mit der Varianzanalyse wird das Streuverhalten eines Systems in Abhängigkeit von Merkmalen beschrieben. Ausgehend von der einfaktoriellen Varianzanalyse wird die mehrfaktorielle Varianzanalyse eingeführt und an Beispielen angewandt. Darauf aufbauend wird die Korrelationsanalyse eingeführt. Abschließend wird in dem Kapitel Regressionsanalyse gezeigt, wie multivariate Stichproben mathematisch durch Funktionen approximiert werden können.

In die Darstellung sind viele Hinweise von Kollegen und Studierenden eingeflossen, für die ich mich an dieser Stelle herzlich bedanken möchte. Für weitere Hinweise bin ich jederzeit dankbar.

Karlsruhe, 09.07.2020

2 Grundlagen der Wahrscheinlichkeitstheorie

Die Wahrscheinlichkeitstheorie ist ein mathematisches Modell zur Beschreibung empirischer Sachverhalte, bei denen der Zufall eine Rolle spielt. Sie entstand aus dem Wunsch, die Gewinnaussichten bei Glücksspielen zu prognostizieren. Aber schon bald zeigte sich, dass sie in sehr vielen anderen Gebieten angewendet werden kann, zum Beispiel im Versicherungswesen oder in der Mess- und Prozesstechnik. Der Versuch, den Begriff der Wahrscheinlichkeit zu definieren, führt zunächst zum klassischen Wahrscheinlichkeitsbegriff von Laplace. Diese Definition ist allerdings für Anwendungen in der Mess- und Prozesstechnik speziell und wird deshalb verallgemeinert.

2.1 Grundbegriffe und Mengenoperationen

Bevor die Grundlagen der Wahrscheinlichkeitstheorie zusammengestellt und diskutiert werden können, müssen einige Grundbegriffe erläutert und die Grundzüge der Mengenlehre eingeführt werden. Zur Veranschaulichung der Begriffe wird parallel ein Würfelexperiment beschrieben.

2.1.1 Ereignisse

Die Wahrscheinlichkeitstheorie hat den Anspruch, die Wahrscheinlichkeit für den Ausgang von zufälligen Prozessen vorauszusagen. Der Zufallsprozess wird auch als Zufallsexperiment bezeichnet. Er muss wiederholbar und das Ergebnis vom Zufall abhängig sein. Das konkrete Ergebnis kann im Voraus nicht eindeutig bestimmt werden.

Bekanntestes Beispiel für ein Zufallsexperiment ist das Werfen eines regelmäßigen Würfels. Der Würfel kann beliebig oft geworfen werden und das Ergebnis ist nicht vorhersagbar. Bei jedem Zufallsexperiment sind unterschiedliche Ergebnisse möglich, die zufällig eintreffen. Diese Ergebnisse werden Zufallereignisse oder Ereignisse genannt. Beim einmaligen Werfen eines Würfels können die Ereignisse

$$1, 2, 3, 4, 5 \text{ oder } 6 \quad (2.1)$$

eintreffen oder realisiert werden. Im Folgenden werden unterschiedliche Ereignisse definiert, die sich aus einem Zufallsexperiment ergeben können. Als Beispiel dient dabei das einmalige Würfeln mit einem Würfel.

Elementarereignis

Sind Ergebnisse von Zufallsexperimenten nicht weiter zerlegbar und schließen sich gegenseitig aus, werden sie als Elementarereignisse bezeichnet. Beim einmaligen Würfeln sind die Ereignisse

$$1, 2, 3, 4, 5 \text{ oder } 6 \quad (2.2)$$

Elementarereignisse. Sie bestehen aus einem Element und sind deshalb nicht weiter zerlegbar, außerdem schließen sie sich gegenseitig aus.

Ereignisraum

Die Menge Ω aller Elementarereignisse eines Zufallsexperimentes wird als Ereignisraum dieses Zufallsexperimentes bezeichnet. In der Literatur wird der Ereignisraum auch Grundraum genannt. Für das einmalige Würfeln ergibt sich ein Ereignisraum Ω von

$$\Omega = \{1, 2, 3, 4, 5, 6\} \quad (2.3)$$

Dabei werden, wie in der Mengenlehre üblich, Mengen mit geschweiften Klammern dargestellt.

Ereignismenge

Eine Menge möglicher Elementarereignisse wird als Ereignismenge A bezeichnet. Sie ist eine Teilmenge des Ereignisraums Ω . Zum Beispiel ist beim einmaligen Würfeln die Menge A der geraden Zahlen

$$A = \{2, 4, 6\} \quad (2.4)$$

eine Teilmenge des Ereignisraums Ω und damit eine Ereignismenge.

Unmögliches Ereignis

Ist ein Ereignis kein Teilelement des Ereignisraums, wird es als unmögliches Ereignis U bezeichnet.

$$A \not\subset \Omega \quad (2.5)$$

Beim einmaligen Würfeln kann die Zahl 7 nicht eintreffen, das Ereignis ist unmöglich.

Sicheres Ereignis

Entspricht die Definition eines Ereignisses dem gesamten Ereignisraum Ω ,

$$A = \Omega \quad (2.6)$$

wird das Ereignis als sicheres Ereignis S bezeichnet. Bei dem Würfeln mit einem Würfel wird eines der Ergebnisse 1, 2, 3, 4, 5 oder 6 sicher eintreffen.

2.1.2 Verknüpfungen von Ereignissen durch Mengenoperationen

Eine Menge kann, wie im vorhergehenden Abschnitt gezeigt wird, als eine Zusammenfassung verschiedener Ereignisse verstanden werden. Zufallsereignisse lassen sich daher mithilfe der Mengenlehre beschreiben und verknüpfen. Der Mengenbegriff wird anhand des Zufallsexperimentes Würfeln mit einem regelmäßigen Würfel verdeutlicht. Das Würfeln führt zu sechs möglichen Ereignissen. Diese Möglichkeiten bilden den Ereignisraum ?, der als Menge dargestellt werden kann.

$$\Omega = \{1, 2, 3, 4, 5, 6\} \quad (2.7)$$

Für das Experiment werden die Mengen A - D definiert:

- A Würfeln einer geraden Zahl, $A = \{2, 4, 6\}$
- B Würfeln einer durch 3 teilbaren Zahl, $B = \{3, 6\}$
- C Würfeln einer 1, $C = \{1\}$
- D Würfeln einer 4, $D = \{4\}$

Die Ereignisse sind in Bild 2.1 grafisch dargestellt:

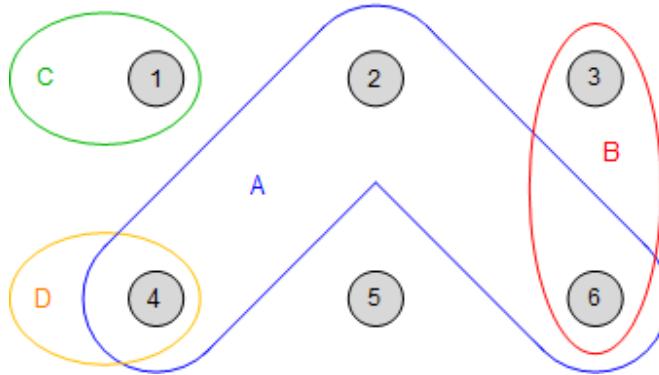


Bild 2.1: Darstellung des Zufallsexperimentes Wurf eines regelmäßigen Würfels

Mit dem Beispiel Wurf eines regelmäßigen Würfels werden im Folgenden die grundlegenden Mengenoperationen beschrieben.

Element der Menge

Ist eine Menge D in einer Menge A vollständig enthalten, wird sie als Element der Menge bezeichnet. Die Eigenschaft wird mit der Schreibweise

$$D \in A \quad (2.8)$$

dargestellt. Ist die Menge C kein Element der Menge A, ergibt sich die Schreibweise

$$C \notin A \quad (2.9)$$

Teilmenge

Ist eine Menge D komplett in einer anderen Menge A enthalten, ist die Menge D eine Teilmenge von der Menge A. Dafür wird die Schreibweise

$$D \subseteq A \quad (2.10)$$

verwendet.

Vereinigungsmenge

Mit $A \cup B$ wird das Ereignis bezeichnet, bei dem das Ereignis A oder das Ereignis B eintrifft. In der Mengenlehre wird von der Vereinigungsmenge der Ereignisse A und B gesprochen. In dem Beispiel aus Bild 2.1 umfasst die Vereinigungsmenge $A \cup B$ die Elemente

$$A \cup B = \{2, 3, 4, 6\} \quad (2.11)$$

Die Vereinigungsmenge $A \cup B$ der Ereignisse A und B sind also Würfe mit den Augenzahlen 2, 3, 4 oder 6.

Schnittmenge

Mit $A \cap B$ wird das Ereignis bezeichnet, bei dem das Ereignis A und das Ereignis B zusammen eintreffen. In der Mengenlehre wird von der Schnittmenge der Ereignisse A und B gesprochen. In dem Beispiel aus Bild 2.1 umfasst die Schnittmenge $A \cap B$ das Element

$$A \cap B = \{6\} \quad (2.12)$$

Die Schnittmenge $A \cap B$ der Ereignisse A und B ist ein Wurf mit einer Augenzahl 6. Diese Augenzahl erfüllt sowohl die Forderung nach einer geraden Zahl als auch die Forderung, durch 3 teilbar zu sein.

Differenzmenge

Die Differenzmenge $A \setminus B$ ist die Menge aller Elemente, die in A, aber nicht in B vorkommen. Für das Beispiel aus Bild 2.1 ergibt sich die Differenzmenge $A \setminus B$ zu

$$A \setminus B = \{2, 4\} \quad (2.13)$$

Komplementäre oder inverse Menge

Die komplementäre oder inverse Menge A' bezeichnet die Ereignisse, die im Ereignisraum liegen, aber kein Element der Menge A sind.

$$A' = \Omega \setminus A \quad (2.14)$$

In dem Beispiel aus Bild 2.1 ergibt sich die komplementäre Menge A' zu

$$A' = \Omega \setminus A = \{1, 3, 5\} \quad (2.15)$$

Disjunkte Menge

Wenn zwei Ereignisse nicht gemeinsam eintreffen können, schließen sich die Ereignisse gegenseitig aus. Ihre Schnittmenge ist eine leere Menge.

$$A \cap C = \{\} \quad (2.16)$$

Die Mengen werden als disjunkte Mengen bezeichnet. In dem Beispiel aus Bild 2.1 schließen sich die Ereignisse A und C gegenseitig aus, weil die Zahl 1 keine gerade Zahl ist.

Rechenregeln für Mengen

Mithilfe von Mengenoperationen lassen sich Rechenregeln für die mit den Ereignissen verbundenen Wahrscheinlichkeiten ableiten. Die Rechenregeln sind in Tabelle 2.1 zusammengestellt. Ihre Gültigkeit kann anhand des Beispiels des einmaligen Würfels plausibilisiert werden.

Tabelle 2.1: Funktionen zur Beschreibung von Einschwingvorgängen

Gesetz	Rechenoperation
Kommutativgesetz	$A \cap B = B \cap A, A \cup B = B \cup A$
Assoziativgesetz	$(A \cap B) \cap C = A \cap (B \cap C)$ $(A \cup B) \cup C = A \cup (B \cup C)$
Distributivgesetz	$(A \cup B) \cap C = (A \cap C) \cup (B \cap C)$ $(A \cap B) \cup C = (A \cup C) \cap (B \cup C)$
Die Morgansche Regeln	$(A \cup B)' = A' \cap B', (A \cap B)' = A' \cup B'$

2.2 Klassische Wahrscheinlichkeit nach Laplace

2.2.1 Definition

Zur Herleitung des Begriffes der klassischen Wahrscheinlichkeit nach Laplace wird wieder das Beispiel des Würfeln mit einem regelmäßigen Würfel betrachtet. Alle sechs Elementarereignisse schließen einander aus. Da der Würfel regelmäßig ist, ist keines dieser sechs Elementarereignisse wahrscheinlicher als ein anderes. Alle Elementarereignisse sind möglich und gleich wahrscheinlich.

In ähnlicher Weise können auch für andere Zufallsexperimente die jeweiligen Elementarereignisse angegeben werden, die einander ausschließen und gleich wahrscheinlich sind. Diese beiden Eigenschaften sind die Voraussetzung für die klassische Wahrscheinlichkeitsrechnung nach Laplace.

Gibt es bei einem Experiment N gleichwahrscheinliche Fälle, können diese Fälle in zwei Gruppen eingeteilt werden. Die günstigen Fälle erfüllen eine definierte Bedingung, die ungünstigen Fälle erfüllen die definierte Bedingung nicht. Zum Beispiel kann die Bedingung für ein günstiges Ereignis A lauten, dass bei einem Wurf mit einem regelmäßigen Würfel eine gerade Zahl gewürfelt wird. Dann ist der Wurf mit der Augenzahl 2, 4 oder 6 ein günstiger Fall, alle anderen Augenzahlen sind ungünstige Fälle.

Wird die Anzahl der günstigen Elementarereignisse mit G bezeichnet und die Anzahl der möglichen Fälle mit N, ergibt sich die Wahrscheinlichkeit P(A) für das günstige Ereignis A nach Laplace zu

$$P(A) = \frac{G}{N} \quad (2.17)$$

Das Ereignis A, bei einem Wurf mit einem regelmäßigen Würfel eine gerade Zahl zu würfeln, ergibt sich aus 3 günstigen Elementarereignissen, nämlich den Augenzahlen 2, 4 und 6. Entsprechend ergibt sich für die Wahrscheinlichkeit P(A)

$$P(A) = \frac{3}{6} = \frac{1}{2} \quad (2.18)$$

Für das Ereignis D, bei einem Wurf mit einem regelmäßigen Würfel eine Augenzahl 4 zu erzielen, ergibt sich die Wahrscheinlichkeit entsprechend aus

$$P(D) = \frac{1}{6} \quad (2.19)$$

Beispiel: Zweimaliges Würfeln

In dem folgenden Beispiel wird mit zwei Würfeln gewürfelt. Wie groß ist die Wahrscheinlichkeit, bei einem einzelnen Wurf mit zwei regelmäßigen Würfeln gleichzeitig zwei gerade Zahlen zu würfeln?

Es gibt insgesamt 36 gleichmögliche Fälle. Die folgenden Kombinationen von Einzelwürfen stellen den Ereignisraum Ω dar.

$$\Omega = \{(1, 1), (1, 2), \dots, (1, 6), (2, 1), (2, 2), \dots, (2, 6), \dots, (6, 1), (6, 2), (6, 3), (6, 4), (6, 5), (6, 6)\} \quad (2.20)$$

Dabei bezeichnet die erste Zahl jeweils die mit dem ersten Würfel erzielte Augenzahl und die zweite Zahl die mit dem zweiten Würfel erzielte Augenzahl. Bei 9 dieser 36 Fälle

$$A = \{(2, 2), (2, 4), (2, 6), (4, 2), (4, 4), (4, 6), (6, 2), (6, 4), (6, 6)\} \quad (2.21)$$

trifft das Ereignis zweier gerader Zahlen ein. Damit ergibt sich die Wahrscheinlichkeit

$$P(A) = \frac{9}{36} = \frac{1}{4} \quad (2.22)$$

Bei der Beschreibung von Zufallsexperimenten mit einer endlichen Anzahl von Ereignissen ist es erforderlich, die Anzahl möglicher und günstiger Varianten zu berechnen. Diese Rechnungen bauen auf wenigen Rechenmodellen auf: Permutationen, Variationen und Kombinationen. Sie werden in den folgenden Abschnitten vorgestellt.

2.2.2 Permutationen

Jede Zusammenstellung von N Elementen, die dadurch entsteht, dass sämtliche Elemente unter Berücksichtigung der Reihenfolge zusammengesetzt werden, heißt Permutation der gegebenen Elemente. Bei der Auswahl des ersten Elementes existieren N Möglichkeiten, den ersten Platz zu besetzen. Anschließend sind noch (N – 1) Elemente übrig, die als zweites Element gewählt werden können. Die Anzahl M der Permutationen, die mit N verschiedenen Elementen generiert werden können, ergibt sich damit aus

$$M = N! = N \cdot (N - 1) \cdot (N - 2) \cdot (N - 3) \cdot \dots \cdot 1 \quad (2.23)$$

Beispiel: Permutationen bei verschiedenfarbigen Kugeln

Als Beispiel wird berechnet, wie viele Permutationen bei einer Gruppe von einer roten, einer blauen und einer gelben Kugel entstehen können.

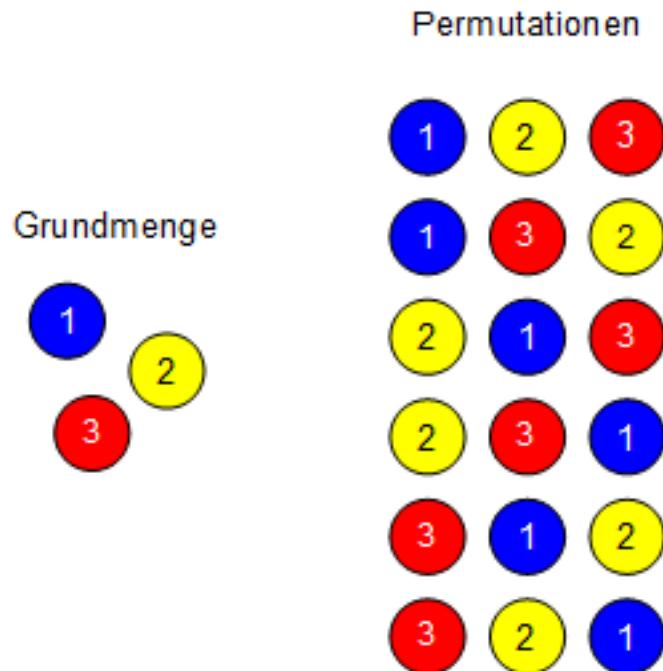


Bild 2.2: Permutationen von drei verschiedenfarbigen Kugeln

Für das Beispiel ergibt sich mit N = 3

$$M = N! = 3 \cdot 2 \cdot 1 = 6 \quad (2.24)$$

Das Ergebnis deckt sich mit dem grafischen Ergebnis in Bild 2.2.

Sind die N Elemente nicht alle verschieden, sondern lassen sie sich in K Klassen gleicher Elemente mit der jeweiligen Anzahl N_1, N_2, \dots, N_K einteilen, berechnet sich die Anzahl M unterscheidbarer Permutationen zu

$$M = \frac{N!}{\prod_{k=1}^K N_k!} = \frac{N!}{N_1! \cdot \dots \cdot N_k! \cdot \dots \cdot N_K!} \quad (2.25)$$

Die Division durch die Faktoren $N_k!$ ergibt sich daraus, dass zwischen den Permutationen identischer Elemente innerhalb einer Klasse nicht unterschieden werden kann.

Beispiel: Permutationen bei zwei gelben und einer roten Kugel

Als Beispiel wird berechnet, wie viele Permutationen bei einer Gruppe von zwei gelben und einer roten Kugel entstehen können. Als Grundmenge stehen die Kugeln 1 - 3 zur Verfügung. Die Kugel 1 ist rot, die Kugeln 2 und 3 sind gelb.

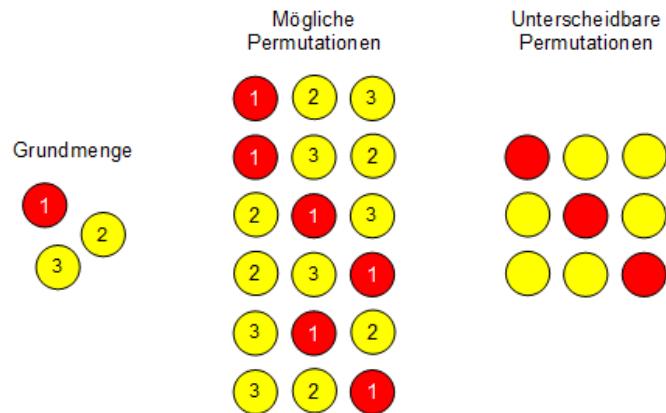


Bild 2.3: Permutationen von zwei blauen und einer roten Kugel

Es liegen $N = 3$ Kugeln vor, die in $K = 2$ Klassen aufgeteilt sind. In der Klasse 1 befinden sich $N_1 = 2$ gelbe Kugeln, in der Klasse 2 befindet sich $N_2 = 1$ rote Kugel. Damit berechnet sich die Anzahl M unterscheidbarer Permutationen aus

$$M = \frac{N!}{N_k! \cdot N_{k-1}! \cdot \dots \cdot N_1!} = \frac{3!}{2! \cdot 1!} = 3 \quad (2.26)$$

Die Anzahl stimmt mit dem grafischen Ergebnis aus Bild 2.3 überein.

Bei Permutationen werden alle Elemente einer Grundmenge angeordnet. Im Gegensatz dazu werden bei Variationen und Kombinationen nur Teile einer Grundmenge angeordnet.

2.2.3 Variationen

Wird aus einer Menge von N Elementen eine Teilmenge von K Elementen herausgegriffen und ist die Reihenfolge dieser Teilmenge von Bedeutung, so wird diese Zusammenstellung als Variation K-ter Ordnung bezeichnet. Es werden Variationen mit und ohne Wiederholung unterschieden.

Bei Variationen ohne Wiederholung stehen bei der ersten Ziehung noch alle N Elemente zur Verfügung. Bei der zweiten Ziehung ist ein Element bereits gezogen worden, es stehen nur noch (N - 1) Elemente zur Verfügung. Für N gegebene, voneinander verschiedene Elemente ergibt sich ohne Wiederholung die Anzahl M der Variationen K-ter Ordnung aus

$$M = N \cdot (N - 1) \cdot (N - 2) \cdot \dots \cdot (N - K + 1) = \frac{N!}{(N - K)!} \quad (2.27)$$

Werden N unbeschränkt oft wiederholbare verschiedene Elemente in Variationen K-ter Ordnung angeordnet, so existieren für jeden der K Plätze N Möglichkeiten. Für Variationen mit Wiederholung ergibt sich damit eine Anzahl von M Möglichkeiten mit

$$M = N^K \quad (2.28)$$

Beispiel: Variationen bei verschiedenenfarbigen Kugeln ohne und mit Wiederholung

Als Beispiel wird berechnet, wie viele Variationen sich bei einer Gruppe von einer roten, einer blauen und einer gelben Kugel entstehen können, von denen zwei Kugeln gezogen werden.



Bild 2.4: Variationen von zwei Kugeln aus einer Grundmenge von drei unterschiedlichen Kugeln ohne und mit Wiederholung

Die Berechnungen ergeben mit $N = 3$ und $K = 2$ die Anzahl der unterscheidbaren Variationen ohne Wiederholung zu

$$M = \frac{3!}{(3 - 2)!} = 6 \quad (2.29)$$

und mit Wiederholungen zu

$$M = 3^2 = 9 \quad (2.30)$$

Die Ergebnisse decken sich mit der grafischen Lösung in Bild 2.4

2.2.4 Kombinationen

Bei einer Kombination von Elementen bleibt die Reihenfolge unberücksichtigt. Wie bei Variationen werden Anwendungen mit und ohne Wiederholung unterschieden. Für Zufallsexperimente ohne Wiederholung wird die Anzahl von möglichen Variationen in Gleichung (2.27) berechnet. Dabei ergibt sich für jede Kombination eine Anzahl von $K!$ Permutationen, die sich nur in der Reihenfolge der Anordnung unterscheiden. Ist die Anordnung der Elemente nicht von Bedeutung, so fallen von den Kombinationen diejenigen zusammen, die die gleichen Elemente in der Anordnung haben. Gemäß Gleichung (2.23) sind das bei einer Kombination K -ter Ordnung $K!$ Elemente. Die Anzahl M von Kombinationen K -ter Ordnung aus einer Grundmenge mit N verschiedenen Elementen ohne Wiederholung beträgt damit

$$M = \binom{N}{K} = \frac{N \cdot (N-1) \cdot (N-2) \cdot \dots \cdot (N-K+1)}{K!} = \frac{N!}{K! \cdot (N-K)!} \quad (2.31)$$

Der Rechenausdruck in Gleichung (2.31) wird als N über K oder als Binomialkoeffizient bezeichnet. Die Anzahl M von Kombinationen K -ter Ordnung aus einer Grundmenge mit N verschiedenen Elementen mit Wiederholung kann auf eine ähnliche Art dargestellt werden. Es ergibt sich

$$M = \binom{N+K-1}{K} = \frac{(N+K-1)!}{K! \cdot (N-1)!} \quad (2.32)$$

Beispiel: Kombinationen bei verschiedenfarbigen Kugeln ohne und mit Wiederholung

Als Beispiel wird berechnet, wie viele Kombinationen sich bei einer Gruppe von einer roten, einer blauen und einer gelben Kugel ergeben können, von denen zwei Kugeln gezogen werden. Mit $N = 3$ und $K = 2$ ergibt sich die Anzahl der unterscheidbaren Kombinationen ohne Wiederholung zu

$$M = \frac{3!}{2! \cdot (3-2)!} = 3 \quad (2.33)$$

und mit Wiederholungen ergibt sich

$$M = \binom{3+2-1}{2} = \frac{4!}{2! \cdot 2!} = 6 \quad (2.34)$$

Die Ergebnisse decken sich mit der grafischen Lösung in Bild 2.5.

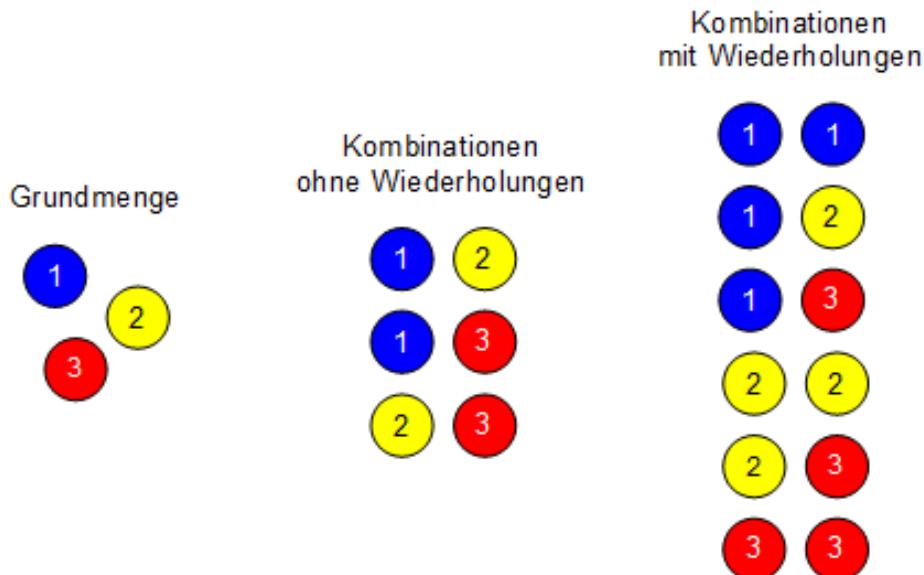


Bild 2.5: Kombinationen von zwei Kugeln aus einer Grundmenge von drei unterschiedlichen Kugeln ohne und mit Wiederholung

Beispiel: Zahlenlotto

Bekanntestes Beispiel für die Berechnung von Kombinationsmöglichkeiten ist das Zahlenlotto, bei dem 6 Kugeln aus einer Menge aus 49 Kugeln gezogen werden. Die Kugeln können nicht mehrfach gezogen werden, die Reihenfolge der Ziehung ist nicht relevant. Es handelt sich um eine Kombination ohne Wiederholung und ohne Berücksichtigung der Reihenfolge. Die Anzahl von Möglichkeiten errechnet sich nach Gleichung (2.35) zu

$$M = \binom{49}{6} = \frac{49!}{6! \cdot 43!} \approx 14\text{Mio.} \quad (2.35)$$

2.2.5 Überblick über Permutationen, Variationen und Kombinationen

Bei der Berechnung von Permutationen, Variationen und Kombinationen müssen folgende Fragen beantwortet werden:

- Welchen Umfang N hat die Ausgangsmenge?
- Wie kann die Auswahlprozedur modelliert werden (Reihenfolge, Wiederholungen, ...)?
- Welchen Umfang K hat die Auswahl?

Ist die Reihenfolge bei der Auswahl nicht relevant, handelt es sich bei dem Auswahlprozess um eine Kombination. Ist die Reihenfolge des Auswahlprozesses relevant und wird nur ein Teil aller Elemente der Grundmenge angeordnet, handelt es sich um eine Variation von Elementen. Werden alle Elemente angeordnet und ist die Reihenfolge der Auswahl wesentlich, handelt es sich um eine Permutation von Elementen. Mit diesen Informationen kann der geeignete Berechnungsprozess ausgewählt werden.

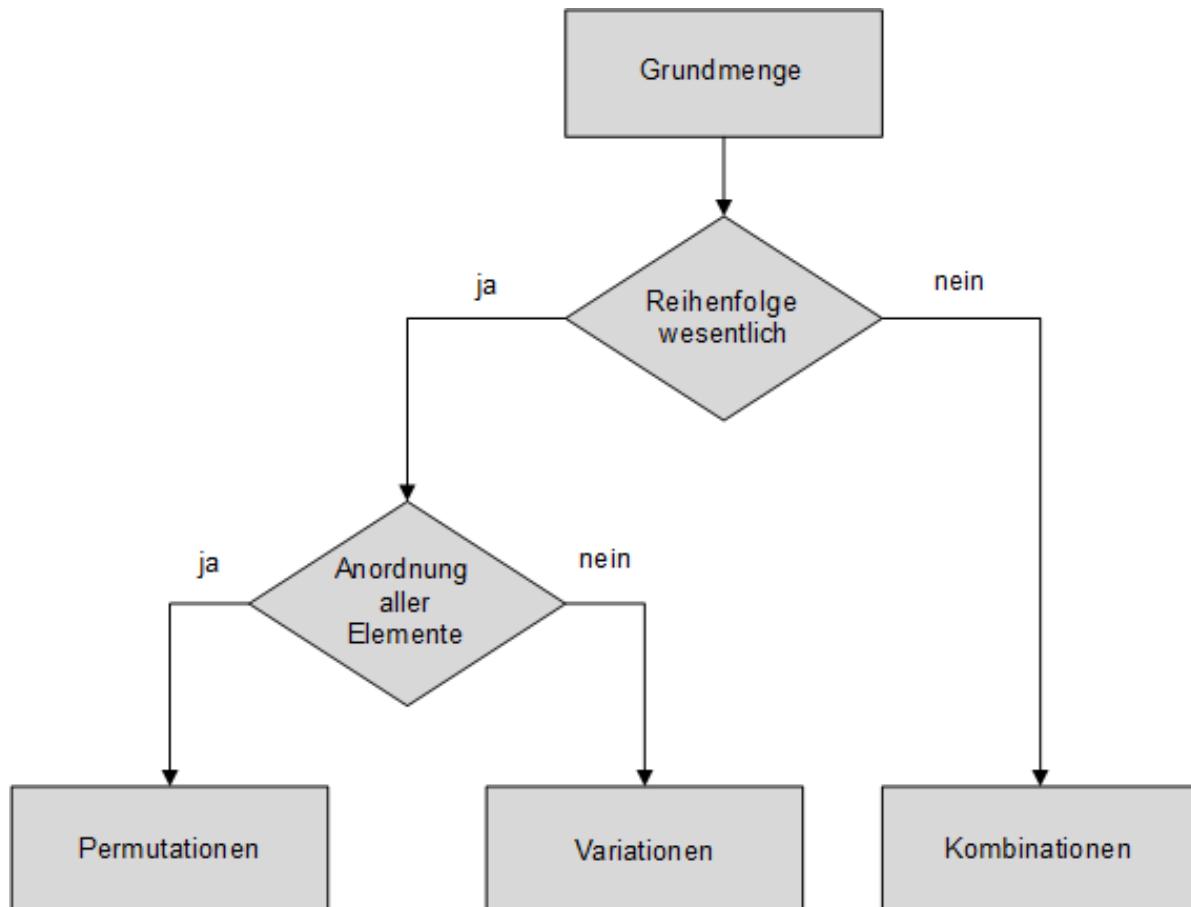


Bild 2.6: Entscheidungsprozess für die verschiedenen Varianten der Kombinatorik

Tabelle 2.2 fasst die Anzahl von Kombinationen und Variationen für die dargestellten Fälle mit und ohne Wiederholung zusammen.

Tabelle 2.2: Zusammenfassung der Berechnung von Permutationen, Kombinationen und Variationen

Variante der Kombinatorik	Berechnungsgleichung
Permutation ohne Klassenbildung	$M = N!$
Permutation mit Klassenbildung	$M = \frac{N!}{\prod_{i=1}^k N_i!} = \frac{N!}{N_k! \cdot N_{k-1}! \cdot \dots \cdot N_1!}$
Variation ohne Wiederholung	$M = \frac{N!}{(N-K)!}$
Variation mit Wiederholung	$M = N^K$
Kombination ohne Wiederholung	$M = \binom{N}{K} = \frac{N!}{K! \cdot (N-K)!}$
Kombination ohne Wiederholung	$M = \binom{N+K-1}{K} = \frac{(M+K-1)!}{K! \cdot (N-1)!}$

Befehle zur Berechnung in MATLAB

Zur Berechnung der Anzahl von Permutationen, Kombinationen und Variationen werden im wesentlichen drei MATLAB-Befehle verwendet:

Tabelle 2.3: MATLAB-Befehle zur Berechnung von Permutationen, Kombinationen und Variationen

Fakultät	$M = N!$	factorial(N)
Potenz	$M = N^K$	N^K
Binomialkoeffizient	$M = \binom{N}{K} = \frac{N!}{K! \cdot (N-K)!}$	nchoosek(N,K)

Befehle zur Berechnung in Python

Zur Berechnung der Anzahl von Permutationen, Kombinationen und Variationen werden im wesentlichen drei Python-Befehle verwendet:

Tabelle 2.4: Python-Befehle zur Berechnung von Permutationen, Kombinationen und Variationen

Fakultät	$M = N!$	scipy.special.factorial
Potenz	$M = N^K$	scipy.special.factorial
Binomialkoeffizient	$M = \binom{N}{K} = \frac{N!}{K! \cdot (N-K)!}$	scipy.special.comb

Mit den berechneten Werten lassen sich die Anzahl möglicher Ereignisse und die Anzahl günstiger Ereignisse bestimmen. Mit den Ergebnissen wird die Wahrscheinlichkeit für das entsprechende Ereignis berechnet.

2.3 Aufbau von Ereignisbäumen

Oftmals können Zufallsprozesse aus mehreren nacheinander ablaufenden Zufallsexperimenten modelliert werden. In dem Fall handelt es sich um ein sogenanntes mehrstufiges Zufallsexperiment. Ein anschauliches, grafisches Hilfsmittel bei der Berechnung von Wahrscheinlichkeiten solcher mehrstufigen Zufallsexperimente ist der Ereignisbaum [Papu01][Papu01]. Er besteht aus einer Wurzel, dem Ausgangspunkt der Zufallsexperimente, mehreren Verzweigungspunkten und einer Vielzahl von Zweigen.

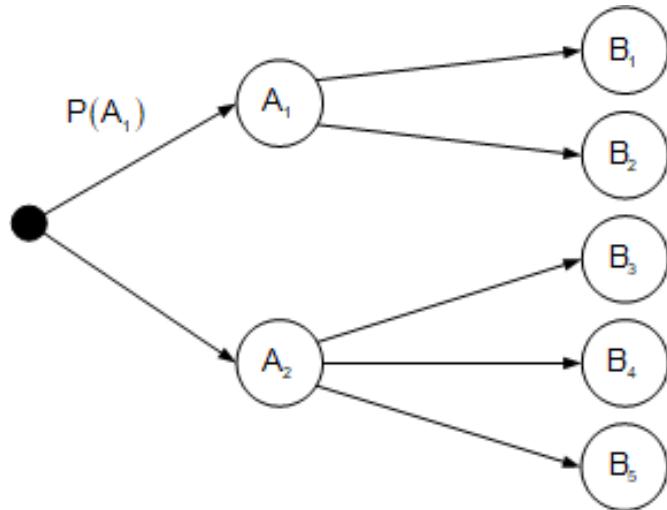


Bild 2.7: Darstellung eines Ereignisbaumes

In Bild 2.7 ist der prinzipielle Aufbau eines zweistufigen Ereignisbaumes zu erkennen. Der Ausgangspunkt des Zufallsexperiments, die Wurzel, wird durch einen schwarzen Punkt dargestellt. Die Verzweigungspunkte A_1 und A_2 charakterisieren die möglichen Ereignisse für das erste Zufallsexperiment. Die von diesen Verzweigungspunkten ausgehenden Zweige führen zu den jeweils möglichen Ergebnissen der folgenden 2. Stufe, die durch B_1 bis B_5 bezeichnet sind.

Die Übergangswahrscheinlichkeit von einem Ereignis zu einem Folgeereignis wird an den betreffenden Zweig geschrieben. So gibt $P(A_1)$ an, mit welcher Wahrscheinlichkeit das Ereignis A_1 eintritt. Ein mögliches Endergebnis des Zufallsprozesses wird dann immer von der Wurzel ausgehend längs eines Pfades erreicht. Dieser besteht meist aus mehreren Zweigen. Die Berechnung der Wahrscheinlichkeit eines bestimmten Endergebnisses erfolgt unter Anwendung folgender Rechenregeln:

- Wahrscheinlichkeiten entlang eines Pfades werden multipliziert
- Führen mehrere Pfade zum gleichen Endergebnis, so addieren sich ihre Wahrscheinlichkeiten

Beispiel: Ereignisbaum für das zweimalige Würfeln

Um den Umgang mit Ereignisbäumen zu verdeutlichen, wird das zweimalige Würfeln eines regelmäßigen Würfels betrachtet. Es handelt sich dabei um ein zweistufiges Zufallsexperiment. Für das Würfeln mit einem Würfel können eine Menge der geraden Zahlen und eine Menge der ungeraden Zahlen definiert werden. Beide Mengen sind disjunkt, zusammen stellen sie das sichere Ereignis dar, weil jede gewürfelte Augenzahl entweder gerade oder ungerade ist.

Als Beispiel wird die Wahrscheinlichkeit dafür ausgerechnet, dass bei zwei Würfen mit einem regelmäßigen Würfel eine gerade Augensumme erzielt wird. Dieses Ereignis wird als B bezeichnet. Es ergibt sich der in Bild 2.8 dargestellte Ereignisbaum.

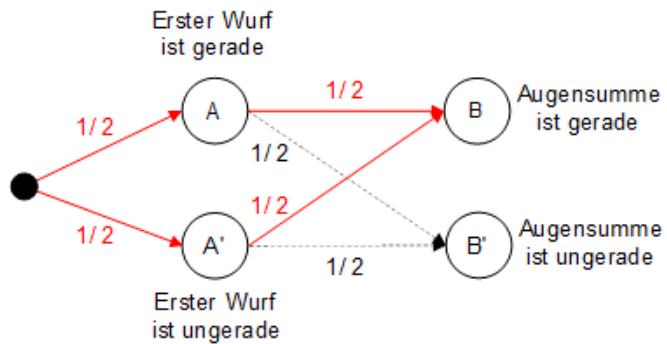


Bild 2.8: Darstellung eines Ereignisbaumes für das zweimalige Würfeln

Die gerade Augensumme kann sich auf zweierlei Arten ergeben, da beim ersten Wurf entweder eine gerade oder eine ungerade Augenzahl gewürfelt werden kann. Wird beim ersten Wurf eine gerade Zahl gewürfelt, muss auch beim zweiten Wurf eine gerade Zahl gewürfelt werden. Wird beim ersten Wurf eine ungerade Zahl gewürfelt, muss auch beim zweiten Wurf eine ungerade Zahl gewürfelt werden. Die Gesamtwahrscheinlichkeit ergibt sich aus der Summe der Wahrscheinlichkeiten dieser beiden Pfade.

$$P(B) = \frac{1}{2} \cdot \frac{1}{2} + \frac{1}{2} \cdot \frac{1}{2} = \frac{1}{2} \quad (2.36)$$

2.4 Wahrscheinlichkeitsbegriff der Statistik nach Kolmogoroff

Die klassische Wahrscheinlichkeitsrechnung nach Laplace beruht darauf, den Begriff der Wahrscheinlichkeit auf die Gleichwahrscheinlichkeit von Elementarereignissen zurückzuführen. Ereignismengen werden mit den vorgestellten Mengenoperationen auf Elementarereignisse zurückgeführt. Es existieren aber auch Zufallsexperimente, bei denen der Ereignisraum unendlich viele Elementarereignisse aufweist. Ein Ereignis wird damit ein stetiges Kontinuum. Soll zum Beispiel die Toleranzverteilung bei der Fertigung von Passstiften analysiert werden, ist der Ereignisraum kontinuierlich. Der Wahrscheinlichkeitsbegriff von Laplace kann deshalb nicht angewendet werden. Aus diesem Grund hat Kolmogoroff ihn verallgemeinert.

Der statistische Begriff der Wahrscheinlichkeit nach Kolmogoroff beruht auf Axiomen. Ein Axiom ist eine grundlegende Aussage, die zu anderen Axiomen widerspruchsfrei ist. Aus Axiomen lassen sich alle sonstigen Sätze des Systems logisch ableiten. Den Ausgangspunkt für den Wahrscheinlichkeitsbegriff der Statistik nach Kolmogoroff bildet die Erfahrung, dass das Eintreffen von Ereignissen bei den meisten Zufallsprozessen bei vielfacher Wiederholung einer gewissen Gesetzmäßigkeit unterliegt. Insbesondere erweist sich die relative Häufigkeit eines Experimentes für große Stichprobenumfänge als stabil. Diese ist definiert als Quotient aus der absoluten Häufigkeit, mit der ein Wert vorliegt, und dem Stichprobenumfang N.

$$h(A) = \frac{h_A(A)}{N} \quad (2.37)$$

Bild 2.9 zeigt für ein Würfelexperiment mit einem regelmäßigen Würfel die relative Häufigkeit, eine gerade Zahl zu würfeln.

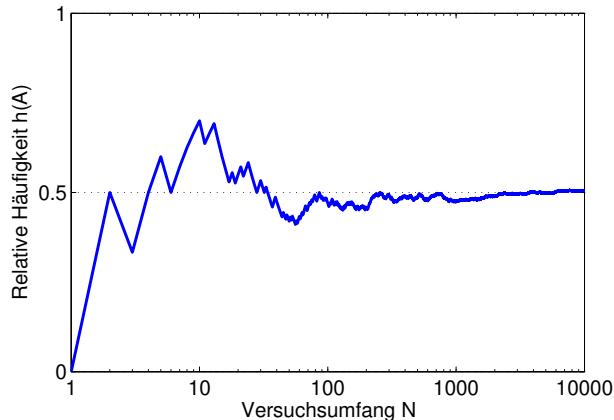


Bild 2.9: Relative Häufigkeit mit einem regelmäßigen Würfel eine gerade Zahl zu würfeln als Funktion der Anzahl von Versuchen

Sind die relativen Häufigkeiten der Ereignisse bei einem Experiment nahezu konstant, weist das Experiment eine statistische Regelmäßigkeit oder Stabilität der relativen Häufigkeit auf. Die meisten praktischen Zufallsexperimente haben diese Stabilitätseigenschaft. Trifft bei wiederholtem Ausführen eines Experimentes das Ereignis A mit der relativen Häufigkeit $h(A)$ ein, so wird die relative Häufigkeit als die Wahrscheinlichkeit $P(A)$ für das Ereignis A definiert. Die Wahrscheinlichkeit $P(A)$ ist damit das theoretische Gegenstück zu der empirischen relativen Häufigkeit $h(A)$, auf die in Kapitel 3 eingegangen wird.

2.4.1 Axiome der Wahrscheinlichkeitsrechnung nach Kolmogoroff

Für die Wahrscheinlichkeit $P(A)$ gelten nach Kolmogoroff folgende Axiome:

Axiom 1

Die Wahrscheinlichkeit $P(A)$ eines Ereignisses A bei einem Experiment ist eine eindeutig bestimmte reelle, nicht negative Zahl, die höchstens gleich 1 ist.

$$0 \leq P(A) \leq 1 \quad (2.38)$$

Dieses Axiom wird von dem Wahrscheinlichkeitsbegriff nach Laplace erfüllt. Die Anzahl der günstigen Fälle eines Experimentes ist stets kleiner oder gleich der Anzahl der möglichen Fälle.

Axiom 2

Ein sicheres Ereignis S bei einem Experiment hat die Wahrscheinlichkeit 1.

$$P(S) = 1 \quad (2.39)$$

Sind zwei Ereignisse B und C äquivalent, so sind ihre Wahrscheinlichkeiten gleich groß.

$$P(B) = P(C) \quad (2.40)$$

Auch dieses Axiom wird von der Laplaceschen Definition der Wahrscheinlichkeit erfüllt. Im Fall des sicheren Ereignisses S entspricht die Anzahl der günstigen Fälle gerade der Anzahl der möglichen Fälle, sodass sich als Wahrscheinlichkeit der Wert 1 ergibt. Sind zwei Ereignisse identisch, so ist die Anzahl der günstigen Fälle gleich groß und in beiden Fällen ergibt sich die gleiche Wahrscheinlichkeit.

Axiom 3

Schließen sich zwei Ereignisse A und B bei einem Experiment gegenseitig aus, so gilt bei dem Experiment

$$P(A \cup B) = P(A) + P(B) \quad (2.41)$$

Beispiel: Würfeexperiment und Axiome nach Kolmogoroff

Das erste Axiom ist für das Würfeexperiment erfüllt. Die Anzahl der günstigen Fälle des Würfeexperimentes A ist stets kleiner oder gleich der Anzahl der möglichen Fälle. Damit liegt die Wahrscheinlichkeit $P(A)$ in dem Bereich $0 \leq P(A) \leq 1$.

Sind alle Ereignisse beim Würfeln günstig, ergibt sich das sichere Ereignis S . Die Anzahl günstiger Fälle entspricht der Anzahl möglicher Fälle und die Wahrscheinlichkeit ist $P(S) = 1$.

Auch das dritte Axiom kann an dem Würfeexperiment mit einem Würfel verdeutlicht werden. Das Würfeln der Zahlen 1 und 2 schließt sich gegenseitig aus. Die Wahrscheinlichkeit für einen Wurf mit der Augenzahl 1 oder 2 entspricht der Vereinigungsmenge $A \cup B$. Die klassische Wahrscheinlichkeit ergibt sich aus der Summe der Wahrscheinlichkeit für den Wurf einer 1 und der einer 2.

2.4.2 Sätze zur Wahrscheinlichkeitsrechnung nach Kolmogoroff

Auf Basis dieser drei Axiome zur Wahrscheinlichkeit lassen sich weitere Sätze ableiten, die im Folgenden dargestellt werden.

Additionssatz der Wahrscheinlichkeit für komplementäre Ereignisse

Ein Ereignis, das genau dann eintrifft, wenn das Ereignis A nicht eintrifft, wird als komplementäres Ereignis A' bezeichnet. Da entweder das Ereignis A oder das komplementäre Ereignis A' eintrifft, gilt

$$P(A \cup A') = P(S) = 1 \quad (2.42)$$

Gemäß Axiom 3 gilt

$$P(A) + P(A') = 1 \quad (2.43)$$

oder

$$P(A') = 1 - P(A) \quad (2.44)$$

Aus diesem Satz folgt für die Wahrscheinlichkeit für ein unmögliches Ereignis U

$$P(U) = 1 - P(S) = 1 - 1 = 0 \quad (2.45)$$

Beispiel: Würfelexperiment und Additionssatz für komplementäre Ereignisse

Zum Beispiel berechnet sich die Wahrscheinlichkeit des Ereignisses D, beim Würfeln eine 4 zu erzielen, aus

$$P(D) = \frac{1}{6} \quad (2.46)$$

Die Wahrscheinlichkeit D', keine 4 zu erzielen, ergibt sich entsprechend zu

$$P(D') = 1 - \frac{1}{6} = \frac{5}{6} \quad (2.47)$$

Additionssatz der Wahrscheinlichkeit für mehrere sich ausschließende Ereignisse

Sind A_1, A_2, \dots, A_N abzählbar unendlich viele Ereignisse, die sich bei einem Experiment gegenseitig ausschließen, so gilt bei diesem Experiment

$$P\left(\bigcup_{n=1}^N A_n\right) = P(A_1 \cup A_2 \cup \dots \cup A_N) = P(A_1) + P(A_2) + \dots + P(A_N) = \sum_{n=1}^N P(A_n) \quad (2.48)$$

Dieser Additionssatz für beliebig viele Ereignisse, die sich gegenseitig ausschließen, kann durch wiederholte Anwendung von Axiom 3 der statistischen Wahrscheinlichkeit hergeleitet werden.

Additionssatz der Wahrscheinlichkeit für beliebige Ereignisse

Schließen sich die Ereignisse A und B eines Zufallsexperiments nicht gegenseitig aus und haben sie die Wahrscheinlichkeiten $P(A)$ und $P(B)$, so gilt für die Wahrscheinlichkeit für das Ereignis $A \cup B$

$$P(A \cup B) = P(A) + P(B) - P(A \cap B) \quad (2.49)$$

Zum Beweis des Satzes wird die Vereinigungsmenge in zwei unabhängige Mengen zerlegt.

$$P(A \cup B) = P(A \cup (A' \cap B)) \quad (2.50)$$

Mit Axiom 3 der Wahrscheinlichkeit nach Kolmogoroff folgt

$$P(A \cup B) = P(A) + P(A' \cap B) \quad (2.51)$$

Addition und Subtraktion des Ausdrucks $P(A \cap B)$ führt zu

$$\begin{aligned} P(A \cup B) &= P(A) + P(A' \cap B) + P(A \cap B) - P(A \cap B) \\ &= P(A) + (P(A' \cap B) + P(A \cap B)) - P(A \cap B) \\ &= P(A) + P(B) - P(A \cap B) \end{aligned} \quad (2.52)$$

Beispiel: Würfelexperiment und Additionssatz für beliebige Ereignisse

Anschaulich kann die Wahrscheinlichkeit wie bereits bei dem Laplaceschen Begriff der Wahrscheinlichkeit erklärt werden.

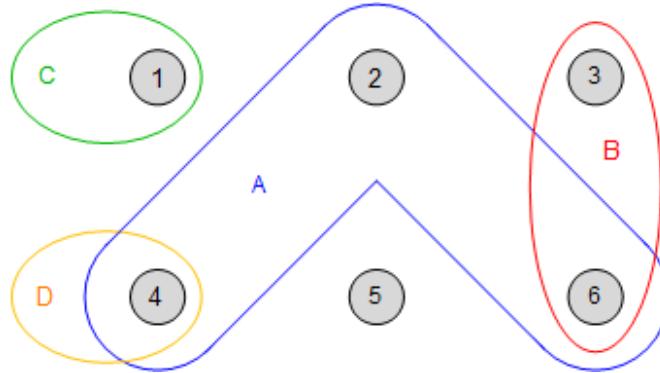


Bild 2.10: Darstellung des Zufallsexperiments Würfeln mit einem regelmäßigen Würfel

Für das Beispiel ergibt sich die Wahrscheinlichkeit für das Ereignis $A \cup B$ aus

$$P(A \cup B) = P(A) + P(B) - P(A \cap B) = \frac{1}{2} + \frac{1}{3} - \frac{1}{6} = \frac{2}{3} \quad (2.53)$$

Das Ergebnis kann dadurch verifiziert werden, dass die Wahrscheinlichkeiten der Würfe mit den Augenzahlen 2, 3, 4 und 6 addiert werden. Das Subtrahieren der Wahrscheinlichkeit $P(A \cap B)$ in Gleichung (2.53) ist notwendig, damit die Wahrscheinlichkeit des Wurfes mit der Augenzahl 6 nicht doppelt in das Ergebnis eingeht.

Bedingte Wahrscheinlichkeit

Sind A und B Ereignisse und ist bei einem Experiment $P(A) \neq 0$, so wird mit der Schreibweise $P(B|A)$ die Wahrscheinlichkeit des Ereignisses B unter der Voraussetzung oder Hypothese verstanden, dass das Ereignis A bereits eingetroffen ist. $P(B|A)$ wird auch als die bedingte Wahrscheinlichkeit des Ereignisses B unter der Hypothese A bezeichnet. Sie berechnet sich aus

$$P(B|A) = \frac{P(A \cap B)}{P(A)} \quad (2.54)$$

Beispiel: Würfeexperiment

Als Beispiel für die Berechnung der bedingten Wahrscheinlichkeit wird wieder das Würfeexperiment herangezogen. Das Ereignis A ist das Würfeln einer geraden Zahl, das Ereignis B_2 ist das Würfeln der Zahl 2. Gesucht ist die Wahrscheinlichkeit $P(B_2|A)$, also die Wahrscheinlichkeit für das Würfeln der Zahl 2 unter der Bedingung, dass eine gerade Augenzahl gewürfelt wurde.

Das Beispiel kann in Form eines Ereignisbaumes modelliert werden. Ausgehend von der Wurzel kann eine gerade oder eine ungerade Zahl gewürfelt werden. Jedes dieser Zwischenergebnisse unterteilt sich in drei mögliche Endergebnisse. Die Wahrscheinlichkeiten ergeben sich daher wie in Bild 2.11 dargestellt.

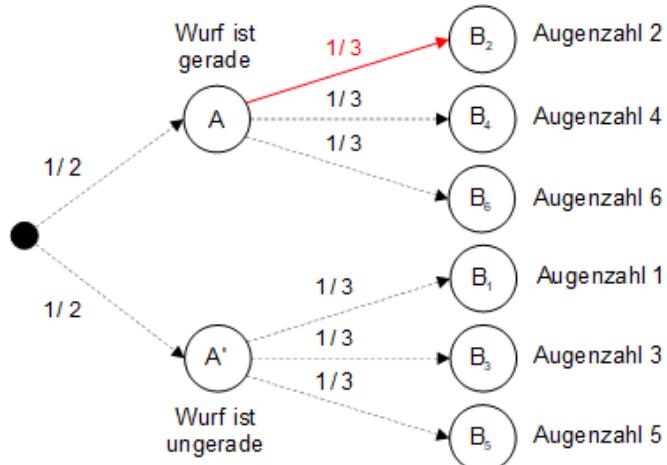


Bild 2.11: Ereignisbaum zur Modellierung eines Würfeperiments

Die Wahrscheinlichkeit für eine gerade Augenzahl ergibt sich zu

$$P(A) = \frac{1}{2} \quad (2.55)$$

Um die Wahrscheinlichkeit $P(A \cap B_2)$ zu berechnen, werden die Wahrscheinlichkeiten des entsprechenden Pfades multipliziert

$$P(A \cap B_2) = \frac{1}{2} \cdot \frac{1}{3} = \frac{1}{6} \quad (2.56)$$

Damit berechnet sich die bedingte Wahrscheinlichkeit von

$$P(B_2|A) = \frac{P(A \cap B_2)}{P(A)} = \frac{1/6}{1/2} = \frac{1}{3} \quad (2.57)$$

Diese bedingte Wahrscheinlichkeit entspricht der Übergangswahrscheinlichkeit in dem Ereignisbaum. Sie kann direkt aus dem Ereignisbaum des Zufallsexperiments abgelesen werden. Unter der Annahme, dass eine gerade Zahl gewürfelt wurde, kann direkt die bedingte Wahrscheinlichkeit $P(B_2|A)$ am rot eingezeichneten Zweig abgelesen werden.

Multiplikationssatz der Wahrscheinlichkeit

Nach Gleichung (2.54) gilt für die bedingte Wahrscheinlichkeit

$$P(B|A) = \frac{P(A \cap B)}{P(A)} \quad (2.58)$$

und bei Tauschen der Variablen A und B

$$P(A|B) = \frac{P(A \cap B)}{P(B)} \quad (2.59)$$

Damit kann die Wahrscheinlichkeit für die Schnittmenge $P(A \cap B)$ angegeben werden. Haben zwei Ereignisse A und B bei einem Experiment die Wahrscheinlichkeit $P(A)$ und $P(B)$, so beträgt die Wahrscheinlichkeit für gleichzeitiges Eintreffen von den Ereignissen A und B

$$P(A \cap B) = P(A|B) \cdot P(B) = P(B|A) \cdot P(A) \quad (2.60)$$

Beispiel: Würfeexperiment

Diese Aussagen lassen sich ebenfalls mit dem Würfeexperiment veranschaulichen. Ein Ereignis A ist dadurch definiert, dass beim Würfeln mit einem regelmäßigen Würfel eine gerade Zahl gewürfelt wird. Das Ereignis B ist dadurch definiert, dass eine durch 3 teilbare Zahl gewürfelt wird. Damit ergibt sich die Wahrscheinlichkeit für die Schnittmenge aus beiden Aussagen zu

$$P(A \cap B) = P(A|B) \cdot P(B) = \frac{1}{2} \cdot \frac{2}{6} = \frac{1}{6} \quad (2.61)$$

oder zu

$$P(B \cap A) = P(B|A) \cdot P(A) = \frac{1}{3} \cdot \frac{1}{2} = \frac{1}{6} \quad (2.62)$$

Beide Ergebnisse sind identisch und geben die Wahrscheinlichkeit an, mit der die Zahl 6 gewürfelt wird, denn es ist das einzige Ereignis, das die Bedingungen A und B erfüllt.

Statistische Unabhängigkeit von Ereignissen

Ist das Eintreffen eines Ereignisses A von dem Eintreffen eines Ereignisses B unabhängig, so gilt

$$P(A|B) = P(A) \quad (2.63)$$

und

$$P(B|A) = P(B) \quad (2.64)$$

Die Ereignisse A und B werden in diesem Fall auch als statistisch unabhängig bezeichnet. Durch Einsetzen vereinfacht sich Gleichung (2.60) für den Fall unabhängiger Variablen zu

$$P(A \cap B) = P(A|B) \cdot P(B) = P(A) \cdot P(B) = P(B) \cdot P(A) = P(B|A) \cdot P(A) \quad (2.65)$$

Die statistische Unabhängigkeit ist bei Würfelexperimenten unmittelbar erkennbar. Bei komplexeren Vorgängen kann eine statistische Unabhängigkeit oft nicht so klar erkannt werden. Gleichung (2.65) ermöglicht es, statistische Unabhängigkeit nachzuweisen. Der Begriff der statistischen Unabhängigkeit lässt sich auch auf mehr als zwei unabhängige Variable erweitern.

$$P\left(\bigcap_{n=1}^N A_n\right) = \prod_{n=1}^N P(A_n) \quad (2.66)$$

Diese Rechenregeln können als Test für die statistische Unabhängigkeit von Größen verwendet werden. Im Fall der statistischen Unabhängigkeit muss die Bedingung

$$P(A \cap B) = P(A) \cdot P(B) = P(B) \cdot P(A) \quad (2.67)$$

erfüllt sein.

Beispiel: Statistische Unabhängigkeit beim Würfelexperiment

Als Beispiel wird die Wahrscheinlichkeit berechnet, bei zweimaligem Würfeln mit einem regelmäßigen Würfel zwei Sechsen zu erhalten. Die Wahrscheinlichkeit, beim ersten Wurf eine 6 zu würfeln beträgt

$$P(A) = \frac{1}{6} \quad (2.68)$$

Da der Würfel nicht verändert wird, ist die Ausgangssituation beim zweiten Wurf dieselbe wie beim ersten. Das Ergebnis des ersten Zuges hat also keinerlei Einfluss auf das Ergebnis des zweiten. Demnach handelt es sich um unabhängige Ereignisse. Die Wahrscheinlichkeit, beim zweiten Wurf eine 6 zu erhalten, beträgt ebenfalls

$$P(B) = \frac{1}{6} \quad (2.69)$$

Die Wahrscheinlichkeit für das zweimalige Würfeln einer 6 beträgt

$$P = \frac{1}{6} \cdot \frac{1}{6} = \frac{1}{36} \quad (2.70)$$

Dasselbe Ergebnis ergibt sich bei der Berechnung der Gesamtwahrscheinlichkeit des Endergebnisses im Ereignisbaum.

Beispiel: Statistische Unabhängigkeit von Spezifikationsmerkmalen

Bei einer Untersuchung von Einspritzdüsen werden 100 Einspritzdüsen auf ihre Toleranzen hin analysiert. Dabei werden der Durchmesser D und die Rauigkeit R gemessen. Ist der Durchmesser außerhalb der Spezifikation oder die Rauigkeit zu hoch, wird das Bauteil als defekt eingestuft. Die Untersuchung führt zu dem in Tabelle 2.5 dargestellten Klassifikationsergebnis.

Tabelle 2.5: Vorgehen bei der Berechnung der Systemantwort mit der z-Transformation

Spezifikation Rauigkeit R	Spezifikation Durchmesser D	
	Erfüllt	Nicht erfüllt
Erfüllt	$91 P(R \cap D) = 0.91$	$3P(R \cap D') = 0.03$
Nicht erfüllt	$5P(R' \cap D) = 0.05$	$1P(R' \cap D') = 0.01$

Die Wahrscheinlichkeit $P(D)$, mit der die Spezifikation des Durchmessers D erfüllt ist, beträgt

$$P(D) = \frac{96}{100} = 0.96 \quad (2.71)$$

Die Wahrscheinlichkeit $P(R)$, mit der die Rauigkeit innerhalb der spezifizierten Werte liegt, errechnet sich zu

$$P(R) = \frac{94}{100} = 0.94 \quad (2.72)$$

Zur Untersuchung der statistischen Unabhängigkeit der Ereignisse D und R wird Gleichung (2.65) herangezogen.

$$P(D \cap R) = \frac{91}{100} = 0.91 \neq 0.9024 = 0.96 \cdot 0.94 = P(D) \cdot P(R) \quad (2.73)$$

Da die Gleichung nicht erfüllt ist, sind die Ereignisse D und R nicht statistisch voneinander unabhängig. Die Aussage zur statistischen Unabhängigkeit wird hier auf Basis einer Stichprobe gefällt. Die Aussage ist damit zunächst nur für die Stichprobe gültig. Eine Verallgemeinerung der Aussage auf die entsprechende Grundgesamtheit wird in Kapitel 5 begonnen.

Totale Wahrscheinlichkeit

In einigen Fällen ist es vorteilhaft, die Wahrscheinlichkeit eines Ereignisses B über die Berechnung der bedingten Wahrscheinlichkeit $P(B|A_n)$ zu berechnen. Für die Herleitung wird von der Schnittmenge aus dem Ereignis B und dem sicheren Ereignis S ausgegangen.

$$B = B \cap S \quad (2.74)$$

Das sichere Ereignis S soll aus N disjunkten Ereignissen A_n bestehen, sodass gilt

$$S = A_1 \cup A_2 \cup \dots \cup A_N \quad (2.75)$$

Wegen der statistischen Unabhängigkeit disjunkter Ereignisse gilt die Beziehung

$$P(S) = P(A_1 \cup A_2 \cup \dots \cup A_N) = 1 \quad (2.76)$$

Die Wahrscheinlichkeit der Schnittmenge der beiden Ereignisse B und S

$$P(B) = P(B \cap S) = P(B \cap (A_1 \cup A_2 \cup \dots \cup A_N)) \quad (2.77)$$

berechnet sich unter den beschriebenen Voraussetzungen mit den Rechenregeln für den Umgang mit Mengen aus Tabelle 2.1 zu

$$P(B) = P((B \cap A_1) \cup (B \cap A_2) \cup \dots \cup (B \cap A_N)) \quad (2.78)$$

Da es sich bei den Ereignissen A_1 bis A_N definitionsgemäß um disjunkte Mengen handelt, kann die Wahrscheinlichkeit aus Gleichung (2.78) nach dem Additionssatz der Wahrscheinlichkeit für mehrere sich ausschließende Ereignisse als Summe dargestellt werden, die mit dem Multiplikationssatz der Wahrscheinlichkeit aus Gleichung (2.58) weiter umgeformt werden kann.

$$\begin{aligned} P(B) &= P(B \cap A_1) + P(B \cap A_2) + \dots + P(B \cap A_N) \\ &= P(B|A_1) \cdot P(A_1) + P(B|A_2) \cdot P(A_2) + \dots + P(B|A_N) \cdot P(A_N) = \sum_{n=1}^N P(B|A_n) \cdot P(A_n) \end{aligned} \quad (2.79)$$

Diese Gleichung wird als Satz der totalen Wahrscheinlichkeit bezeichnet. Sie trägt der Tatsache Rechnung, dass ein Ereignis auf verschiedene Arten zustande kommen kann. Der Satz der totalen Wahrscheinlichkeit findet sich auch in der Berechnung der Wahrscheinlichkeit eines Ereignisses mit einem Ereignisbaum wieder. Hierbei werden die Wahrscheinlichkeiten der einzelnen Zweige ebenfalls addiert.

Beispiel: Zulieferer und Defekte von Transistoren

Ein Betrieb bezieht Transistoren für den Bau von Wechselrichtern von zwei Zulieferern. Zulieferer 1 liefert insgesamt 60 % der Transistoren, die restlichen 40 % stammen von Zulieferer 2. In den Lieferverträgen wird dem Zulieferer 1 eine Ausschussquote von 0.03 % zugestanden, in der Lieferung von Zulieferer 2 dürfen maximal 0.05 % defekte Transistoren enthalten sein. Für die beschriebene Ausgangssituation kann der Ereignisbaum aufgestellt werden.

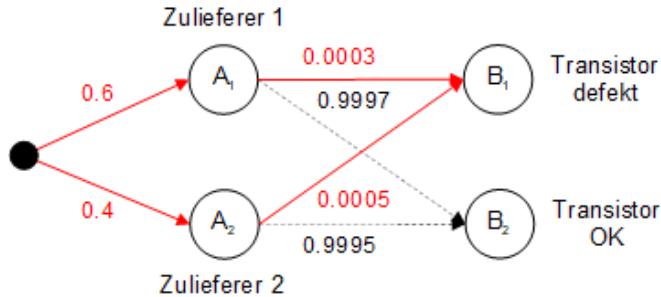


Bild 2.12: Ereignisbaum zur Liefersituation der Transistoren

Es soll die Wahrscheinlichkeit für den Einbau eines defekten Transistors berechnet werden, also die Wahrscheinlichkeit, mit der das Ereignis B_1 eintritt. In dem Ereignisbaum in Bild 2.12 ist zu erkennen, dass das Ereignis B_1 auf zwei Pfaden erreicht wird, die in rot eingezzeichnet sind. Die Berechnung erfolgt mit dem Satz der totalen Wahrscheinlichkeit zu

$$P(B_1) = P(B_1|A_1) \cdot P(A_1) + P(B_1|A_2) \cdot P(A_2) = 0.0003 \cdot 0.6 + 0.0005 \cdot 0.4 = 0.00038 \quad (2.80)$$

Die Wahrscheinlichkeit, dass ein defekter Transistor eingebaut wird, beträgt somit 380 ppm.

Satz von Bayes

Die Berechnung der Wahrscheinlichkeit $P(B|A)$ für ein Ereignis B unter der Bedingung, dass ein Ereignis A eingetroffen ist, ist teilweise einfacher zu berechnen als die Wahrscheinlichkeit $P(B|A)$. Durch Anwendung der Regeln für die bedingte Wahrscheinlichkeit aus Gleichung (2.60)

$$P(B|A) \cdot P(A) = P(A|B) \cdot P(B) \quad (2.81)$$

ergibt sich die Wahrscheinlichkeit $P(B|A)$ aus der Wahrscheinlichkeit $P(B|A)$ zu

$$P(A|B) = \frac{P(B|A) \cdot P(A)}{P(B)} \quad (2.82)$$

Wird die Wahrscheinlichkeit $P(B)$ ausgedrückt über die totale Wahrscheinlichkeit, ergibt sich der Satz von Bayes.

$$P(A|B) = \frac{P(B|A) \cdot P(A)}{P(B)} = \frac{P(B|A) \cdot P(A)}{\sum_{n=1}^N P(B|A_n) \cdot P(A_n)} \quad (2.83)$$

Beispiel: Zulieferer und Defekte von Transistoren

Als Anwendung des Satzes von Bayes wird die Frage diskutiert, mit welcher Wahrscheinlichkeit ein eingebauter defekter Transistor von Zulieferer 2 stammt, wenn die im vorigen Beispiel beschriebene Situation vorliegt. Die Berechnung der gesuchten Wahrscheinlichkeit $P(A_2|B_1)$ erfolgt über den Satz von Bayes zu

$$\begin{aligned} P(A_2|B_1) &= \frac{P(B_1|A_2) \cdot P(A_2)}{P(B_1)} \\ &= \frac{P(B_1|A_2) \cdot P(A_2)}{\sum_{n=1}^2 P(B_1|A_n) \cdot P(A_n)} = \frac{0.0005 \cdot 0.4}{0.0003 \cdot 0.6 + 0.0005 \cdot 0.4} = 52.63\% \end{aligned} \quad (2.84)$$

Unter der Bedingung, dass ein defekter Transistor vorliegt, stammt er mit einer Wahrscheinlichkeit von 52.63 % von Zulieferer 2, obwohl dieser weniger Transistoren liefert als Zulieferer 1.

Zusammenfassung der Sätze zur Wahrscheinlichkeit nach Kolmogoroff

Tabelle 2.6 fasst die Sätze zur Wahrscheinlichkeit nach Kolmogoroff zusammen.

Tabelle 2.6: Tabellarische Übersicht der an der Übertragungsfunktion ablesbaren Systemeigenschaften

Regel	Rechenvorschrift
Additionssatz	$P(A \cup B) = P(A) + P(B) - P(A \cap B)$
Additionssatz komplementärer Ereignisse	$P(A) + P(\bar{A}) = 1$
Additionssatz unabhängiger Ereignisse	$P(A_1 \cup \dots \cup A_N) = P(A_1) + \dots + P(A_N)$
Bedingte Wahrscheinlichkeit	$P(A B) = \frac{P(A \cap B)}{P(B)}$
Multiplikationssatz	$P(A \cap B) = P(A B) \cdot P(B) = P(B A) \cdot P(A)$
Statistische Unabhängigkeit	$P(A B) = P(A)$ $P(B A) = P(B)$ $P(A \cap B) = P(A) \cdot P(B)$
Totale Wahrscheinlichkeit	$P(B) = \sum_{n=1}^N P(B A_n) \cdot P(A_n)$
Satz von Bayes	$P(A B) = \frac{P(B A) \cdot P(A)}{\sum_{n=1}^N P(B A_n) \cdot P(A_n)}$

Mit den Regeln und Beispielen aus diesem Kapitel ist der Begriff der Wahrscheinlichkeit eingeführt. Außerdem sind grundlegende Rechenregeln zur Berechnung der Wahrscheinlichkeit bekannt, sie werden im Folgenden in einem Projekt angewandt.

2.5 Anwendungsbeispiel: Sensordiagnose im Steuergerät

Zur Erkennung von Sensorfehlern werden in Steuergeräten Diagnoseverfahren eingesetzt. Sie überwachen Ausgangsspannungen von Sensoren und plausibilisieren unterschiedliche Sensorsignale miteinander.



Bild 2.13: Motorsteuergerät

Diagnosefunktionen sollten idealerweise so arbeiten, dass sie eine Fehlfunktion erkennen, wenn sie tatsächlich vorliegt und nicht reagieren, wenn der entsprechende Sensor nicht fehlerhaft ist. Die Diagnose von Fehlern ist jedoch nicht ideal. Reale Diagnosekonzepte haben eine endliche Aussagesicherheit. Die Auswirkungen werden im Folgenden mithilfe der Wahrscheinlichkeitsrechnung analysiert.

Von einer Diagnosefunktion ist bekannt, dass sie mit einer Wahrscheinlichkeit von 99.99 % einen defekten Sensor als defekt erkennt. Außerdem ist bekannt, dass mit einer Wahrscheinlichkeit von 0.02 % funktionsfähige Sensoren als defekt eingestuft werden. Außerdem ist aus einer Ausfallstatistik bekannt, dass die Wahrscheinlichkeit für einen Sensordefekt bei 100 ppm liegt.

Es wird die Frage diskutiert, mit welcher Wahrscheinlichkeit ein Sensor wirklich defekt ist, wenn er mit der oben diskutierten Diagnosefunktion als defekt erkannt wurde. Um die Aufgabe zu lösen, werden zu Beginn zwei Ereignisse definiert:

- Ereignis A: Sensor ist defekt
- Ereignis B: Diagnoseergebnis ist positiv, Sensor wird als defekt eingestuft

Die Aufgabenstellung kann auch mithilfe eines Ereignisbaumes wie in Bild 2.14 beschrieben werden. An die entsprechenden Zweige werden die bekannten Wahrscheinlichkeiten gekennzeichnet. Da ein sicheres Ereignis S bei einem Experiment die Wahrscheinlichkeit 1 besitzt, sind auch die Wahrscheinlichkeiten der übrigen Zweige bekannt.

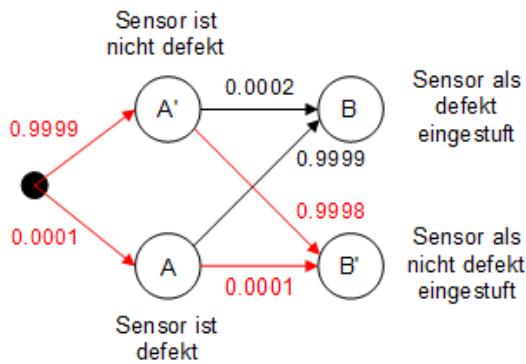


Bild 2.14: Ereignisbaum zum Beispiel der Sensordiagnose

In der Fragestellung ist die Wahrscheinlichkeit gesucht, mit der unter der Voraussetzung eines positiven Diagnoseergebnisses der Sensor tatsächlich defekt ist. Dies entspricht der bedingten Wahrscheinlichkeit $P(A|B)$.

Bekannt ist die Wahrscheinlichkeit $P(B|A) = 99.99\%$, mit der ein defekter Sensor als defekt erkannt wird, und die Wahrscheinlichkeit $P(B|A') = 0.02\%$, mit der ein funktionstüchtiger Sensor als defekt eingestuft wird. Außerdem beträgt die Wahrscheinlichkeit für einen Sensordefekt $P(A) = 100\text{ppm} = 10^{-4}$.

Da es sich um eine bedingte Wahrscheinlichkeit handelt, kann mit dem Satz von Bayes gearbeitet werden.

$$\begin{aligned} P(A|B) &= \frac{P(B|A) \cdot P(A)}{P(B|A) \cdot P(A) + P(B|A') \cdot P(A')} \\ &= \frac{0.9999 \cdot 10^{-4}}{0.9999 \cdot 10^{-4} + 0.0002 \cdot (1 - 10^{-4})} = 33.33\% \end{aligned} \quad (2.85)$$

Im Folgenden soll die Vorgehensweise bei der Berechnung mit dem Ereignisbaum aus Bild 2.14 gezeigt werden. Um die Wahrscheinlichkeit zu berechnen, mit der unter der Voraussetzung eines positiven Diagnoseergebnisses der Sensor tatsächlich defekt ist, muss der Ereignisbaum umgestellt werden.

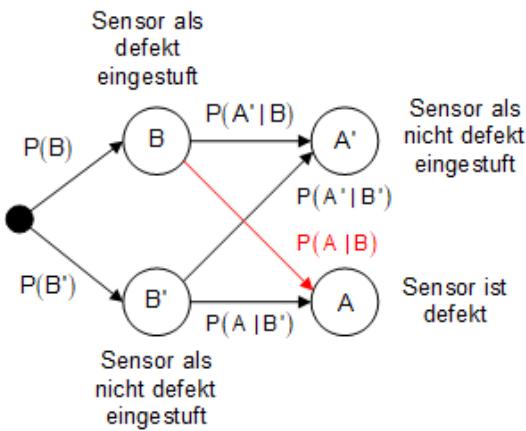


Bild 2.15: Umgestellter Ereignisbaum zum Beispiel der Sensordiagnose

Zur Berechnung der gesuchten Wahrscheinlichkeit muss zunächst die Wahrscheinlichkeit für ein positives Testergebnis $P(B)$ bestimmt werden. Mit dem Ereignisbaum aus Bild 2.14 oder dem Satz der totalen Wahrscheinlichkeit folgt diese zu

$$P(B) = P(B|A') \cdot P(A') + P(B|A) \cdot P(A) = 0.0002 \cdot 0.9999 + 0.9999 \cdot 10^{-4} = 3 \cdot 10^{-4} \quad (2.86)$$

Da die Wahrscheinlichkeit des Pfades 0-A-B im Ereignisbaum in Bild 2.14 gleich der Wahrscheinlichkeit

0-B-A in Bild 2.15 sein muss, kann die Wahrscheinlichkeit für einen defekten Sensor bei einem positiven Diagnoseergebnis berechnet werden.

$$P(A|B) = \frac{P(B|A) \cdot P(A)}{P(B)} = \frac{0.9999 \cdot 10^{-4}}{3 \cdot 10^{-4}} = 33.33\% \quad (2.87)$$

Bei näherer Betrachtung entspricht diese Rechnung dem Satz von Bayes.

Das Ergebnis der Berechnung besagt, dass trotz des positiven Diagnoseergebnisses die Wahrscheinlichkeit für einen Sensordefekt lediglich bei 33.33 % liegt. Ursache dafür ist, dass der Anteil funktionsfähiger Sensoren sehr groß ist. In Bild 2.16 ist diese Aussagesicherheit als Funktion der Wahrscheinlichkeit $P(B|A')$ dargestellt, mit der ein funktionstüchtiger Sensor als defekt eingestuft wird. Der in die Grafik eingezeichnete Punkt gibt das Ergebnis für $P(B|A') = 0.02\%$ an.

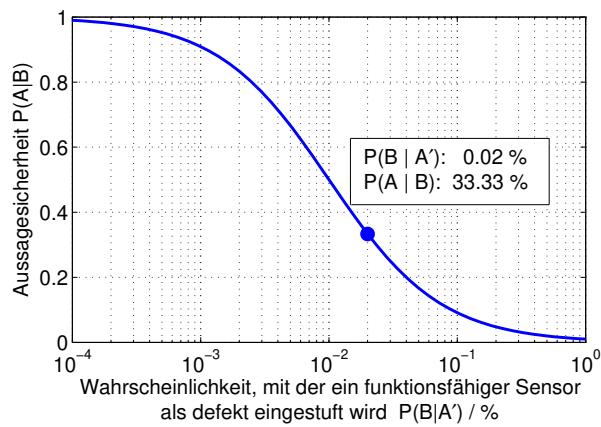


Bild 2.16: Wahrscheinlichkeit mit der ein Sensor wirklich defekt ist, wenn er mit einer Diagnosefunktion als defekt eingestuft wurde

Die Rechnung zeigt, dass sich die Wahrscheinlichkeit $P(B|A')$, mit der ein funktionstüchtiger Sensor als defekt eingestuft wird, stark in das Ergebnis einwirkt. Entwickler von Diagnosefunktionen sind deshalb bestrebt, gerade diesen Fehler zu minimieren.

2.6 Literatur

- [Krey91] Kreyszig, Erwin: Statistische Methoden und ihre Anwendungen
4., unveränderter Nachdruck der 7. Auflage
Vandenhoeck & Ruprecht, Göttingen, 1991
- [Papu01] Papula, Lothar: Mathematik für Ingenieure und Naturwissenschaftler Band 3
4., verbesserte Auflage
Vieweg Teubner, Braunschweig / Wiesbaden, 2008
- [Fahr06] Fahrmeir, Ludwig; Künstler, Rita; Pigeot, Iris; Tutz, Gerhard: Der Weg zur Datenanalyse
6. Auflage
Springer Berlin Heidelberg New York, 2006

3 Beschreibende Statistik univariater Daten

In der Serienfertigung und der automatisierten Messtechnik entsteht eine Vielzahl von Daten. Die Datensätze sind oft komplex und unübersichtlich, eine Interpretation aller Daten im Detail ist zudem zeitaufwändig. Vor diesem Hintergrund ist es erforderlich, die Daten übersichtlich darstellen oder auf Kenngrößen komprimieren zu können.

Als Einstieg in die Statistik beschäftigt sich dieses Kapitel deshalb mit der Frage, wie Daten numerisch und grafisch aufbereitet werden können. Anschließend wird die zusammenfassende Beschreibung von Datensätzen mit Hilfe statistischer Kenngrößen eingeführt.

3.1 Merkmalstypen

Der Design For Six Sigma (DFSS) Prozess zeichnet sich dadurch aus, dass über den gesamten Prozess quantitative Methoden eingesetzt werden. Die Ergebnisse sind dabei von definierten Merkmalen abhängig. Die Merkmale können unterschiedlicher Natur sein. Vor dem Einsatz statistischer Methoden ist es notwendig, die unterschiedlichen Merkmalstypen zu klassifizieren, da sie einen Einfluss auf die Methodik und die Genauigkeit der Aussage haben. Die für den Design For Six Sigma Prozess relevanten Merkmalstypen lassen sich in stetige und diskrete Größen, sowie ordinale und gruppierende Größen aufteilen.

3.1.1 Stetige Merkmale

Stetige Merkmale können eine Eigenschaft beliebig fein wiedergeben. Es entsteht kein Fehler durch die Darstellung des Ergebnisses, allenfalls durch die Aufzeichnung des Messergebnisses. Beispiele für stetige Merkmale sind Temperaturen, elektrische Spannungen und Ströme, geometrische Maße wie Strecken oder Flächeninhalte sowie die Zeit. Stetige Merkmale werden bei einer Verarbeitung der Werte im Rechner diskretisiert. Sind die Stufen der Diskretisierung eine Größenordnung kleiner als die kleinste darzustellende Größe, kann die Quantisierung vernachlässigt werden. Die Merkmale werden als quasi-stetig bezeichnet.

3.1.2 Diskrete Merkmale

Diskrete Merkmale haben nur endlich viele Ausprägungen. Zum Beispiel kann ein Wurf mit einem Würfel nur die Zahlen eins bis sechs annehmen, und er weist nur endlich viele unterschiedliche Ereignisse auf. Eine Erfassung von stetigen Größen mit diskreten Messmitteln führt zu einer diskreten Messgröße. Beispielsweise führt die Messung einer Spannung mit einem 6 Bit Analog-Digital-Wandler und einem Messbereich von 5 V zu einem Quantisierungsintervall von

$$\Delta U = \frac{5V}{2^6} = 78mV \quad (3.1)$$

Ist diese Diskretisierung größer als ein Zehntel der interessierenden kleinsten Spannung, wird das ursprünglich stetige Merkmal als diskretes Merkmal bezeichnet.

Ein diskretes Merkmal entsteht auch bei der Klassenbildung von stetigen Merkmalen. Zum Beispiel ist es denkbar, die Rohwerte einer Widerstandsmessung in Klassen zusammenzufassen, die einem Widerstandsbereich entsprechen. Nach der Klassenbildung kann nicht mehr entschieden werden, ob der Widerstand am unteren oder oberen Ende des Intervalls lag. Aus dem stetigen Merkmal ist durch die Klassenbildung ein diskretes Merkmal entstanden.

3.1.3 Ordinale Merkmale

Ordinale Datentypen werden für Daten verwendet, die nach ihrer Ausprägung geordnet werden können, deren Abstände aber nicht interpretiert werden können. Ein typisches Beispiel dafür sind Kontrollergebnisse, die zu einer Aussage „gut“, „mäßig“ oder „schlecht“ führen. Diese Aussage kann in Zahlen

wiedergegeben werden, zum Beispiel kann der Eigenschaft „gut“ die Zahl 1, „mäßig“ die Zahl 2 und „schlecht“ die Zahl 3 zugeordnet werden. Diese Zuordnung gibt jedoch nur ein Ordnungsschema an. Im Gegensatz zu den stetigen und diskreten Datentypen kann mit ordinalen Datentypen nicht sinnvoll gerechnet werden, vielleicht mit Ausnahme der Fuzzy Logik. Außerdem sind die Aussagen „gut“, „mäßig“ oder „schlecht“ deutlich größer und schlechter zu interpretieren als numerische Angaben mit stetigen oder diskreten Daten.

3.1.4 Gruppierende Merkmale

Gruppierende Merkmale sind zum Beispiel bei der Charakterisierung von Verfahren zu finden. Wird ein Verfahren geändert, erfolgt ein Vergleich des alten Verfahrens mit dem neuen. Auch Zulieferer lassen sich nicht ordnen, sie existieren parallel und können nicht nach einer Ausprägung geordnet werden. Gruppierende Merkmale werden deshalb im Allgemeinen auch nicht mit Zahlen bezeichnet.

3.1.5 Merkmalstypen und Aussagesicherheit

Aus der Beschreibung der unterschiedlichen Merkmalstypen ergibt sich, dass die Genauigkeit bei Messungen von stetigen Merkmalen zu gruppierenden Merkmalen kontinuierlich abnimmt. Ordinalen und gruppierenden Datentypen kann eine mathematisch berechnete Kenngröße nicht mehr sinnvoll zugeordnet werden. Aus diesem Grund unterscheiden sich auch die statistischen Methoden, die für die stetigen und diskreten Datentypen eingesetzt werden, von denen, die für die ordinalen und gruppierten Datentypen verwendet werden.

Der Schwerpunkt liegt in den folgenden Kapiteln auf stetigen und diskreten Datentypen, die nur von einer Einflussgröße abhängen.

3.2 Häufigkeitsverteilungen

Statistische Daten werden durch Aufzeichnen von Beobachtungsergebnissen gewonnen. Die dabei entstehende Liste wird als Urliste bezeichnet. Tabelle 3.1 stellt die Messwerte von $N = 100$ Widerständen mit einem Sollwert von $R = 1k\Omega$ dar. Die Daten weisen eine Auflösung von $\Delta R = 1\Omega$ auf. Es handelt sich deshalb um einen diskreten Merkmalstyp.

Tabelle 3.1: Beispiel für eine Urliste: Messwerte von 100 Widerständen mit einem Sollwert von $R = 1k\Omega$

Index	Messwert R / Ω									
1 - 10	983	988	985	987	988	987	986	985	986	991
11 - 20	987	986	987	986	985	988	986	986	988	985
21 - 30	985	989	986	986	985	992	988	989	986	986
31 - 40	985	986	986	986	989	988	986	986	986	987
41 - 50	989	986	986	985	988	990	986	986	988	987
51 - 60	985	989	987	985	986	990	986	985	986	988
61 - 70	985	988	984	988	986	985	987	989	986	987
71 - 80	987	987	985	987	986	986	986	987	985	989
81 - 90	988	992	985	986	987	987	985	988	984	988
91 - 100	987	988	985	986	986	985	987	989	986	985

Die in Tabelle 3.1 dargestellten Größen bilden eine Stichprobe mit dem Umfang $N = 100$, die einzelnen Messwerte werden allgemein als Stichprobenwerte bezeichnet. Mithilfe der Wahrscheinlichkeitsrechnung wird später versucht, von der Stichprobe auf die Grundgesamtheit, zum Beispiel aller Widerstände in einem definierten Fertigungszeitraum, zu schließen.

Zur Übersicht können die Daten in einem sogenannten Streudiagramm dargestellt werden. Dabei wird der Stichprobenindex als Abszisse und der Stichprobenwert als Ordinate dargestellt.

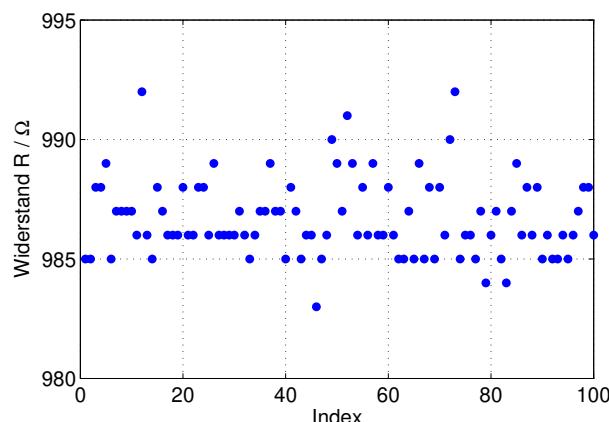


Bild 3.1: Darstellung der Stichprobe in Tabelle 3.1 als Streudiagramm

3.2.1 Absolute und relative Häufigkeit diskreter Merkmalstypen

Zur übersichtlichen Darstellung der Stichprobenwerte diskreter Merkmalstypen werden die Stichproben der Größe nach geordnet, und die Häufigkeit der einzelnen Werte wird ausgewertet. Bei manueller Ausführung ergibt sich zunächst eine Strichliste wie in Tabelle 3.2, aus der anschließend eine Häufigkeitsverteilung wie in Tabelle 3.3 abgeleitet werden kann.

Tabelle 3.2: Häufigkeitsbewertung über eine Strichliste

R / Ω	Absolute Häufigkeit $h_A(R)$	R / Ω	Absolute Häufigkeit $h_A(R)$
983	I	988	III IIII IIII
984	II	989	III III
985	III IIII IIII IIII	990	II
986	III IIII IIII IIII IIII II	991	I
987	III IIII IIII II	992	II

Tabelle 3.3: Häufigkeitsverteilung der Stichprobe

R / Ω	Absolute Häufigkeit $h_A(R)$	Relative Häufigkeit $h_A(R)$	R / Ω	Absolute Häufigkeit $h_A(R)$	Relative Häufigkeit $h_A(R)$
983	1	0.01	988	15	0.15
984	2	0.02	989	8	0.08
985	20	0.20	990	2	0.02
986	32	0.32	991	1	0.01
987	17	0.17	992	2	0.02

Die absolute Häufigkeit gibt an, wie oft der entsprechende Messwert x in der Stichprobe die Ausprägung x_n annimmt. Diese Anzahl wird als absolute Häufigkeit $h_A(x)$ bezeichnet. Die relative Häufigkeit $h(x)$ ergibt sich aus dem Quotient aus absoluter Häufigkeit $h_A(x)$ und dem Stichprobenumfang N .

$$h(x) = \frac{h_A(x)}{N} \quad (3.2)$$

Zum Beispiel kommt die Ausprägung mit einem Widerstandswert von $R = 988\Omega$ in der Stichprobe 15-mal vor. Da die Stichprobe insgesamt $N = 100$ Werte aufweist, ergibt sich eine relative Häufigkeit von 15%.

Kommt der Wert x_0 in der Stichprobe nicht vor, hat er die absolute Häufigkeit $h_A(x_0) = 0$ und nach Gleichung (2.37) auch die relative Häufigkeit $h(x_0) = 0$. Im anderen Extremfall könnten alle Stichprobenwerte die Ausprägung x_1 haben. In diesem Fall wäre die absolute Häufigkeit $h_A(x_1) = N$ und damit die relative Häufigkeit $h(x_1) = 1$. Die relative Häufigkeit $h(x)$ ist somit eine nicht negative Zahl, die höchstens den Wert 1 annehmen kann.

$$0 \leq h(x) \leq 1 \quad (3.3)$$

Die Zahlenwerte in Tabelle 3.3 stellen die Häufigkeitsverteilung $h(x)$ der Stichprobe dar. Sie ordnet jedem Wert x eine relative Häufigkeit $h(x)$ zu. Die Summe aller absoluten Häufigkeiten muss die Anzahl

von Stichprobenwerten N ergeben. Die Summe aller relativen Häufigkeiten ist damit 1. Es gilt:

$$h(x_1) + h(x_2) + \dots + h(x_N) = \sum_{n=1}^N h(x_n) = 1 \quad (3.4)$$

Zur besseren Übersicht können die absolute oder die relative Häufigkeit in Form von Stab- oder Linien-diagrammen dargestellt werden. Bild 3.2 stellt die relative Häufigkeit für die Stichprobe in Tabelle 3.1 als Histogramm dar.

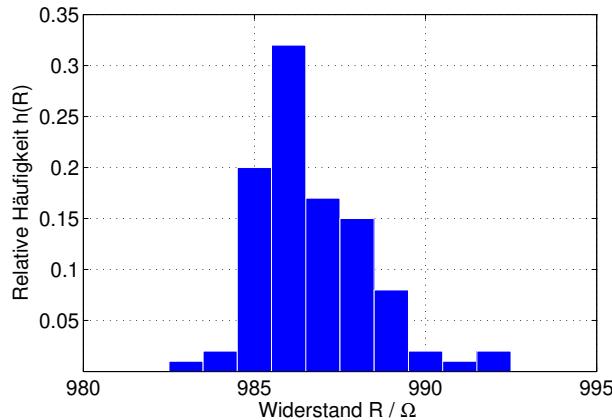


Bild 3.2: Darstellung der relativen Häufigkeiten als Stabdiagramm für die Stichprobe in Tabelle 3.1

Bei sehr vielen unterschiedlichen Zahlenwerten der Stichprobe wird die Häufigkeitsverteilung unübersichtlich. Aus diesem Grund können die Stichprobenwerte in Klassen eingeteilt werden. Ausgehend von dem Gesamtintervall, in dem die Stichprobenwerte liegen, werden Teilintervalle oder Klassenintervalle gebildet. Die Mittenwerte c_n der Intervalle heißen Klassenmitten. Die einzelnen Stichprobenwerte werden den entsprechenden Teilintervallen oder Klassen zugeordnet. Es ergibt sich die absolute und analog zu Gleichung (3.2) die relative Klassenhäufigkeit. Die Häufigkeit in Abhängigkeit der Klassenmitten heißt Häufigkeitsverteilung der in Klassen eingeteilten Stichprobe.

Nach der Aufteilung in Klassen treten die ursprünglichen Stichprobenwerte nicht mehr einzeln in Erscheinung, sie gehen nur als Summe in die Häufigkeitsverteilung der in Klassen eingeteilten Stichprobe ein. Je weniger Klassen gebildet werden, desto mehr Information geht verloren. Eine ungeeignete Anzahl oder Einteilung von Klassen führt zu unübersichtlichen oder falschen Interpretationsergebnissen. Bei der Definition von Stichprobenklassen haben sich folgende Regeln als sinnvoll erwiesen:

- Die Klassenintervalle sind gleich groß zu wählen.
- Die Klassenmitten sollen möglichst einfache Zahlen mit möglichst wenigen Ziffern sein.
- Die Anzahl von Klassen sollte zwischen 10 und 20 liegen, sinnvoll ist eine Anzahl von Klassen mit

$$\text{Anzahl} \approx \sqrt{N} \quad (3.5)$$

Werden die Widerstandswerte aus Tabelle 3.1 in drei oder sechs Klassen eingeteilt, ergibt sich die in Bild 3.3 dargestellte relative Häufigkeit.

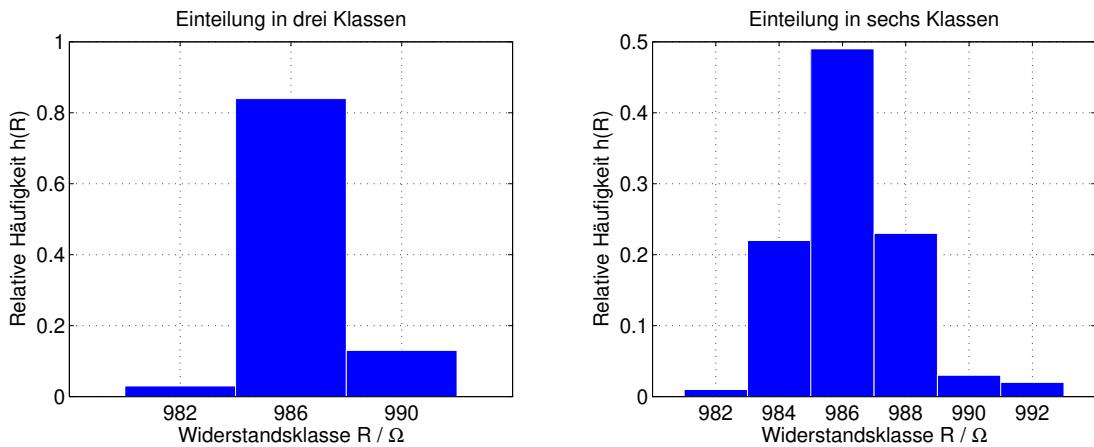


Bild 3.3: Darstellung der relativen Häufigkeiten für die Stichprobe in Tabelle 3.1

Dabei wird bei der Darstellung im Stabdiagramm das Prinzip der Flächentreue eingehalten. Es besagt, dass die Flächen direkt proportional zu den absoluten beziehungsweise relativen Häufigkeiten sein müssen. Wird der Abstand zwischen zwei benachbarten Klassenmitten c_n und c_{n-1} als Breite d mit

$$d = c_n - c_{n-1} \quad (3.6)$$

bezeichnet, so muss die Höhe h_n im Stabdiagramm den Wert

$$h_n = \frac{h(x_n)}{d} \quad (3.7)$$

aufweisen, damit die Fläche proportional zur relativen Häufigkeit wird

$$A_n = h_n \cdot d = \frac{h(x_n)}{d} \cdot d = h(x_n) \quad (3.8)$$

3.2.2 Absolute und relative Summenhäufigkeit diskreter Merkmalstypen

Die Häufigkeitsverteilung $h(x)$ der Stichprobe gibt die relativen Häufigkeiten an, mit der die einzelnen Zahlenwerte in der Stichprobe vorkommen. Oft stellt sich aber die Frage, wie viele Stichprobenwerte unter oder auf einem Grenzwert liegen. Soll zum Beispiel für das Beispiel aus Tabelle 3.1 die Frage beantwortet werden, wie viele Widerstände kleiner oder gleich 985Ω sind, muss die Summe

$$h(R \leq 985\Omega) = h(R = 983\Omega) + h(R = 984\Omega) + h(R = 985\Omega) = 0.23 \quad (3.9)$$

ausgewertet werden. Wird diese Summe für beliebige Werte x durchgeführt, ergibt sich die relative Summenhäufigkeit $H(x)$ der Stichprobe. $H(x)$ ist die Summe der relativen Häufigkeiten aller Stichprobenwerte, die kleiner oder gleich dem Wert x sind.

$$H(x) = \sum_{x_n=-\infty}^x h(x_n) \quad (3.10)$$

Bild 3.4 stellt die relative Häufigkeit $h(R)$ und die relative Summenhäufigkeit $H(R)$ für das Beispiel aus Tabelle 3.1 als Stabdiagramm dar.

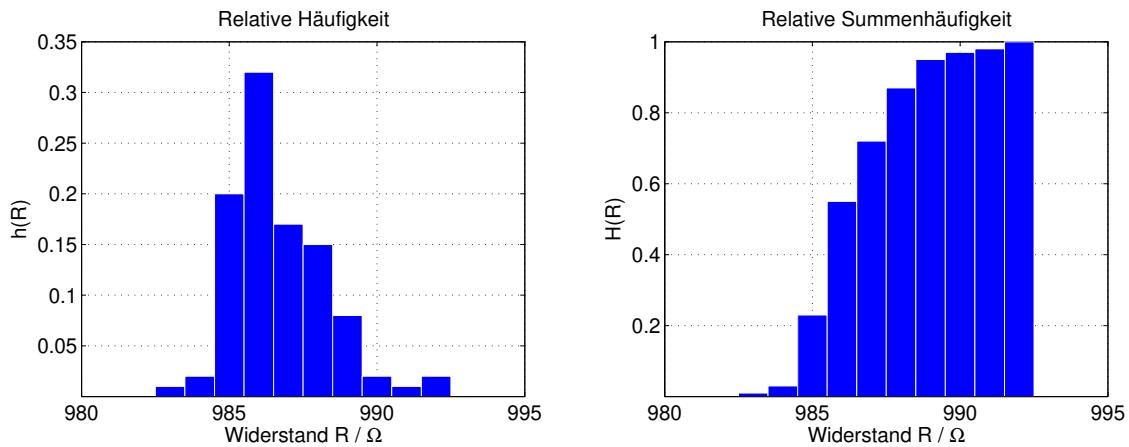


Bild 3.4: Darstellung der relativen Häufigkeit und Summenhäufigkeit für die Stichprobe in Tabelle 3.1

Die relative Summenhäufigkeit erscheint zunächst weniger anschaulich. In Abschnitt 3.2.3 wird sich aber zeigen, dass sie bei dem Übergang zu kontinuierlichen Stichprobenwerten einige Vorteile hat. Der Informationsgehalt ist bei der relativen Häufigkeit und der relativen Summenhäufigkeit identisch. Nach Gleichung (3.10) kann die Summenhäufigkeit aus der relativen Häufigkeit berechnet werden.

$$H(x) = \sum_{x_n=-\infty}^x h(x_n) \quad (3.11)$$

Die Summenhäufigkeit für den Wert $(x - d)$ ergibt sich entsprechend aus

$$H(x-d) = \sum_{x_n=-\infty}^{x-d} h(x_n) \quad (3.12)$$

Damit kann $h(x)$ bestimmt werden aus der Differenz

$$h(x) = H(x) - H(x-d) \quad (3.13)$$

Die Darstellungen können also mithilfe dieser Gleichungen ineinander überführt werden.

3.2.3 Beschreibung stetiger Merkmalstypen

Die anschauliche Beschreibung von Merkmalen mit einer relativen Häufigkeit versagt bei stetigen Merkmalstypen, weil die Messwerte beliebig fein aufgelöst sind und jeder Messwert typischerweise nur einmal vorkommt. Liegen Stichproben mit stetigen Merkmalen vor, können die Werte gruppiert werden. Damit ergibt sich eine Auswertung, wie sie in den Abschnitten 3.2.1 und 3.2.2 beschrieben ist. Allerdings gehen mit der Gruppierung der Merkmale Informationen verloren.

Alternativ können stetige Merkmale mit einer relativen Summenhäufigkeit beschrieben werden. Jeder Stichprobenwert einer Stichprobe mit stetigen Merkmalen kommt wegen der beliebig hohen Auflösung nur einmal vor. Bei einem Stichprobenumfang von N Werten weist jeder dieser Werte eine relative Häufigkeit von $1/N$ auf. Alle anderen Werte weisen die relative Häufigkeit von 0 auf. Durch Sortieren der Werte x nach der Größe ergibt sich eine geordnete Stichprobe. Die relative Summenhäufigkeit der Stichprobe ergibt sich aus

$$H(x) = \sum_{x_n=-\infty}^x h(x_n) = \sum_{x_n=-\infty}^x \frac{1}{N} \quad (3.14)$$

Je nach Werteverlauf der Stichprobe ergibt sich eine relative Summenhäufigkeit, die grafisch als Liniendiagramm dargestellt werden kann.

Bild 3.5 vergleicht die beiden Vorgehensweisen. Im linken Bildteil werden die Toleranzwerte einer Stichprobe mit $N = 1000$ Sensoren in Klassen eingeteilt und die relative Summenhäufigkeit als Balkendiagramm dargestellt. Im rechten Bildteil ist die relative Summenhäufigkeit ohne Gruppierung der Merkmale als Liniendiagramm eingezeichnet.

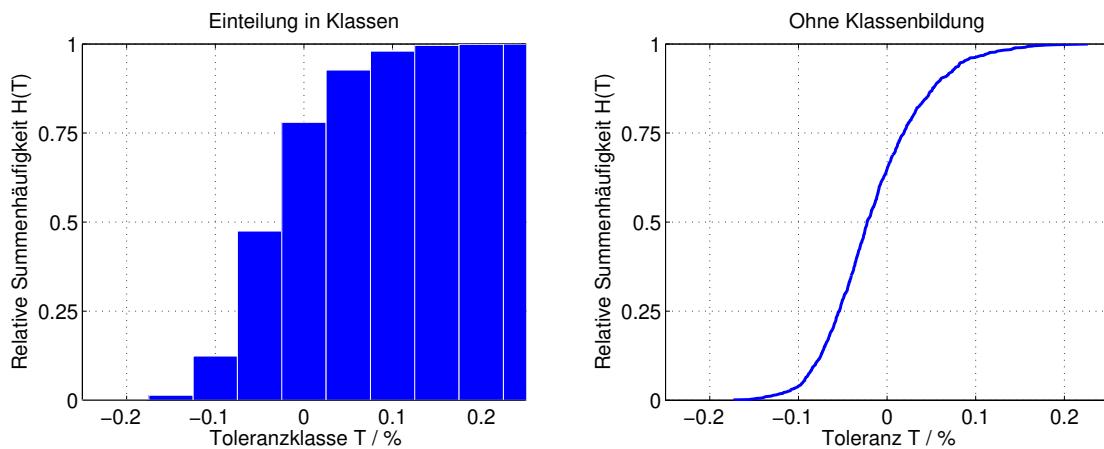


Bild 3.5: Darstellung der Toleranzwerte eines Sensors

Anhand der Darstellung in Bild 3.5 wird unmittelbar deutlich, dass durch die Klassenbildung Information verloren gegangen ist, die Auflösung der Summenhäufigkeit ist schlechter. Aus diesem Grund wird bei stetigen Merkmalstypen wenn möglich eine Klassenbildung vermieden.

3.2.4 Beschreibung ordinaler oder gruppierender Merkmalstypen

Gruppierende oder ordinale Merkmale können nicht als Stabdiagramm dargestellt werden, da keine numerische Einteilung der Abszissenachse möglich ist. Deshalb werden gruppierende oder ordinale Merkmale meist als Kreis- oder Balkendiagramm dargestellt. Ein Beispiel für einen gruppierenden Datensatz ist die Aufteilung einer Gesamtliefermenge von $N = 10000$ Teilen auf vier Zulieferer. In Tabelle 3.4 ist die Liefermenge der einzelnen Zulieferer aufgelistet.

Tabelle 3.4: Auswahl von Funktionen zur Signalverknüpfung

Zulieferer Z	Absolute Liefermenge $h_A(Z)$	Relative Liefermenge $h(Z)$
A	4000	40 %
B	2000	20 %
C	3000	30 %
D	1000	10 %

In Bild 3.6 ist links die grafische Darstellung der relativen Liefermengen der einzelnen Zulieferer Z als Kreisdiagramm zu sehen. Jeder Teilsektor entspricht dabei einer relativen Liefermenge, die dem einzelnen Zulieferer zugeordnet werden kann. Die Fläche der einzelnen Kreissektoren ist dabei proportional zu der relativen Häufigkeit der Liefermenge $h(Z)$. Der gesamte Kreis stellt die Summe aller Teilwerte dar, die Fläche bei der Darstellung von relativen Häufigkeiten ist damit 1. Gleicher gilt für die Darstellung der relativen Liefermenge $h(Z)$ als Stapelbalkendiagramm, wie sie rechts in Bild 3.6 dargestellt ist.

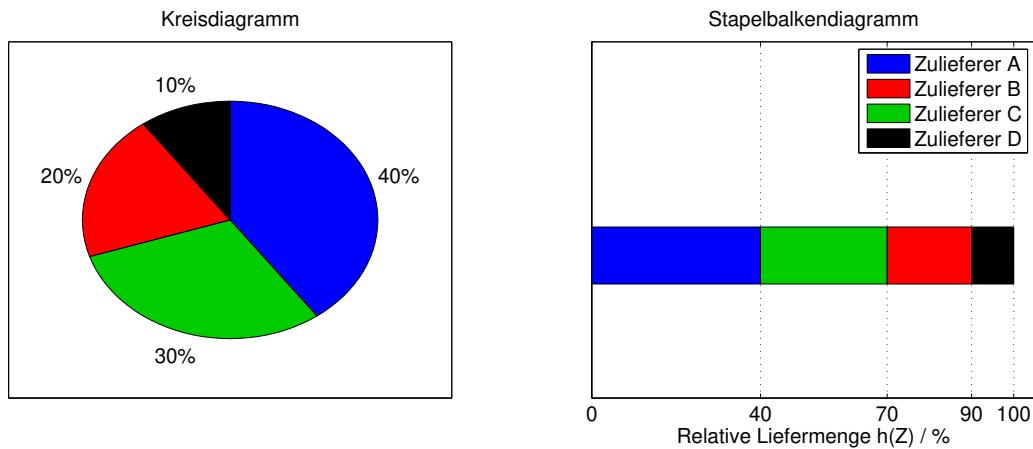


Bild 3.6: Darstellung der relativen Liefermenge als Kreisdiagramm und als Stapelbalkendiagramm

Bei Darstellungen mithilfe von Kreis- oder Stapelbalkendiagrammen sollte darauf geachtet werden, dass der Kreis beziehungsweise der Balken nicht in zu viele Sektoren unterteilt wird, da sonst die Übersichtlichkeit des Diagrammes verloren geht. Es empfiehlt sich in diesem Fall, mehrere Sektoren zusammenzufassen.

Ein Vergleich der einzelnen Zulieferer lässt sich am besten mit einem einfachen Balkendiagramm erzielen. Dieses ist für die Werte aus Tabelle 3.4 in Bild 3.7 dargestellt.

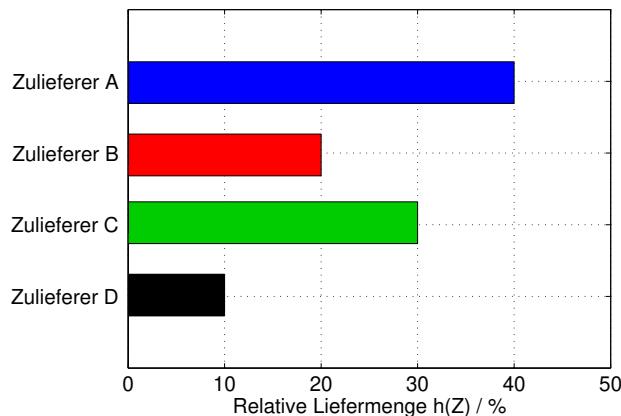


Bild 3.7: Darstellung der relativen Liefermenge als Balkendiagramm

Bei dem Balkendiagramm in Bild 3.7 ist im Vergleich zu Bild 3.6 gut zu erkennen, dass Zulieferer A mit einem Lieferanteil von 40 % das Vierfache von Zulieferer D liefert.

3.2.5 Befehle zur Beschreibung von Häufigkeiten in MATLAB

Zur Berechnung und Darstellung von Häufigkeitsverteilungen stehen in MATLAB diverse Funktionen zur Verfügung. Die wichtigsten sind in Tabelle 3.5 zusammengefasst.

Tabelle 3.5: Berechnung und Darstellung von Häufigkeitsverteilungen in MATLAB

MATLAB Befehl	Funktionsbeschreibung
sort(x)	Sortiert die Werte des Vektors x nach der Größe
cumsum(x)	Berechnet die kumulative Summe des Vektors x
hist(x)	Erzeugt ein Histogramm mit der absoluten Häufigkeit von x
tabulate(x)	Erstellt eine Häufigkeitstabelle aus x
bar(x)	Erzeugt ein Balkendiagramm ausx mit vertikaler Ausrichtung
barh(x)	Erzeugt ein Balkendiagramm ausX mit horizontaler Ausrichtung
pie(x)	Erzeugt ein Kreisdiagramm aus x

3.2.6 Befehle zur Beschreibung von Häufigkeiten in Python

Zur Berechnung und Darstellung von Häufigkeitsverteilungen stehen in Python diverse Funktionen zur Verfügung. Die wichtigsten sind in Tabelle 3.6 zusammengefasst.

Tabelle 3.6: Berechnung und Darstellung von Häufigkeitsverteilungen in Python

Python Befehl	Funktionsbeschreibung
sorted	Sortiert die Werte des Vektors x nach der Größe
numpy.cumsum	Berechnet die kumulative Summe des Vektors x
numpy.histogram	Erzeugt ein Histogramm mit der absoluten Häufigkeit von x
count	Erstellt eine Häufigkeitstabelle aus x
matplotlib.pyplot.bar	Erzeugt ein Balkendiagramm ausx mit vertikaler Ausrichtung
matplotlib.pyplot.barh	Erzeugt ein Balkendiagramm ausX mit horizontaler Ausrichtung
matplotlib.pyplot.pie	Erzeugt ein Kreisdiagramm aus x

3.3 Kennwerte einer Stichprobe

Jede Stichprobe wird mit ihrer Häufigkeitsverteilung oder Summenhäufigkeitsverteilung in allen Einzelheiten beschrieben. Gerade bei dem Vergleich größerer Datenmengen sind diese Häufigkeitsverteilungen aber unhandlich. Die Methoden der Statistik werden dazu verwendet, Stichproben über charakteristische Kenngrößen oder Maßzahlen abstrakt zu beschreiben. Dabei sind Kenngrößen für die Lage, die Streuung und die Symmetrie einer Stichprobe zu unterscheiden.

3.3.1 Lagekennwerte einer Stichprobe

Lagekennwerte beschreiben die Lage des Zentrums einer Verteilung durch einen numerischen Wert.

Arithmetisches Mittelwert einer Stichprobe

Der arithmetische Mittelwert einer Stichprobe ist wohl die bekannteste Lagekenngröße. Er ist definiert als

$$\bar{x} = \frac{x_1 + \dots + x_N}{N} = \frac{1}{N} \cdot \sum_{n=1}^N x_n \quad (3.15)$$

Liegen die Daten in Klassen mit ihren Klassenmittnen c_n vor, berechnet sich der arithmetische Mittelwert aus

$$\bar{x} = \frac{c_1 \cdot h_A(c_1) + \dots + c_N \cdot h_A(c_N)}{N} = \frac{1}{N} \cdot \sum_{n=1}^N (c_n \cdot h_A(c_n)) = \sum_{n=1}^N (c_n \cdot h(c_n)) \quad (3.16)$$

Der arithmetische Mittelwert hat zwei wichtige Eigenschaften. Zum einen ist die Summe der Differenzen aller Werte von ihrem Mittelwert null. Diese Eigenschaft lässt sich durch Zerlegen der Summe beweisen.

$$\sum_{n=1}^N (x_n - \bar{x}) = \sum_{n=1}^N x_n - \sum_{n=1}^N \bar{x} = \sum_{n=1}^N x_n - N \cdot \bar{x} = N \cdot \bar{x} - N \cdot \bar{x} = 0 \quad (3.17)$$

Zum anderen kann gezeigt werden, dass die Summe der Quadrate der Differenzen aller Werte von ihrem Mittelwert kleiner ist als die Summe der Quadrate der Differenzen aller Werte zu irgendeinem anderen Wert z.

$$\sum_{n=1}^N (x_n - \bar{x})^2 < \sum_{n=1}^N (x_n - z)^2 \quad (3.18)$$

Der arithmetische Mittelwert ist eine Kenngröße, die gerade bei kleinem Stichprobenumfang stark von einzelnen Ausreißern abhängig sein kann.

Beispiel: Kapazitätsmessung

Als Beispiel sind in Tabelle 3.7 zwei Messreihen dargestellt, in denen jeweils $N = 10$ Kondensatoren mit einem nominalen Kapazitätswert von $C = 100 \text{ nF}$ vermessen wurden. In der zweiten Messreihe wurde ein Messwert durch einen Ausreißer ersetzt.

Tabelle 3.7: Beispiel für eine Urliste: Messwerte von 100 Widerständen mit einem Sollwert von $R = 1\text{k}\Omega$

Messreihe 1: Kapazitätswerte C / nF									
101	102	99	98	100	97	99	100	101	103
Messreihe 2: Kapazitätswerte C / nF									
101	153	99	98	100	97	99	100	101	103

Der arithmetische Mittelwert für Messreihe 1 ergibt sich aus

$$\bar{C}_1 = \frac{1}{N} \cdot \sum_{n=1}^N C_{n,1} = \frac{101 + 102 + \dots + 103}{10} \text{ nF} = 100 \text{ nF} \quad (3.19)$$

Wegen des Messfehlers in der zweiten Messreihe weicht der arithmetische Mittelwert stark von dem nominellen Wert ab. Für Messreihe 2 ergibt sich der Mittelwert zu

$$\bar{C}_2 = \frac{1}{N} \cdot \sum_{n=1}^N C_{n,2} = \frac{101 + 153 + \dots + 103}{10} \text{ nF} = 105.1 \text{ nF} \quad (3.20)$$

Das Beispiel bestätigt die Empfindlichkeit des arithmetischen Mittelwertes gegenüber Ausreißern. Mit fallendem Stichprobenumfang steigt der Einfluss der Fehlmessung weiter an.

Median einer Stichprobe

Ein Lagemaß, das weniger empfindlich auf Ausreißer reagiert als der arithmetische Mittelwert, ist der Median. Der Median x_{MED} ist der Wert, bei dem die Hälfte der Stichprobenwerte x größer und die Hälfte der Stichprobenwerte x kleiner ist.

$$H(x_{MED}) = 0.5 \quad (3.21)$$

Bei ungeradem Stichprobenumfang n ist der Median x_{MED} der mittlere Wert der nach Größe geordneten Stichprobe. Bei geradem Stichprobenumfang ergibt sich der Median aus dem Mittelwert der beiden Stichprobenwerte, die in der Mitte der geordneten Stichprobe liegen.

Beispiel: Kapazitätsmessung

Zur Bestimmung des Medians für die Messreihen aus Tabelle 3.7 müssen die Daten zunächst sortiert werden.

Tabelle 3.8: Beispiel für eine Urliste: Messwerte von 100 Widerständen mit einem Sollwert von $R = 1k\Omega$

Messreihe 1: Kapazitätswerte C / nF									
97	98	99	99	100	100	101	101	102	103
Messreihe 2: Kapazitätswerte C / nF									
97	98	99	99	100	100	101	101	103	153

Aus den Messreihen 1 und 2 ergeben sich die beiden Mediane aus den arithmetischen Mittelwerten des fünften und sechsten Elementes zu

$$C_{MED,1} = \frac{C_{5,1} + C_{6,1}}{2} = 100nF \quad (3.22)$$

beziehungsweise

$$C_{MED,2} = \frac{C_{5,2} + C_{6,2}}{2} = 100nF \quad (3.23)$$

Das Beispiel zeigt, dass der Median deutlich unempfindlicher auf Messausreißer reagiert als der arithmetische Mittelwert, in diesem Beispiel bleibt er sogar konstant. Für den Median ist zum Beispiel der Absolutwert des größten und des kleinsten Stichprobenwertes völlig gleichgültig. Der Median wird deshalb als resistenter oder robuster Lagekennwert bezeichnet.

Liegen die Daten in gruppieter Form mit konstanter Klassenbreite d vor, kann die relative Lage des Medians in der Klasse berücksichtigt werden. Dazu wird zwischen den benachbarten Klassenmittnen c_n ein linearer Zusammenhang zwischen der Summenhäufigkeit $H(x)$ und dem Merkmal x angenommen. Bild 3.8 stellt diese Annahme grafisch dar.

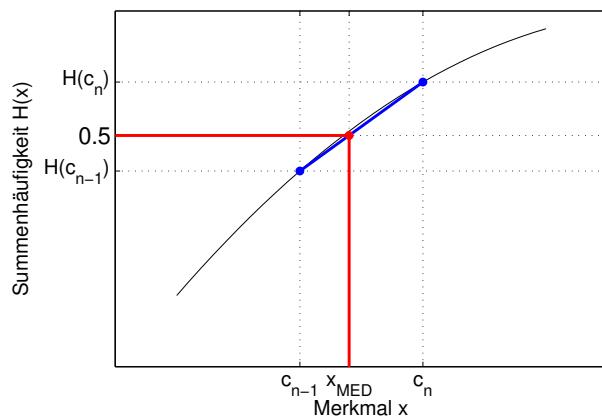


Bild 3.8: Lineare Interpolation zwischen Summenhäufigkeit $H(x)$ und Merkmal x

Die Steigung der Geraden zwischen den bekannten Werten für die Klassenmittnen ergibt sich aus

$$m = \frac{H(c_n) - H(c_{n-1})}{c_n - c_{n-1}} = \frac{h(c_n)}{d} \quad (3.24)$$

Daraus folgt für die Summenhäufigkeit des Medians die Geradengleichung

$$H(x_{MED}) = 0.5 = H(c_{n-1}) + m \cdot (x_{MED} - c_{n-1}) = H(c_{n-1}) + \frac{h(c_n)}{d} \cdot (x_{MED} - c_{n-1}) \quad (3.25)$$

Sie kann nach dem Median aufgelöst werden. Es ergibt sich

$$x_{MED} = c_{n-1} + \frac{d \cdot (0.5 - H(c_{n-1}))}{h(c_n)} \quad (3.26)$$

Beispiel: Widerstandsmessung

Für das Beispiel aus Tabelle 3.3 ergibt sich damit als Median

$$R_{MED} = 985\Omega + \frac{1\Omega \cdot (0.5 - 0.23)}{0.32} = 985.84\Omega \quad (3.27)$$

Der Wert erreicht eine Auflösung, die unterhalb der Klassenbreite d liegt. Sowohl das arithmetische Mittel als auch der Median müssen bei diskreten Datentypen nicht mit einer möglichen Ausprägung x_n übereinstimmen. Liegt der Median genau auf der Grenze zweier Klassen, kann diese Rechnung nicht durchgeführt werden. In diesem Fall entspricht der Median der Grenze der beiden Klassen.

Weitere Lagekennwerte von Stichproben

Neben dem arithmetischen Mittelwert und dem Median sind zwei weitere Kennwerte gebräuchlich. Bei Wachstumsprozessen wird das geometrische Mittel x_G verwendet, das sich bei einem Stichprobenumfang N aus der N -ten Wurzel des Produktes der einzelnen Stichprobenwerte berechnet.

$$x_G = \sqrt[N]{x_1 \cdot \dots \cdot x_N} = \sqrt[N]{\prod_{i=1}^N x_n} \quad (3.28)$$

Der Modus ist der am häufigsten vorkommende Wert einer Stichprobe, also der Wert mit der größten absoluten oder relativen Häufigkeit. Auf beide Kennwerte wird hier nicht weiter eingegangen.

Zusammenfassung der Lagekennwerte von Stichproben

Zur besseren Übersicht fasst Tabelle 3.9 die Lagekennwerte einer Stichprobe zusammen.

Tabelle 3.9: Lagekennwerte einer Stichprobe

Lagekennwert	Definition	Bemerkungen
Mittelwert	$\bar{x} = \frac{x_1 + \dots + x_N}{N} = \frac{1}{N} \cdot \sum_{n=1}^N x_n$	Empfindlich gegenüber Ausreißern
Mittelwert von Daten in Klassen	$\bar{x} = \sum_{n=1}^N (x_n \cdot h(c_n))$	
Median	$H(x_{MED}) = 0.5$	
Median von Daten in Klassen	$x_{MED} = c_{n-1} + \frac{d \cdot (0.5 - H(c_{n-1}))}{h(c_n)}$	Weniger empfindlich gegenüber Ausreißern als der Mittelwert
Geometrisches Mittel	$x_G = \sqrt[N]{x_1 \cdot \dots \cdot x_N}$	
Modus	häufigster Wert einer Stichprobe	

Da zur Berechnung der Lagekennwerte meist entsprechende Software verwendet wird, entfällt in vielen Fällen eine Einteilung der Daten in Klassen. MATLAB bietet einige Funktionen, mit denen die Lagekennwerte von Stichproben bestimmt werden können. Sie sind in Tabelle 3.10 aufgelistet.

Tabelle 3.10: Berechnung der Lagekennwerte von Stichproben in MATLAB

Lagekennwert	MATLAB-Befehl
Mittelwert	mean(X)
Median	median(x)
Geometrisches Mittel	geomean(X)
Modus	mode(X)

Vergleichbare Befehle existieren in Python, sie sind in Tabelle 3.11 zusammengestellt.

Tabelle 3.11: Berechnung der Lagekennwerte von Stichproben in Python

Lagekennwert	MATLAB-Befehl
Mittelwert	numpy.median
Median	numpy.mean
Geometrisches Mittel	scipy.stats.mstats.gmean
Modus	scipy.stats.mode

3.3.2 Streuungskennwerte einer Stichprobe

Der Mittelwert und der Median sind Kennwerte für die Lage einer Stichprobe. Für eine zusammenfassende Beschreibung ist es notwendig, zusätzlich die Streuung der Stichprobe zu charakterisieren. Dazu können Spannweite, Varianz und Quantile der Verteilung verwendet werden.

Spannweite

Die Spannweite oder Streu- beziehungsweise Variationsbreite einer Stichprobe ergibt sich aus der Differenz von größtem und kleinstem Stichprobenwert.

$$\Delta x = x_{MAX} - x_{MIN} \quad (3.29)$$

Beispiel: Kapazitätsmessung

Für die Messreihen 1 und 2 aus Tabelle 3.7 ergeben sich die beiden Spannweiten zu

$$\Delta C_1 = 103nF - 97nF = 6nF \quad (3.30)$$

beziehungsweise

$$\Delta C_2 = 153nF - 97nF = 56nF \quad (3.31)$$

Die Spannweite reagiert offensichtlich extrem auf Ausreißer und besitzt damit bezüglich der Streuung aller Stichprobenwerte nur eine geringe Aussagekraft.

Varianz und Standardabweichung

Die Varianz betrachtet die Streuung aller Stichprobenwerte um den Mittelwert. Die Summe der einzelnen Abweichungen heben sich auf.

$$\sum_{n=1}^N (x_n - \bar{x}) = \sum_{n=1}^N x_n - N \cdot \bar{x} = 0 \quad (3.32)$$

Der Mittelwert kann deshalb nicht zur Bewertung der Streuung herangezogen werden. Eine bessere Möglichkeit zur Bestimmung der Streuung einer Stichprobe ergibt sich aus der Summe der quadrierten Abweichungen. Durch das Quadrieren werden alle Elemente der Summe positiv und können sich nicht mehr gegenseitig kompensieren. Es ergibt sich die Varianz s^2 , die definiert ist als

$$s^2 = \frac{1}{N-1} \cdot \sum_{n=1}^N (x_n - \bar{x})^2 \quad (3.33)$$

Für eine manuelle Berechnung kann Gleichung (3.33) mit der binomischen Formel umgeformt werden zu

$$\begin{aligned} s^2 &= \frac{1}{N-1} \cdot \sum_{n=1}^N (x_n - \bar{x})^2 = \frac{1}{N-1} \cdot \sum_{n=1}^N (x_n^2 - 2 \cdot x_n \cdot \bar{x} + \bar{x}^2) \\ &= \frac{1}{N-1} \cdot \left(\sum_{n=1}^N x_n^2 - 2 \cdot N \cdot \bar{x}^2 + N \cdot \bar{x}^2 \right) = \frac{1}{N-1} \cdot \left(\sum_{n=1}^N x_n^2 - N \cdot \bar{x}^2 \right) \end{aligned} \quad (3.34)$$

Im Gegensatz zu dem Mittelwert wird bei der Varianz nicht durch die Anzahl N der Stichprobenwerte, sondern durch den Wert $(N - 1)$ geteilt. Mathematisch ergibt sich die Division durch den Wert $(N - 1)$ dadurch, dass bei der Bildung der Varianz die Differenzen zum Mittelwert ausgewertet werden. Der Mittelwert wird bereits mit diesen Daten berechnet, sodass ein Summand in Gleichung (3.32) sich aus den übrigen $(N - 1)$ Werten ergibt. Es sind also nur $(N - 1)$ Summanden unabhängig oder frei. Es geht ein sogenannter Freiheitsgrad verloren, sodass die Division durch den Wert $(N - 1)$ gerechtfertigt ist.

Beispiel: Kapazitätsmessung

Für die Messreihen 1 und 2 aus Tabelle 3.7 ergeben sich die beiden Varianzen zu

$$s_{C1}^2 = 3.33nF^2 \quad (3.35)$$

beziehungsweise

$$s_{C2}^2 = 286.1nF^2 \quad (3.36)$$

Die Varianz ist ein Maß für die Streuung, der physikalische Gehalt der Größe ist aber unanschaulich. Aus diesem Grund wird für die Kennzeichnung einer Streuung oft die Standardabweichung verwendet. Die Standardabweichung s ergibt sich aus der positiven Quadratwurzel der Varianz s^2 und berechnet sich aus

$$s = \sqrt{\frac{1}{N-1} \cdot \sum_{n=1}^N (x_n - \bar{x})^2} \quad (3.37)$$

Die Standardabweichung hat dieselbe physikalische Dimension wie die Stichprobenwerte x und eignet sich damit gut für eine anschauliche Interpretation.

Beispiel: Kapazitätsmessung

Für die Messreihen 1 und 2 aus Tabelle 3.7 ergeben sich die beiden Varianzen zu

$$s_{C1}^2 = 1.825nF \quad (3.38)$$

beziehungsweise

$$s_{C2}^2 = 16.915nF \quad (3.39)$$

Liegt die Stichprobe in Klassen vor, ergibt sich die Varianz aus

$$s^2 = \frac{1}{N-1} \cdot \sum_{n=1}^N (h_A(c_n) \cdot (c_n - \bar{x})^2) \quad (3.40)$$

und die Standardabweichung berechnet sich entsprechend zu

$$s = \sqrt{\frac{1}{N-1} \cdot \sum_{n=1}^N (h_A(c_n) \cdot (c_n - \bar{x})^2)} \quad (3.41)$$

Quantilabstände einer Stichprobe

Auch wenn die Empfindlichkeit der Varianz gegenüber Ausreißern kleiner ist als die der Spannweite einer Stichprobe, reagiert die Standardabweichung auf Ausreißer ähnlich stark wie der arithmetische Mittelwert. Aus diesem Grund werden Streuungskenngrößen definiert, die sich an der Definition des Medians orientieren und als Quantile bezeichnet werden. Das P-Quantil x_P einer Verteilung trennt die Daten so in zwei Teile, dass ein Anteil P unterhalb des Quantils liegt und ein Anteil $(1 - P)$ über dem Quantil liegt.

$$H(x_P) = P \quad (3.42)$$

Der Median entspricht dem 50%-Quantil. Werden die Quantile die Stichprobe in vier Intervalle teilen, werden Sie als Quartile bezeichnet. Die Quartile einer Stichprobe lassen sich ähnlich wie der Median einer Stichprobe bestimmen. Auch die Rechenregeln für in Klassen eingeteilte Daten mit konstanter Klassenbreite d gelten sinngemäß, sodass die Quartile bestimmt werden durch

$$x_{0.25} = c_{n-1} + \frac{d \cdot (0.25 - H(c_{n-1}))}{h(c_n)} \quad (3.43)$$

$$x_{0.75} = c_{n-1} + \frac{d \cdot (0.75 - H(c_{n-1}))}{h(c_n)} \quad (3.44)$$

Darüber hinaus ist eine grafische Bestimmung der Quartile aus der Summenhäufigkeit möglich. Der Abstand zwischen dem 75%- und dem 25%-Quartil wird als Interquartilabstand (inter quartile range) IQR bezeichnet.

$$IQR = x_{0.75} - x_{0.25} \quad (3.45)$$

Der Interquartilabstand ist unempfindlich gegen Ausreißer, weil er von der absoluten Lage der Stichprobenwerte, die am Rand der Verteilung liegen, unabhängig ist.

Beispiel: Widerstandsmessung

Die Berechnung wird am Beispiel der Widerstandsmessung veranschaulicht. Bild 3.9 zeigt die Summenhäufigkeit der Stichprobe aus Tabelle 3.1 sowie die 25%--, 50%- und 75%-Quantile der Stichprobe.

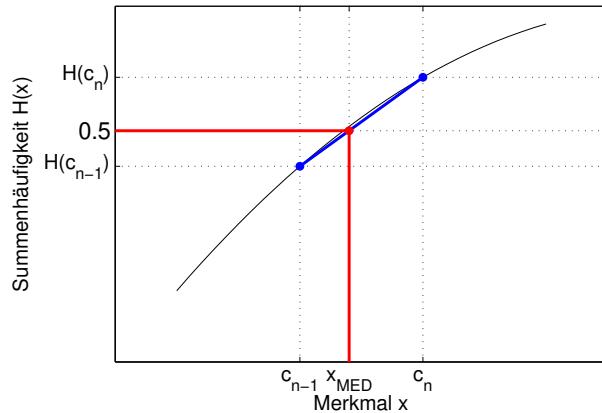


Bild 3.9: Darstellung der Summenhäufigkeit für die Stichprobe in Tabelle 3.3 mit Quartilen

Für die Daten in Tabelle 3.3 ergeben sich das 25%- und 75%-Quartil zu

$$R_{0.25} = 985\Omega + \frac{1\Omega \cdot (0.25 - 0.23)}{0.32} = 985.06\Omega \quad (3.46)$$

$$R_{0.75} = 987\Omega + \frac{1\Omega \cdot (0.75 - 0.72)}{0.15} = 987.20\Omega \quad (3.47)$$

Damit errechnet sich der Inter-Quartil-Range zu

$$IQR = R_{0.75} - R_{0.25} = 987.20\Omega - 985.06\Omega = 2.14\Omega \quad (3.48)$$

Zusammenfassung der Streuungskennwerte von Stichproben

Zur besseren Übersicht fasst Tabelle 3.12 die Streuungskennwerte einer Stichprobe zusammen.

Tabelle 3.12: Berechnung der Lagekennwerte von Stichproben in Python

Streuungskennwert	Definition	Bemerkungen
Spannweite	$\Delta x = x_{MAX} - x_{MIN}$	Extrem empfindlich gegenüber Ausreißern
Varianz	$s^2 = \frac{1}{N-1} \cdot \sum_{n=1}^N (x_n - \bar{x})^2$	Empfindlich gegenüber Ausreißern
Varianz in Klassen eingeteilter Daten	$s^2 = \frac{1}{N-1} \cdot \sum_{n=1}^N (h_A(c_n) \cdot (c_n - \bar{x})^2)$	
Standardabweichung	$s = \sqrt{\frac{1}{N-1} \cdot \sum_{n=1}^N (x_n - \bar{x})^2}$	Vergleichbar zur Varianz, aber in Einheiten der Zielgröße x
Standardabweichung in Klassen eingeteilter Daten	$s = \sqrt{\frac{1}{N-1} \cdot \sum_{n=1}^N (h_A(c_n) \cdot (c_n - \bar{x})^2)}$	
P-Quantil	$H(x_P) = P$	
P-Quantil in Klassen eingeteilter Daten	$x_P = c_{n-1} + \frac{d \cdot (P - H(c_{n-1}))}{h(c_n)}$	
Interquartilabstand	$IQR = x_{0.75} - x_{0.25}$	Robuster Streuungskennwert

Gerade bei größeren Stichprobenumfängen ist es aufwendig, die Streuungskennwerte manuell zu berechnen. Daher wird zur Auswertung meist entsprechende Software herangezogen. Die Befehle zur Berechnung der Streuungskenngrößen mit MATLAB sind in Tabelle 3.13 zusammengefasst.

Tabelle 3.13: Berechnung der Streuungskennwerte von Stichproben in MATLAB

Streuungskennwert	MATLAB-Befehl
Spannweite	range(x) oder max(x)-min(x)
Varianz	var(x)
Standardabweichung	std(x)
p-Quantil	quantile(x,p)
Interquartilabstand	iqr(x)

Vergleichbare Befehle existieren in Python, sie sind in Tabelle 3.14 zusammengestellt.

Tabelle 3.14: Berechnung der Streuungskennwerte von Stichproben in Python

Streuungskennwert	Python-Befehl
Spannweite	<code>max - min</code>
Varianz	<code>numpy.var</code>
Standardabweichung	<code>numpy.std</code>
p-Quantil	<code>numpy.quantile</code>
Interquartilabstand	<code>numpy.quantile</code>

3.3.3 Schiefe oder Symmetrie einer Stichprobe

Neben der Lage und Streuung einer Verteilung kann die Schiefe einer Verteilung angegeben werden. Die Schiefe ist ein Maß für die Symmetrie der Verteilung zum Mittelwert.

Definition von Schiefe und Symmetrie einer Stichprobe

Eine Stichprobe wird als symmetrisch bezeichnet, wenn es eine Symmetrieeachse gibt, zu der die rechte und linke Seite der Verteilung annähernd zueinander spiegelbildlich sind. Exakte Symmetrie ist bei empirischen Verteilungen selten gegeben. Ein Beispiel für eine symmetrische Verteilung ist in Bild 3.10 Mitte gegeben. Eine unsymmetrische Verteilung wird als schiefe Verteilung bezeichnet. Eine Verteilung heißt rechtsschief, wenn der überwiegende Teil der Daten linksseitig konzentriert ist. Dann fällt die Verteilung wie die linke Grafik in Bild 3.10 nach links deutlich steiler ab als nach rechts. Entsprechend heißt eine Verteilung linksschief, wenn wie in der Verteilung der rechten Grafik in Bild 3.10 der überwiegende Teil der Daten rechtsseitig konzentriert ist und die Verteilung nach rechts deutlich steiler abfällt.

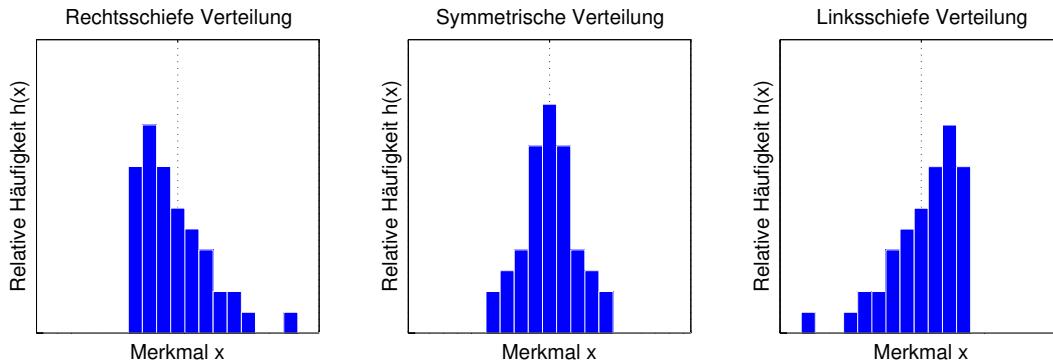


Bild 3.10: Darstellung von Stichprobenverteilungen mit gleichem Mittelwert und unterschiedlicher Schiefe

Die Schiefe kann mit zwei unterschiedlichen Kenngrößen charakterisiert werden, dem Momentenkoef- fizient und dem Quantilkoeffizient der Schiefe.

Momentenkoeffizient der Schiefe

In Analogie an die Varianz s^2 wird ein Momentenkoeffizient der Schiefe definiert.

$$g_M = \frac{N}{(N-1) \cdot (N-2)} \cdot \sum_{n=1}^N \left(\frac{x_n - \bar{x}}{s} \right)^3 \quad (3.49)$$

Durch die dritte Potenz der Abweichung der Stichprobenwerte vom Mittelwert bleiben im Vergleich zur Standardabweichung die Vorzeichen bei den Abweichungen erhalten. Bei rechtsschiefen Verteilungen überwiegen positive Abweichungen, sodass der Momentenkoeffizient g_M positiv wird. Bei linksschiefen Verteilungen weist er entsprechend negative Werte auf. Wegen der Division durch die dritte Potenz der Standardabweichung ist der Momentenkoeffizient der Schiefe dimensionslos und unabhängig von der Messgröße. Für den Fall von in Klassen aufgeteilten Daten errechnet sich der Momentenkoeffizient der Schiefe aus

$$g_M = \frac{\frac{1}{N} \cdot \sum_{n=1}^N (h_a(c_n) \cdot (c_n - \bar{x})^3)}{s^3} \quad (3.50)$$

Positive Werte für den Momentenkoeffizienten der Schiefe weisen auf rechtsschiefe Verteilungen, negative Werte weisen auf linksschiefe Verteilungen hin. Ist der Momentenkoeffizient der Schiefe nahe null, handelt es sich um eine symmetrische Verteilung.

Tabelle 3.15: Bewertung des Momentenkoeffizienten der Schiefe g_M

Kennwert	Symmetrieeigenschaft
$g_M > 0$	Rechtsschiefe Verteilung
$g_M = 0$	Symmetrische Verteilung
$g_M < 0$	Linksschiefe Verteilung

Nachteil des Momentenkoeffizienten der Schiefe ist die durch seine Definition gegebene Abhängigkeit gegenüber Ausreißern. Der Befehl zur Berechnung der Schiefe in MATLAB ist in Tabelle 3.16 aufgeführt.

Tabelle 3.16: Berechnung der Symmetriekennwerte von Stichproben in MATLAB

Symmetriekennwert	MATLAB-Befehl
Momentenkoeffizient der Schiefe	Rechtsschiefe skewness(x)

Tabelle 3.17 zeigt den Befehl zur Berechnung der Schiefe in Python.

Tabelle 3.17: Berechnung der Symmetriekennwerte von Stichproben in Python

Symmetriekennwert	Python-Befehl
Momentenkoeffizient der Schiefe	scipy.stats.skew

Quantilkoeffizient der Schiefe

Die Symmetrie oder Schiefe einer Verteilung kann über eine Kenngröße bewertet werden, die die Symmetrie der Quantile einer Stichprobe bewertet. Dazu wird der Quantilkoeffizient der Schiefe berechnet aus

$$g_P = \frac{(x_{1-P} - x_{MED}) - (x_{MED} - x_P)}{x_{1-P} - x_P} \quad (3.51)$$

Der Quantilkoeffizient mit $P = 25\%$ wird als Quartilkoeffizient der Schiefe bezeichnet. Er ergibt sich zu

$$g_{0.25} = \frac{(x_{0.75} - x_{MED}) - (x_{MED} - x_{0.25})}{x_{0.75} - x_{0.25}} \quad (3.52)$$

Die Quartilkoeffizienten bewerten im Zähler den Unterschied zwischen der Entfernung des 25%- beziehungsweise 75%-Quartils zum Median. Bei symmetrischen Verteilungen ist der Abstand gleich groß, der Unterschied ist null. Damit gilt für symmetrische Verteilungen

$$g_{0.25} = 0 \quad (3.53)$$

Mit steigender Asymmetrie steigt der Betrag des Quartilkoeffizienten. Für die Bewertung gilt die gleiche Klassifizierung wie bei dem Momentenkoeffizient der Schiefe. Positive Quartilkoeffizienten weisen auf eine rechtsschiefe, negative Quartilkoeffizienten weisen auf eine linksschiefe Verteilung hin. Durch den Nenner wird der Quartilkoeffizient so normiert, dass er nur Zahlenwerte im Bereich $-1 \leq g_P \leq 1$ annehmen kann. Der Quartilkoeffizient reagiert weniger sensitiv auf Ausreißer, da dessen Definition mithilfe der Quartile erfolgt. Für das Beispiel aus Bild 3.10 ergeben sich die in Tabelle 3.18 dargestellten Koeffizienten.

Tabelle 3.18: Charakterisierung der Schiefe für die Stichprobe aus Bild 3.10

Bild 3.10	Momentenkoeffizient der Schiefe g_M	Quartilkoeffizient der Schiefe g_P	Folgerung für die Verteilung
Links	1.1017	0.3333	Rechtsschief
Mitte	0	0	Symmetrisch
Rechts	-1.1017	-0.3333	Linksschief

Beide Maße für die Schiefe weisen dasselbe Vorzeichen, aber unterschiedliche Zahlenwerte auf und sind deshalb nicht direkt miteinander vergleichbar.

3.3.4 Lageregeln zur Interpretation der Symmetrie einer Stichprobe

Die Symmetrieeigenschaften der Häufigkeitsverteilung einer Stichprobe können auch an der Lage von Median und Mittelwert abgelesen werden. Die Verteilung der Stichprobe in Bild 3.10 links fällt nach links steil ab und läuft nach rechts flach aus, sie ist also rechtsschief. Mittelwert und Median berechnen sich zu

$$\bar{x} = 10 \quad (3.54)$$

und

$$x_{MED} = 9.26 \quad (3.55)$$

Für die mittlere Häufigkeitsverteilung stimmen Mittelwert und Median überein.

$$\bar{x} = x_{MED} = 10 \quad (3.56)$$

Die rechte Häufigkeitsverteilung ist linksschief, für sie errechnen sich die Lagekennwerte zu

$$\bar{x} = 10 \quad (3.57)$$

und

$$x_{MED} = 10.74 \quad (3.58)$$

Dieses Ergebnis kann verallgemeinert werden. Tabelle 3.19 fasst die Lageregeln zur Beschreibung der Symmetrie einer Häufigkeitsverteilung zusammen. Je stärker sich die Lagekennwerte voneinander unterscheiden, desto schiefer ist die Verteilung.

Tabelle 3.19: Lageregeln von Median und arithmetischem Mittelwert zur Beschreibung der Symmetri

Lagekennwerte	Symmetrieeigenschaft
$\bar{x} > x_{MED}$	Rechtsschiefe Verteilung
$\bar{x} < x_{MED}$	Symmetrische Verteilung
$\bar{x} < x_{MED}$	Linksschiefe Verteilung

Beispiel: Dicke einer Schutzlackbeschichtung

Als Beispiel für eine schiefe Verteilung soll die Dicke einer Schutzlackbeschichtung betrachtet werden, die von einer Automatisierungseinrichtung auf Platinen aufgebracht wird. Der Sollwert der Schichtdicke ist spezifiziert auf 10 µm. Bei 100 Platinen wurde die aufgetragene Schutzschicht vermessen. Dabei ergab sich die in Tabelle 3.20 aufgelistete Häufigkeitsverteilung.

Tabelle 3.20: Häufigkeitsverteilung der Stichprobe

D / µm	Anzahl Häufigkeit h _A (D)	Relative Häufigkeit(D)	D / µm	Anzahl Häufigkeit h _A (D)	Relative Häufigkeit(D)
4	13	0.13	24	5	0.05
8	33	0.33	28	2	0.02
12	21	0.21	32	2	0.02
16	16	0.16	36	0	0.00
20	7	0.07	40	1	0.01

In Bild 3.11 ist zu erkennen, dass die Verteilung nach links wesentlich steiler abfällt als nach rechts. Es handelt sich daher um eine rechtsschiefe Verteilung.

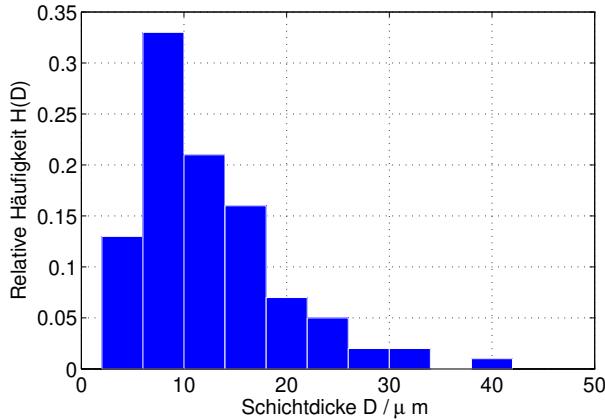


Bild 3.11: Relative Häufigkeitsverteilung der Lackdicke

Der Momentenkoeffizient der Schiefe berechnet sich für das Beispiel der Schichtdicken mit der Standardabweichung

$$s = \sqrt{\frac{1}{N-1} \cdot \sum_{n=1}^N (h_A(c_n) \cdot (c_n - \bar{D})^2)} = 7.007 \mu m \quad (3.59)$$

zu

$$g_M = \frac{\frac{1}{N} \cdot \sum_{n=1}^N (h_A(c_n) \cdot (c_n - \bar{D})^3)}{s^3} = 1.3 \quad (3.60)$$

Wegen des positiven Vorzeichens des Momentenkoeffizienten handelt es sich um eine rechtsschiefe Verteilung, was die Einschätzung anhand Bild 3.11 bestätigt. Für die Berechnung des 25%-Quartilkoeffizienten der Schiefe werden die Quartile der Verteilung benötigt. Sie berechnen sich aus den angegebenen Klassen zu

$$D_{0.25} = c_{n-1} + \frac{d \cdot (0.25 - H(c_{n-1}))}{h(c_n)} = 4 \mu m + \frac{4 \mu m \cdot (0.25 - 0.13)}{0.33} = 5.455 \mu m \quad (3.61)$$

$$D_{0.5} = D_{MED} = c_{n-1} + \frac{d \cdot (0.5 - H(c_{n-1}))}{h(c_n)} = 8 \mu m + \frac{4 \mu m \cdot (0.5 - 0.46)}{0.21} = 8.763 \mu m \quad (3.62)$$

und

$$D_{0.75} = c_{n-1} + \frac{d \cdot (0.75 - H(c_{n-1}))}{h(c_n)} = 12 \mu m + \frac{4 \mu m \cdot (0.75 - 0.67)}{0.16} = 14 \mu m \quad (3.63)$$

Der 25%-Quartilkoeffizient der Schiefe folgt damit nach Gleichung (3.61) zu

$$g_{0.25} = \frac{(D_{0.75} - D_{0.5}) - (D_{0.5} - D_{0.25})}{D_{0.75} - D_{0.25}} = \frac{(14 - 8.763) - (8.763 - 5.455)}{14 - 5.455} = 0.2259 \quad (3.64)$$

Das positive Vorzeichen des 25%-Quartilkoeffizienten weist auf eine rechtsschiefe Verteilung hin und bestätigt damit das Ergebnis des Momentenkoeffizienten der Schiefe und der grafischen Beurteilung. Zusätzlich kann die Lage des arithmetischen Mittelwertes

$$\bar{D} = \sum_{n=1}^N (D_n \cdot h(c_n)) = 12.44 \mu m \quad (3.65)$$

in Relation zum Median zur Bewertung der Schiefe nach Tabelle 3.19 ausgewertet werden. Auch nach diesem Kriterium ist die Stichprobe rechtsschief. Alle Bewertungsmöglichkeiten der Schiefe führen damit zum gleichen Ergebnis.

Schiefe oder asymmetrische Häufigkeitsverteilungen treten insbesondere dann auf, wenn das untersuchte Merkmal durch einen natürlichen einseitigen Grenzwert eingeschränkt wird. Im vorigen Beispiel der Dicke einer Schutzlackschicht wird die Häufigkeitsverteilung nach links begrenzt, da die Schichtdicke nie kleiner als null werden kann. Weitere Beispiele wären die Rauheit einer Oberfläche oder der Durchmesser einer Bohrung, der durch den Durchmesser des verwendeten Bohrers nach unten begrenzt ist. In Kapitel 4 wird sich zeigen, dass asymmetrische Verteilungen auch zur Abschätzung von Lebensdauern oder Ausfallzeiten herangezogen werden.

3.3.5 Box-Plot

In den Abschnitten 3.3.1 und 3.3.2 werden Kenngrößen für die numerische Beschreibung von Stichproben und die Charakterisierung von Verteilungen vorgestellt. Die wesentlichen Größen können aus dem sogenannten Box-Plot abgelesen werden, der im Folgenden vorgestellt wird. Der Box-Plot fasst fünf charakteristische Punkte einer Verteilung zusammen:

- Minimaler Stichprobenwert x_{MIN}
- 25%-Quartil $x_{0.25}$
- Median x_{MED}
- 75%-Quartil $x_{0.75}$
- Maximaler Stichprobenwert x_{MAX}

Die Idee des Box-Plots ist in Bild 3.12 dargestellt. Anfang und Ende der Box stellen die 25%- und 75%-Quartile dar. Die Länge der Box repräsentiert damit den Interquartilabstand. Der Median wird als Balken in der Box eingezeichnet. Zwei Linien außerhalb der Box, die sogenannten Whisker, zeigen die minimalen und maximalen Werte x_{MIN} und x_{MAX} der Stichprobe. Ausreißer werden nicht zur Bestimmung des minimalen und maximalen Wertes verwendet. Als Ausreißer gelten Werte, die erheblich kleiner als das 25%-Quartil und erheblich größer als das 75%-Quartil sind. Mathematisch wird diese Aussage durch die Bedingungen

$$x_{OUT} < x_{0.25} - 1.5 \cdot (x_{0.75} - x_{0.25}) \quad (3.66)$$

beziehungsweise

$$x_{OUT} > x_{0.75} + 1.5 \cdot (x_{0.75} - x_{0.25}) \quad (3.67)$$

formuliert. Ausreißer werden als separates Kreuz dargestellt.

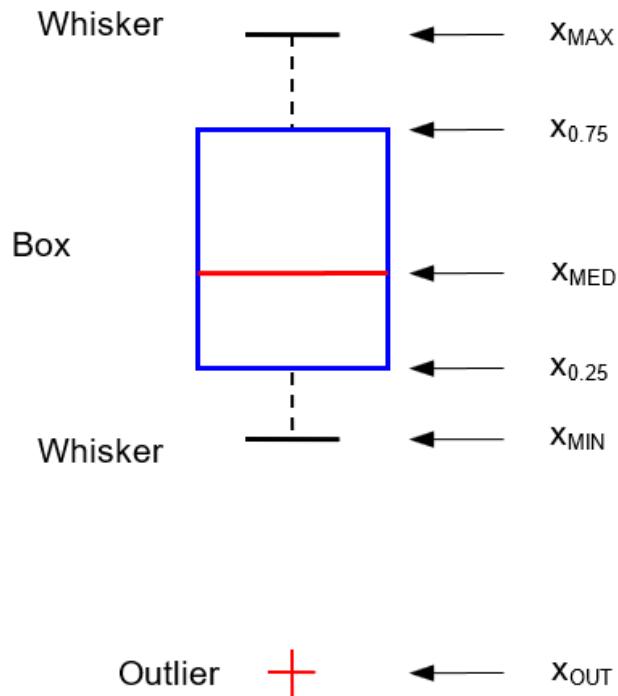


Bild 3.12: Grundidee des Box-Plots

Neben der grafischen Darstellung eignet sich der Box-Plot für eine Interpretation der Stichprobenkennwerte.

Bild 3.13 stellt als Beispiel den Box-Plot für die Stichproben aus Bild 3.10 dar. Die unsymmetrische Lage des Median in der Box weist auf eine unsymmetrische Stichprobe hin.

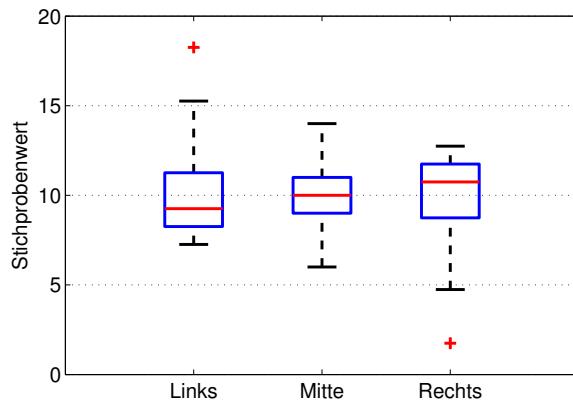


Bild 3.13: Box-Plot für die Stichproben aus 3.10

MATLAB und python Python bieten zur Erstellung eines Box-Plots eine separate Funktion an.

Tabelle 3.21: Darstellung des Box-Plots einer Stichprobe in MATLAB

Darstellung	MATLAB-Befehl
Box-Plot	<code>boxplot(x)</code>

Tabelle 3.22: Darstellung des Box-Plots einer Stichprobe in Python

Darstellung	MATLAB-Befehl
Box-Plot	<code>matplotlib.pyplot.boxplot</code>

3.4 Anwendungsbeispiel: Charakterisierung eines Klebeprozesses

Eine Automatisierungseinrichtung hat zum Befestigen einer Folie im Automobilbau Klebemengen dosiert. Die in Tabelle 3.23 Tabelle 3.23 enthaltenen Zahlenwerte stellen eine Stichprobe mit einem Umfang von $N = 40$ abgefüllten Klebemengen dar.

Tabelle 3.23: Stichprobenwerte zur Bewertung eines Klebeprozesses

Messung	Messwerte Klebemenge m / mg				
1 - 5	50.18	51.85	51.09	50.09	51.03
6 - 10	50.69	51.76	51.23	51.49	51.62
11- 15	50.52	51.33	51.18	51.76	52.62
16 - 20	51.49	51.19	51.28	50.82	50.01
21 - 25	51.6	50.94	51.54	51.3	50.91
26 - 30	51.44	51.78	51.37	51.36	50.54
31 - 35	52.2	52.12	49.4	49.76	51.54
36 - 40	50.64	50.37	50.16	50.61	50.38

Der Datensatz wird im Folgenden ohne Aufteilung der Daten in Klassen und mit Aufteilung der Daten in Klassen ausgewertet. Abschließend werden die Ergebnisse miteinander verglichen.

3.4.1 Datenanalyse in MATLAB ohne Aufteilung der Daten in Klassen

Bei dem untersuchten Klebeprozess handelt es sich um einen kontinuierlichen Prozess, jeder Messwert besitzt hierbei die absolute Häufigkeit 1. Eine Darstellung als Häufigkeitsverteilung lässt somit keine Rückschlüsse auf den zu untersuchenden Prozess zu. Deshalb werden die Daten mit einem Streudiagramm und der relativen Summenhäufigkeit beschrieben.

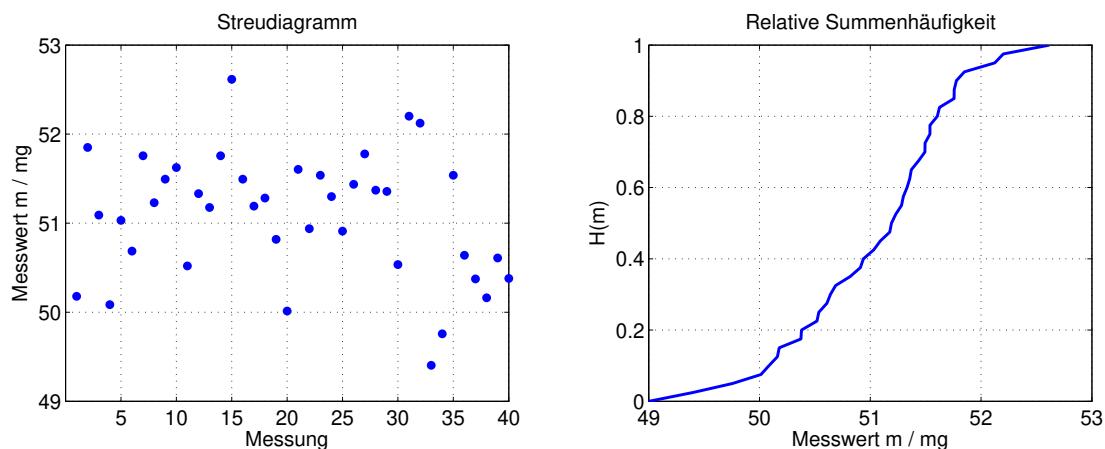


Bild 3.14: Darstellung der Messwerte als Streudiagramm und als relative Summenhäufigkeit

Im Streudiagramm in Bild 3.14 ist zu erkennen, dass sich die Messwerte in einem Bereich von 49 bis 53 mg befinden. In der Darstellung der relativen Summenhäufigkeit kann abgelesen werden, dass der Median mit einer relativen Summenhäufigkeit von 0.5 bei einer Klebemenge von 51.25 mg liegt. Die Grafik wurde mit dem folgenden MATLAB-Code erzeugt.

```

1 % Einlesen der Messdaten
2 load Klebermenge.mat
3
4 % Grafische Darstellung
5 f = figure(1);
6
7 % Messwerte als Streudiagramm
8 subplot(1,2,1);
9 scatter(1:length(Klebermenge),Klebermenge,'bo','filled');
10
11 % Relative Summenhäufigkeit
12 subplot(1,2,2);
13 plot(sort(Klebermenge),(1:length(Klebermenge))/length(Klebermenge));

```

Um die Lage der Messwerte genauer zu beschreiben, werden die in Abschnitt 3.3.1 eingeführten Lagekennwerte berechnet. Für die Messwerte aus Tabelle 3.23 ergibt sich der arithmetische Mittelwert zu

$$\bar{m} = \frac{1}{N} \cdot \sum_{n=1}^N m_n = 51.08 \text{mg} \quad (3.68)$$

Der Median wird nach Sortieren der Messwerte aus dem Mittelwert der beiden mittleren Messwerte bestimmt.

$$m_{MED} = \frac{m_{17} + m_8}{2} = \frac{51.19 \text{mg} + 51.23 \text{mg}}{2} = 51.21 \text{mg} \quad (3.69)$$

Zusätzlich zu den Kennwerten zur Beschreibung der Lage werden in Abschnitt 3.3.2 Streuungskennwerte definiert. Die Varianz der Messwerte berechnet sich zu

$$s^2 = \frac{1}{N-1} \cdot \sum_{n=1}^N (m_n - \bar{m})^2 = 0.49 \text{mg}^2 \quad (3.70)$$

Aus der Wurzel der Varianz folgt die Standardabweichung von

$$s = \sqrt{s^2} = 0.70 \text{mg} \quad (3.71)$$

Der Interquartilabstand ergibt sich aus der Differenz des 75%-Quartils und des 25%-Quartils der Messwerte zu

$$IQR = m_{0.75} - m_{0.25} = 50.57 \text{mg} - 51.54 \text{mg} = 0.97 \text{mg} \quad (3.72)$$

Der verwendete Programmcode zur Berechnung der Lage- und Streuungskennwerte mithilfe von MATLAB ist im Folgenden dargestellt.

```

1 % Berechnung der Lagekennwerte
2 average = mean(Klebermenge)
3 median = quantile(Klebermenge , 0.5)
4
5 % Berechnung der Streuungskennwerte
6 variance = var(Klebermenge)
7 standardDeviation = std(Klebermenge)
8 interQuartilRange = iqr(Klebermenge)
```

Da aus Bild 3.15 keine Aussage hinsichtlich der Schiefe der Verteilung gemacht werden kann, wird der 25%-Quartilkoeffizient und der Momentenkoeffizient der Schiefe zur Bewertung herangezogen. Der 25%-Quartilkoeffizient der Schiefe berechnet sich zu

$$g_{0.25} = \frac{(m_{0.75} - m_{MED}) - (m_{MED} - m_{0.25})}{m_{0.75} - m_{0.25}} = \frac{(51.54 - 51.21) - (51.21 - 50.57)}{51.54 - 50.57} = -0.32 \quad (3.73)$$

und der Momentenkoeffizient der Schiefe folgt zu

$$g_M = \frac{\frac{1}{N} \cdot \sum_{n=1}^N (m_n - \bar{m})^3}{s^3} = -0.28 \quad (3.74)$$

Beide Werte weisen auf eine linksschiefe Verteilung hin. Die Berechnung der Kennwerte der Schiefe wird mit MATLAB mit der folgenden Befehlssequenz durchgeführt

```

1 % Berechnung des 25%-Quartilkoeffizienten der Schiefe
2 q = quantile(Klebermenge ,[0.25 0.5 0.75]);
3 g25 = ((q(3)-q(2))-(q(2)-q(1)))/(q(3)-q(1))
4
5 % Berechnung des Momentenkoeffizienten der Schiefe
6 gm = skewness(Klebermenge)
```

Der Klebeprozess wird mit den Stichprobenwerten aus Tabelle 3.23 sowohl grafisch dargestellt als auch durch Kennwerte beschrieben. Dabei werden die Daten nicht in Klassen eingeteilt. Um den Unterschied zu zeigen, wird auf Basis der gleichen Stichprobenwerte nun eine Datenanalyse durchgeführt, bei der die Messwerte in Klassen eingeteilt werden.

3.4.2 Datenanalyse in MATLAB mit Aufteilung der Daten in Klassen

Zunächst muss für die Daten eine sinnvolle Klasseneinteilung gefunden werden. Hierbei wird insbesondere darauf geachtet, dass die Klassenmitten möglichst Zahlen mit wenig Nachkommastellen sind. Bei dem Datensatz aus Tabelle 3.23 mit einem Minimalwert von $m_{MIN} = 49.40$ mg und einem Maximalwert von $m_{MAX} = 52.62$ mg bietet es sich an, eine Klassenbreite d von 1 mg zu wählen. Die Messwerte lassen sich damit in 5 Klassen einteilen, die in Tabelle 3.24 zusammen mit ihrer absoluten und ihrer relativen Häufigkeit angegeben sind.

Tabelle 3.24: Stichprobenwerte zur Überprüfung eines Klebeprozesses eingeteilt in Klassen

c / mg	Anzahl Häufigkeit $h_A(c)$	Relative Häufigkeit $h(c)$	Absolute Summenhäufigkeit $H_A(c)$	Relative Summenhäufigkeit $H(c)$
49	1	0.025	1	0.025
50	7	0.175	8	0.2
51	21	0.525	29	0.725
52	10	0.25	39	0.975
53	1	0.025	40	1

Die relative Häufigkeit und die relative Summenhäufigkeit der in Klassen eingeteilten Stichprobenwerte sind in Bild 3.15 dargestellt.

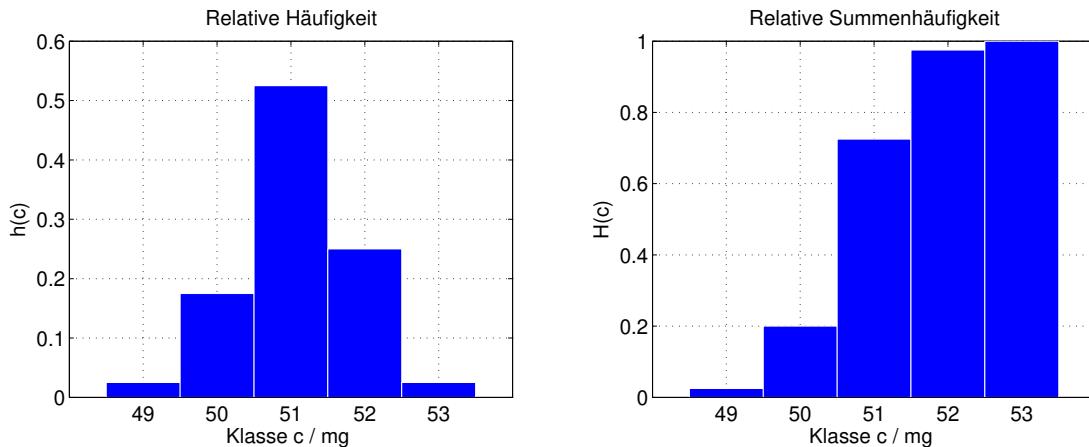


Bild 3.15: Darstellung der relativen Häufigkeit und der relativen Summenhäufigkeit

Die Einteilung der Messwerte in Klassen erfolgt durch MATLAB. Die Berechnung der Häufigkeiten und die Darstellung in Bild 3.15 werden mit der folgenden MATLAB-Befehlssequenz erstellt.

```

1 % Klasseneinteilung
2 class = 49:1:53
3
4 % Berechnung der Häufigkeiten
5 absFreq = hist(Klebermenge, class);
6 absSumFreq = cumsum(absFreq);
7 relFreq = absFreq ./ length(Klebermenge);
8 relSumFreq = cumsum(relFreq);
9
10 % Grafische Darstellung
11 f = figure(2);
12
13 % Relative Häufigkeit
14 subplot(1,2,1)
15 bar(class, relFreq, 'b');
16
17 % Relative Summenhäufigkeit
18 subplot(1,2,2)
19 bar(class, relSumFreq, 'b');
```

Zur numerischen Beschreibung der Häufigkeitsverteilung werden analog zur Datenanalyse ohne Klasseneinteilung die Lage- und Streuungskennwerte berechnet. Hierbei muss die Klasseneinteilung der Messwerte berücksichtigt werden. Damit berechnet sich der arithmetische Mittelwert zu

$$\bar{m} = \sum_{n=1}^N (c_n \cdot h(c_n)) = 51.08 \text{mg} \quad (3.75)$$

Der Median folgt nach den Darstellungen in Abschnitt 3.3.1 zu

$$m_{MED} = c_{n-1} + \frac{d \cdot (0.5 - H(c_{n-1}))}{h(c_n)} = 50 \text{mg} + \frac{1 \text{mg} \cdot (0.5 - 0.2)}{0.525} = 50.57 \text{mg} \quad (3.76)$$

Die Berechnung der Streuungskennwerte führen zu einem Wert für die Varianz von

$$s^2 = \frac{1}{N-1} \cdot \sum_{n=1}^N (h_a(c_n) \cdot (c_n - \bar{m})^2) = 6.19 \text{mg}^2 \quad (3.77)$$

und für die Standardabweichung von

$$s = \sqrt{s^2} = 2.49 \text{mg} \quad (3.78)$$

Der Interquartilabstand berechnet sich aus der Differenz aus dem 75%-Quartil

$$m_{0.75} = c_{n-1} + \frac{d \cdot (0.75 - H(c_{n-1}))}{h(c_n)} = 51 \text{mg} + \frac{1 \text{mg} \cdot (0.75 - 0.725)}{0.25} = 51.10 \text{mg} \quad (3.79)$$

und dem 25%-Quartil

$$m_{0.25} = c_{n-1} + \frac{d \cdot (0.25 - H(c_{n-1}))}{h(c_n)} = 50 \text{mg} + \frac{1 \text{mg} \cdot (0.25 - 0.2)}{0.525} = 50.0952 \text{mg} \quad (3.80)$$

zu

$$IQR = m_{0.75} - m_{0.25} = 51.10 \text{mg} - 50.0952 \text{mg} = 1.00 \text{mg} \quad (3.81)$$

Zur numerischen Auswertung der Daten wird die folgende Befehlssequenz verwendet.

```

1 % Berechnung der Lagekennwerte
2 average = sum(class.*relFreq)
3 median = class(2)+1*(0.5-relSumFreq(2))/relFreq(3)
4
5 % Berechnung der Streuungskennwerte
6 variance = 1/(length(class)-1)*sum(absFreq.*(class-average).^2)
7 standardDeviation = sqrt(variance)
8 quartil75 = class(3)+1*(0.75 - relSumFreq(3))/relFreq(4)
9 quartil25 = class(2)+1*(0.25 - relSumFreq(2))/relFreq(3)
10 interQuartilRange = quartil75-quartil25

```

Abschließend wird die Schiefe der Häufigkeitsverteilung beurteilt. Bereits in Bild 3.15 ist ersichtlich, dass die Verteilung der relativen Häufigkeit nahezu symmetrisch zu der Klasse von 51 mg ist. Es ist daher zu erwarten, dass sowohl der 25%-Quartilkoeffizient als auch der Momentenkoeffizient der Schiefe einen Wert nahe 0 annehmen. Die Berechnung zeigt, dass sowohl der 25%-Quartilkoeffizient der Schiefe mit

$$g_{0.25} = \frac{(m_{0.75} - m_{med}) - (m_{med} - m_{0.25})}{m_{0.75} - m_{0.25}} = \frac{(51.10 - 50.57) - (50.57 - 50.0952)}{51.10 - 50.0952} = 0.05 \quad (3.82)$$

als auch der Momentenkoeffizient der Schiefe mit

$$g_M = \frac{\frac{1}{N} \cdot \sum_{n=1}^N (h_a(c_n) \cdot (c_n - \bar{m})^3)}{s^3} = -0.03 \quad (3.83)$$

diese Aussage bekräftigt. Die Berechnung mit MATLAB ergibt sich aus folgender Befehlssequenz.

```

1 % Berechnung des 25%-Quartilkoeffizienten der Schiefe
2 quartil75 = class(3)+1*(0.75 - relSumFreq(3))/relFreq(4)
3 quartil50 = class(2)+1*(0.5 - relSumFreq(2))/relFreq(3)
4 quartil25 = class(2)+1*(0.25 - relSumFreq(2))/relFreq(3)
5
6 g25 = ((quartil75 - quartil50) - (quartil50 - quartil25)) / (quartil75 - quartil25)
7
8 % Berechnung des Momentenkoeffizienten der Schiefe
9 gm = 1/length(class)*sum(absFreq.* (class-average).^3)/(standardDeviation^3)

```

3.4.3 Vergleich der beiden Datenanalysen

Die Daten aus Tabelle 3.23 werden ohne eine Einteilung in Klassen und mit einer Einteilung in fünf Klassen bewertet. Hierzu werden die eingeführten Lage- und Streuungskennwerte berechnet und die Größen zur Bewertung der Schiefe bestimmt. Dabei weichen die Ergebnisse zwischen den beiden Analysemethoden voneinander ab. Um den Unterschied grafisch zu verdeutlichen, werden die Kennwerte der Datenanalyse mit und ohne eine Einteilung in Klassen als Box-Plot in 3.16 zusammengefasst und gegenübergestellt.

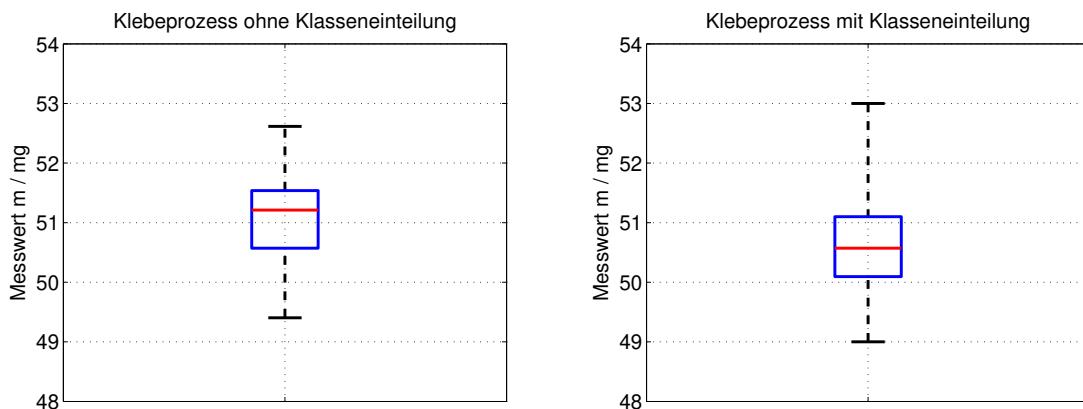


Bild 3.16: Gegenüberstellung der beiden Datenanalysemethoden

Anhand des blauen Rechtecks in der Grafik des Box-Plots kann abgelesen werden, in welchem Bereich 50 % der Stichprobenwerte liegen, die rote Linie kennzeichnet den Median der Verteilung. In 3.16 ist gut zu erkennen, dass die Auswertungen mit und ohne Klassenbildung stark voneinander abweichen. Während der Median für den Klebeprozess ohne Klassenbildung bei $m_{MED} = 51.21 \text{ mg}$ liegt, liegt er nach der Einteilung in Klassen bei $m_{MED} = 50.57 \text{ mg}$, der Wert differiert um 0.7 mg. Außerdem ist zu erkennen, dass die Spannweite der Verteilung bei einer Betrachtung mit einer Einteilung in Klassen bei der vorliegenden Wahl der Klassenmitten größer ist als bei einer Betrachtung ohne eine Einteilung in Klassen.

Die unterschiedlichen Ergebnisse erklären sich dadurch, dass mit der Klassenbildung die einzelnen Werte nicht mehr erfasst werden und dadurch Informationen über die Verteilung verloren gehen. Die Genauigkeit der Auswertung ist von der Klassenanzahl abhängig. Je weniger Klassen aus der Urliste gebildet werden, desto mehr Informationen gehen verloren und desto stärker können die berechneten Kenngrößen von den wahren Kenngrößen der zu Grunde liegenden Verteilung abweichen. Aus diesem Grund sollte nach Möglichkeit stets mit der Urliste gearbeitet werden.

Der verwendete Programmcode zur Darstellung des Box-Plots für die Messdaten ohne eine Einteilung in Klassen kann in MATLAB der folgenden Auflistung entnommen werden.

```

1 % Grafische Darstellung
2 f = figure(3);
3 boxplot(Klebermenge);

```

Der Box-Plot für die Zusammenfassung der Datenanalyse mit Einteilung in Klassen muss unter Verwendung der berechneten Kenngrößen manuell programmiert werden, da MATLAB keine Funktionen zur Berechnung von Kenngrößen gruppierter Daten zur Verfügung stellt.

3.4.4 Programmbeispiel zur Datenanalyse in Python

Die Datenanalyse in Python erfordert zunächst das Laden der erforderlichen Module und eine Definition des Ortes, an dem die Grafiken angezeigt werden sollen.

```

1 Bibliotheken importieren
2 import numpy as np
3 import pandas as pd
4 import matplotlib.pyplot as plt
5 from scipy.stats import skew
6 from scipy.io import loadmat
7 from IPython import get_ipython
8 get_ipython().run_line_magic('matplotlib', 'inline')

```

Die Daten liegen als mat-Datei vor. Python bietet einen Befehl zum Laden dieser Daten an. Die Daten werden in ein eindimensionales Array gewandelt.

```

1 Einlesen und Umsortieren der Daten aus dem .mat-file
2 data = loadmat('Klebermenge')[['data']]
3 X = np.array(data).reshape(data.shape[0]*data.shape[1])

```

Der Datenumfang und die statistischen Kennwerte werden mit den angegebenen Befehlen berechnet. Dabei werden die in MATLAB berechneten Ergebnisse bestätigt.

```

1 Bestimmung Datenumfang
2 Xmin = np.amin(X)
3 Xmax = np.amax(X)
4 N = X.shape[0]
5
6 Berechnen der Lagekennwerte Mittelwert und Median
7 X_mean = np.mean(X)
8 print(' ')
9 print('Arithmetisches Mittelwert: ', X_mean)
10 X_med = np.median(X)
11 print('Median: ', X_med)
12
13 Berechnen der Streuungskennwerte Standardabweichung, Varianz und
14 Inter-Quartil-Range, bei Varianz muss die Anzahl der Freiheitsgrade
15 auf N - 1 angepasst werden
16 X_var = np.var(X, ddof=1)
17 print(' ')
18 print('Varianz: ', X_var)
19 X_std = np.std(X, ddof=1)
20 print('Standardabweichung: ', X_std)
21 X_iqr = np.quantile(X, 0.75) - np.quantile(X, 0.25)
22 print()
23
24 Berechnen der Schiefe

```

```

25 X_skew_mom = skew(X)
26 print(' ')
27 print('Momentenkoeffizient der Schiefe: ', X_skew_mom )
28 X_skew_qua = (np.quantile(X,0.75) - 2*np.quantile(X,0.5) + np.quantile(X
29 ,0.25))/X_iqr
print('Quartilkoeffizient der Schiefe: ', X_skew_qua )

```

Die Daten werden mit einem Streudiagramm sowie einem Boxplot visualisiert.

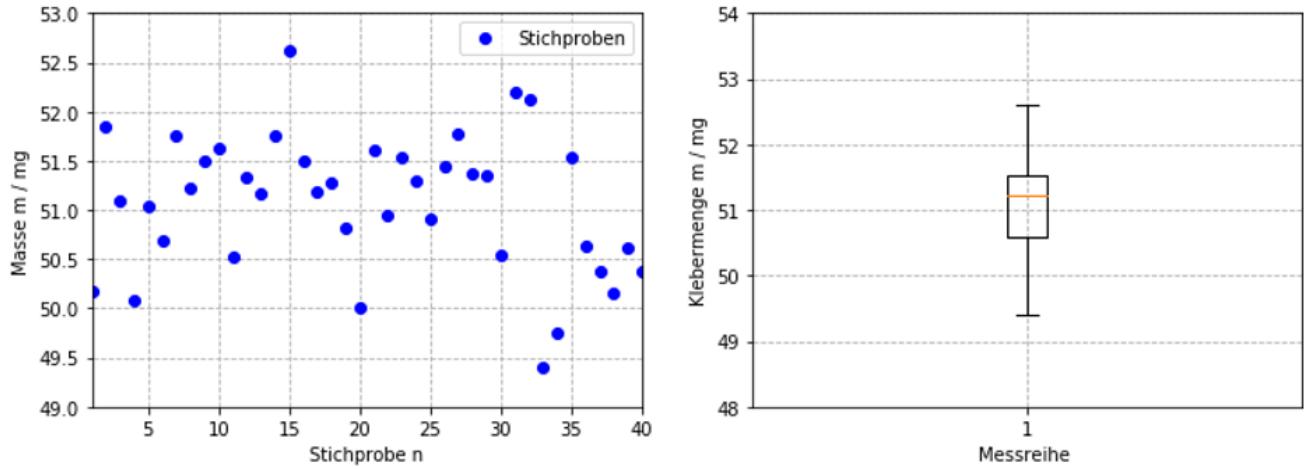


Bild 3.17: Darstellung der Messwerte als Streudiagramm und Boxplot

```

1 Grafische Darstellung der einzelnen Messwerte als Streudiagramm
2 fig = plt.figure(1, figsize=(12, 4))
3 ax1, ax2 = fig.subplots(1,2)
4 n = np.arange(1,N+1)
5 ax1.plot(n,X, 'bo', Linewidth = 2, label = 'Stichproben')
6 ax1.grid(True, which= 'both', axis='both', linestyle='--')
7 ax1.axis([1, N, 49, 53])
8 ax1.legend(loc = 'upper right')
9 ax1.set_xlabel('Stichprobe n')
10 ax1.set_ylabel('Masse m / mg')
11
12 Erstellen eines Boxplot
13 ax2.boxplot(X)
14 ax2.grid(True, which='both', axis='both', linestyle='--')
15 ax2.set_xlabel('Messreihe ')
16 ax2.set_ylabel('Klebermenge m / mg')
17 ax2.axis([0, 2, 48, 54])

```

Der steigende Datensatz wird mit Hilfe von dem Modul Pandas in eine Tabelle gewandelt und ausgegeben. Dabei werden absolute und relative Häufigkeit sowie die absolute und relative Häufigkeit berechnet.

```

1 Berechnung der absoluten und relativen Häufigkeit sowie der absoluten
2 und relativen Summenhäufigkeit
3 X_freq , Klassengrenzen = np.histogram(X, bins=np.arange(np.floor(X_min)
4 -0.5, np.ceil(X_max)+1.5))
5 X_rel_freq = X_freq/N
6 X_sum_freq = np.cumsum(X_freq)
7 X_rel_sum_freq = X_sum_freq/N
8 Klassenmitten = np.arange(np.floor(X_min),np.ceil(X_max)+1)

9 Generieren einer Tabelle in Pandas und Ausgabe der Tabelle
10 Tabelle = pd.DataFrame({ 'Gruppenwert' : Klassenmitten ,
11                         'Absolute Häufigkeit hA(x)' : X_freq ,
12                         'Relative Häufigkeit h(x)' : X_rel_freq ,
13                         'Absolute Summenhäufigkeit HA(x)' : X_sum_freq ,
14                         'Relative Summenhäufigkeit H(x)' : X_rel_sum_freq })
15 print(' ')
16 print(Tabelle)

```

Das Ergebnis wird als Histogramm und mit seiner relativen Summenhäufigkeit dargestellt.

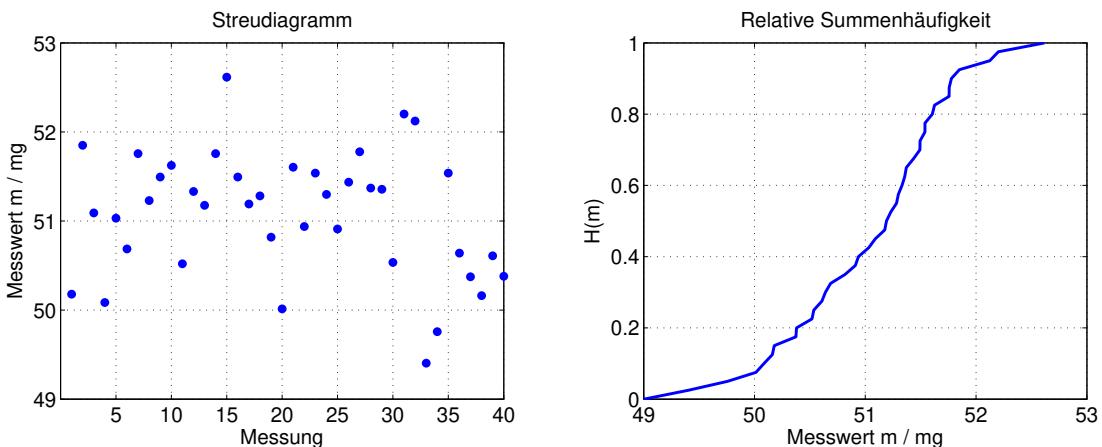


Bild 3.18: Darstellung der Messwerte als Histogramm und relative Summenhäufigkeit

```

1 Grafische Darstellung der relativen Häufigkeiten als Histogramm
2 fig = plt.figure(2, figsize=(12, 4))
3 ax1, ax2 = fig.subplots(1,2)
4 ax1.hist(X, Klassengrenzen, histtype='bar' , color='b' , weights=np.ones(N)/
5 N, rwidth=1)
6 ax1.grid(True, which='both' , axis='both' , linestyle='--')
7 ax1.set_xlabel('Klebermenge m / mg')
8 ax1.set_ylabel('Relative Häufigkeit h(m)')
9 ax1.axis([48, 54, 0, 0.6])

10 Grafische Darstellung der relativen Summenhäufigkeit
11 Xsort = np.append(np.append(48,np.sort(X)),54)
12 Psum = np.append(np.append([0,],np.arange(1,N+1)/N),1)
13 ax2.step(Xsort,Psum, color='b' , where='post' , linewidth=2)
14 ax2.grid(True, which='both' , axis='both' , linestyle='--')
15 ax2.set_xlabel('Klebermenge m / mg')
16 ax2.set_ylabel('Relative Summenhäufigkeit H(m)')
17 ax2.axis([48, 54, 0, 1])

```

3.5 Literatur

- [Krey91] Kreyszig, Erwin: Statistische Methoden und ihre Anwendungen
4., unveränderter Nachdruck der 7. Auflage
Vandenhoeck & Ruprecht, Göttingen, 1991
- [Fahr06] Fahrmeir, Ludwig; Künstler, Rita; Pigeot, Iris; Tutz, Gerhard: Der Weg zur Datenanalyse
6. Auflage
Springer Berlin Heidelberg New York, 2006
- [Ross06] Ross, M. Sheldon: Statistik für Ingenieure und Naturwissenschaftler
3. Auflage
Spektrum Akademischer Verlag, München, 2006

4 Univariate Wahrscheinlichkeitstheorie

Wahrscheinlichkeitsverteilungen geben an, mit welcher Wahrscheinlichkeit unterschiedliche Zufallsergebnisse eintreffen. Sie beschreiben diese Eintreffwahrscheinlichkeit als Funktion einer sogenannten Zufallsvariable. Die Wahrscheinlichkeitsverteilung stellt damit das theoretische Gegenstück zur empirischen Häufigkeitsverteilung dar, die sich aus der Analyse vorhandener Daten wie zum Beispiel Messwerten ergibt.

Wie bei Häufigkeitsverteilungen wird zwischen diskreten und stetigen Wahrscheinlichkeitsverteilungen unterschieden. Beispiel für eine diskrete Verteilung ist die Hypergeometrische Verteilung, die in der Qualitätssicherung bei Stichproben-Eingangsprüfungen eingesetzt wird. Viele Prozesse und Produktmerkmale werden aber über stetige Zufallsvariablen beschrieben. Deshalb basieren viele Methoden des Design For Six Sigma auf stetigen Zufallsvariablen und damit auf stetigen Verteilungen. Der wichtigste Vertreter von stetigen Verteilungen ist die Normalverteilung, mit der sich viele reale Zufallsprozesse approximativ beschreiben lassen.

Nach der Einführung des Begriffes der Zufallsvariablen und ihrer Verteilungen werden Kenngrößen von Wahrscheinlichkeitsverteilungen berechnet. Diese Erkenntnisse werden anschließend an speziellen diskreten und stetigen Verteilungen angewendet.

4.1 Zufallsvariablen und Wahrscheinlichkeitsverteilungen

4.1.1 Zufallsvariablen

Als Zufallsvariable wird allgemein eine Variable bezeichnet, die das Ergebnis eines Zufallsexperiments repräsentiert. Zum Beispiel kann bei einem Würfelexperiment mit zwei Würfeln das Zufallsergebnis über die Summe der beiden Augenzahlen dargestellt werden. Bei einem Prozess mit stetigen Ergebniswerten, wie zum Beispiel der Fertigung von Widerständen, wird die Abweichung vom Sollwert durch eine Zufallsvariable beschrieben. Aber auch, wenn die bei einem Experiment denkbaren Ereignisse nicht direkt mit Zahlen beschrieben werden, kann jedem möglichen Ereignis eine Zahl zugeordnet werden. Diese Zahl gibt im einfachsten Fall den Index der Menge an, zu dem das Ergebnis des Zufallsexperiments gehört. Zum Beispiel können die Permutationen der Buchstaben a, b und c in Gruppen eingeteilt werden, die einer Zufallsvariable x zugeordnet werden.

Tabelle 4.1: Zuordnung von Zufallsereignissen zu einer Zufallsvariablen x

Zufallsvariable x	Ergebnis des Experiments
1	abc bac
2	cab acb
3	bca cba

In jedem dieser Beispiele kann die Zufallsvariable x verschiedene Werte annehmen. Aber es lässt sich nicht vorhersagen, welchen Wert die Zufallsvariable annehmen wird, da dieser Wert vom Einfluss unkontrollierbarer Umstände abhängt.

Zufallsvariablen können in die gleichen Gruppen eingeteilt werden, wie die Merkmalstypen bei der beschreibenden Statistik. Stetige Merkmalstypen entsprechen stetigen Zufallsvariablen, diskrete Merkmalstypen sowie ordinale und gruppierende Merkmalstypen werden über diskrete Zufallsvariablen

abgebildet.

Mathematisch wird einem Zufallsexperiment eine Zufallsvariable x zugeordnet, die folgende Eigenschaften besitzen muss:

- Die Werte von x sind reelle Zahlen.
- Für jede Zahl a und für jedes Intervall I ist die Wahrscheinlichkeit des Ereignisses $x = a$ und $x \in I$ im Einklang mit den Axiomen der Wahrscheinlichkeit.

Beispiel: Würfeexperiment

Anhand des Zufallsexperiments Würfeln mit zwei Würfeln wird der zweite Teil der Definition erläutert. Die Zufallsvariable x stellt die Summe der beiden Augenzahlen dar. Die Wahrscheinlichkeit $P(x = 2)$ ist die Wahrscheinlichkeit, dass das Ergebnis des Zufallsexperiments die Augenzahl 2 ist. Da dieses Ergebnis nur erreicht wird, wenn zweimal die Augenzahl 1 erscheint, ist die Wahrscheinlichkeit

$$P(x = 2) = \frac{1}{6} \cdot \frac{1}{6} = \frac{1}{36} < 1 \quad (4.1)$$

Auch für jedes andere Zufallsereignis des Experimentes liegt die Wahrscheinlichkeit im Bereich $0 \leq P \leq 1$. Das sichere Ereignis ist, dass die Summe der Augenzahlen zwischen 2 und 12 liegt. Nach Axiom 2 ist die Wahrscheinlichkeit für dieses Ereignis 1.

$$P(S) = P(2 \leq x \leq 12) = 1 \quad (4.2)$$

Die Ereignisse $x = 2$ und $x \neq 2$ schließen sich gegenseitig aus. Nach Axiom 3 gilt damit für die Wahrscheinlichkeit, dass $x \neq 2$ ist:

$$P(x \neq 2) = P(2 \leq x \leq 12) - P(x = 2) = 1 - \frac{1}{36} = \frac{35}{36} \quad (4.3)$$

Das Zufallsexperiment erfüllt damit die Axiome der Wahrscheinlichkeit, die Summe der beiden Augenzahlen ist eine Zufallsvariable.

Durch die Zuordnung von Ergebnissen eines Zufallsexperiments zu Zufallsvariablen ist es möglich, die Wahrscheinlichkeit von Zufallsexperimenten als Wahrscheinlichkeitsverteilung der Zufallsvariable x darzustellen und mit ihr zu rechnen.

4.1.2 Diskrete Zufallsvariablen und Verteilungen

Eine Zufallsvariable ist diskret, wenn die Variable x nur endliche viele Werte x_1, x_2, \dots, x_N annehmen kann, die jeweils eine positive Wahrscheinlichkeit aufweisen. Außerdem ist in jedem Intervall $a < x \leq b$, in dem kein Wert für x liegt, die Wahrscheinlichkeit null. Jedem vorkommenden Wert von x_n ist eine Wahrscheinlichkeit $f(x_n) = P(x_n) = P_n$ zugeordnet. Die Funktion $f(x)$ wird als Wahrscheinlichkeitsverteilung bezeichnet. Sie weist jedem Wert x_n der Zufallsvariable x einen Wahrscheinlichkeitswert $f(x_n)$ zu. Ist die Wahrscheinlichkeitsverteilung $f(x)$ bekannt, kann sie zur Berechnung der Wahrscheinlichkeit $P(a < x \leq b)$ verwendet werden.

$$P(a < x \leq b) = \sum_{x_n > a}^b f(x_n) = \sum_{x_n > a}^b P_n \quad (4.4)$$

Insbesondere gilt nach Gleichung (4.4) für die Wahrscheinlichkeit, dass die Werte x_n der Zufallsvariable nicht oberhalb eines Wertes x liegt

$$P(x_n \leq x) = \sum_{x_n = -\infty}^x f(x_n) = F(x) \quad (4.5)$$

Diese Funktion wird als Verteilungsfunktion $F(x)$ bezeichnet.

Beispiel: Würfelexperiment

Als Beispiel wird erneut das Würfeln mit zwei Würfeln aufgegriffen. Für das Würfeln mit zwei Würfeln ergibt sich die in Tabelle 3.2 dargestellte Wahrscheinlichkeitsverteilung $f(x)$ und Verteilungsfunktion $F(x)$.

Tabelle 4.2: Verteilungen für die Augensumme beim Würfeln mit zwei Würfeln

Index	2	3	4	5	6	7	8	9	10	11	12
$f(x)$	$\frac{1}{36}$	$\frac{2}{36}$	$\frac{3}{36}$	$\frac{4}{36}$	$\frac{5}{36}$	$\frac{6}{36}$	$\frac{5}{36}$	$\frac{4}{36}$	$\frac{3}{36}$	$\frac{2}{36}$	$\frac{1}{36}$
$F(x)$	$\frac{1}{36}$	$\frac{3}{36}$	$\frac{6}{36}$	$\frac{10}{36}$	$\frac{15}{36}$	$\frac{21}{36}$	$\frac{26}{36}$	$\frac{30}{36}$	$\frac{33}{36}$	$\frac{35}{36}$	$\frac{36}{36}$

Bild 4.1 stellt die Wahrscheinlichkeitsverteilung $f(x)$ und die Verteilungsfunktion $F(x)$ zum Beispiel Augenzahl beim Würfeln mit zwei Würfeln dar.

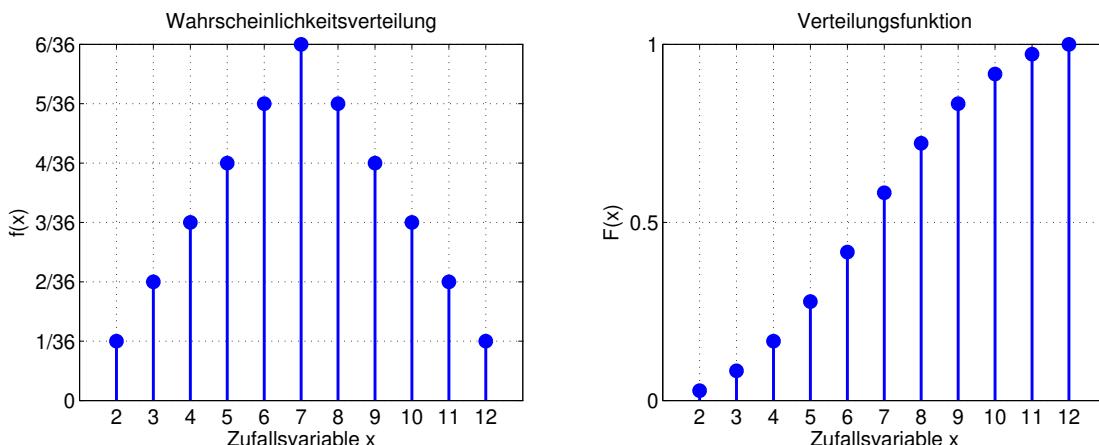


Bild 4.1: Grafische Darstellung der Wahrscheinlichkeitsverteilung $f(x)$ und Verteilungsfunktion $F(x)$ für die Augenzahl beim Würfeln mit zwei Würfeln

Zum Beispiel ist für das Würfeln die Wahrscheinlichkeit für einen Zahlenwert $3 < x \leq 7$

$$P(3 < x \leq 7) = \sum_{x_n > 3}^7 f(x_n) = \frac{3}{36} + \frac{4}{36} + \frac{5}{36} + \frac{6}{36} = \frac{18}{36} = \frac{1}{2} \quad (4.6)$$

Sie kann aber auch aus der Verteilungsfunktion berechnet werden.

$$P(3 < x \leq 7) = F(7) - F(3) = \frac{21}{36} - \frac{3}{36} = \frac{1}{2} \quad (4.7)$$

Die Werte x_n , an denen die diskrete Zufallsvariable x eine positive Wahrscheinlichkeit besitzt, werden als mögliche Werte von x bezeichnet. In jedem Intervall, in dem kein möglicher Wert von x liegt, bleibt die Verteilungsfunktion $F(x)$ konstant. $F(x)$ ist also eine Funktion, die an den Stellen $x = x_n$ um die Wahrscheinlichkeit $P_n = f(x_n)$ springt und ansonsten konstant bleibt. Für Werte der Zufallsvariable x , die kleiner als der kleinste mögliche Wert x_{min} sind, gilt

$$F(x < x_{min}) = 0 \quad (4.8)$$

Für den größten vorkommenden Wert x_{max} nimmt die Verteilungsfunktion den Wert

$$F(x \geq x_{max}) = 1 \quad (4.9)$$

an. Der Informationsgehalt der Wahrscheinlichkeitsverteilung $f(x)$ und der Verteilungsfunktion $F(x)$ sind identisch. Nach Gleichung (4.5) gilt für zwei Zufallsvariablen mit dem Abstand d

$$F(x) = \sum_{x_n=-\infty}^x f(x_n) \quad (4.10)$$

und

$$F(x-d) = \sum_{x_n=-\infty}^{x-d} f(x_n) \quad (4.11)$$

Damit kann die Wahrscheinlichkeitsverteilung $f(x)$ bestimmt werden aus der Differenz

$$f(x) = F(x) - F(x-d) \quad (4.12)$$

Die Darstellungen können also mithilfe dieser Gleichungen ineinander überführt werden. Obwohl beide Funktionen denselben Informationsgehalt haben, ist die Wahrscheinlichkeitsverteilung $f(x)$ für diskrete Verteilungen die anschaulichere Darstellung. Im Gegensatz dazu muss bei Rechnungen mit stetigen Verteilungen die Verteilungsfunktion $F(x)$ verwendet werden, weshalb beide Formen der Darstellung ihre Berechtigung haben.

Die Wahrscheinlichkeitsverteilung $f(x)$ und die Verteilungsfunktion $F(x)$ sind vergleichbar mit den Darstellungen zur relativen Häufigkeit $h(x)$ und der relativen Summenhäufigkeit $H(x)$ in der beschreibenden Statistik. Allerdings liegt bei der beschreibenden Statistik eine konkrete Stichprobe vor, während die Wahrscheinlichkeitsverteilung $f(x)$ und die Verteilungsfunktion $F(x)$ die Grundgesamtheit eines Wahrscheinlichkeitsexperimentes charakterisieren.

4.1.3 Stetige Zufallsvariablen und Verteilungen

Eine Zufallsvariable x ist stetig, wenn die zugehörige Verteilungsfunktion $F(x)$ in Integralform dargestellt werden kann.

$$P(\xi \leq x) = F(x) = \int_{-\infty}^x f(\xi) d\xi \quad (4.13)$$

Da in Gleichung (4.13) die obere Grenze des Integrals der Wert x ist, wird die Integrationsvariable mit dem griechischen ξ bezeichnet. Der Integrand ist eine positive Funktion, die stetig ist oder einzelne Sprünge aufweist. Dadurch ist die Verteilungsfunktion $F(x)$ stetig. Der Integrand $f(x)$ heißt Wahrscheinlichkeitsdichte der Zufallsvariable x . Aus Gleichung (4.13) folgt durch Differentiation

$$f(x) = \frac{dF}{dx} \quad (4.14)$$

Die Wahrscheinlichkeitsdichte $f(x)$ ist demnach die Ableitung der Verteilungsfunktion $F(x)$. Auch bei stetigen Verteilungen ist der Informationsgehalt identisch, sie lassen sich mit den Gleichungen (4.13) und (4.14) ineinander umrechnen. Da das Ereignis $-\infty < x \leq \infty$ ein sicheres Ereignis ist, ergibt sich

$$P(-\infty < x \leq \infty) = F(\infty) = \int_{-\infty}^{\infty} f(x) dx = 1 \quad (4.15)$$

Weiterhin gilt für das Ereignis $a < x \leq b$ analog zur Gleichung (4.4)

$$P(a < x \leq b) = F(b) - F(a) = \int_a^b f(x) dx \quad (4.16)$$

Die Wahrscheinlichkeit des Ereignisses $a < x \leq b$ entspricht der Fläche unter der Kurve der Wahrscheinlichkeitsdichte $f(x)$ zwischen den Punkten $x = a$ und $x = b$. Bild 4.2 stellt diesen Zusammenhang grafisch dar.

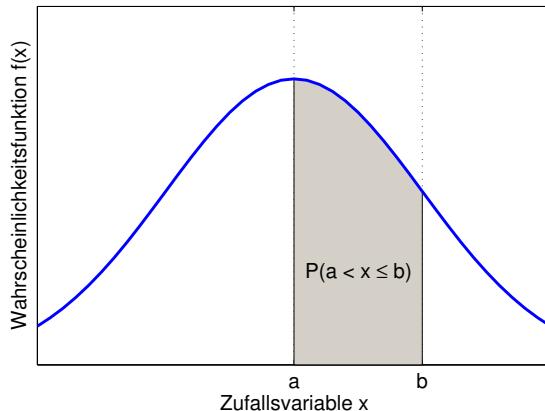


Bild 4.2: Grafische Darstellung der Wahrscheinlichkeit $P(a < x \leq b)$

Für ein singuläres Ereignis $x = a$ ist die Abszisse unendlich klein. Damit wird auch die Fläche unter der Kurve unendlich klein und die Wahrscheinlichkeit für den Wert $x = a$ ist

$$P(x = a) = 0 \quad (4.17)$$

Das bedeutet nicht, dass $x = a$ ein unmögliches Ergebnis ist, sondern dass die Wahrscheinlichkeit unendlich klein ist, bei einer stetigen Zufallsvariablen genau einen definierten Wert zu erzielen.

Beispiel: Glücksrad

Ein Glücksrad, das einen beliebigen Winkel x zwischen 0 und 2π einnehmen kann, kann als Zufallsexperiment aufgefasst werden. Alle Winkel x sind gleich wahrscheinlich, deshalb ist die Wahrscheinlichkeitsdichte $f(x)$ konstant P_0 . Das sichere Ereignis ist, dass der Winkel x in dem Intervall $0 < x \leq 2\pi$ liegt. Es gilt:

$$F(2\pi) = \int_0^{2\pi} f(x)dx = \int_0^{2\pi} f_0 dx = 2\pi \cdot f_0 = 1 \quad (4.18)$$

Daraus ergibt sich durch Auflösen nach P_0 f_0 die Wahrscheinlichkeitsdichte in dem Bereich $0 < x \leq 2\pi$ von

$$f(x) = f_0 = \frac{1}{2\pi} \quad (4.19)$$

und die Verteilungsfunktion

$$F(x) = \int_0^x \frac{1}{2\pi} d\xi = \frac{x}{2\pi} \quad (4.20)$$

Bild 4.3 stellt die Wahrscheinlichkeitsdichte $f(x)$ und die Verteilungsfunktion $F(x)$ für das Experiment dar.

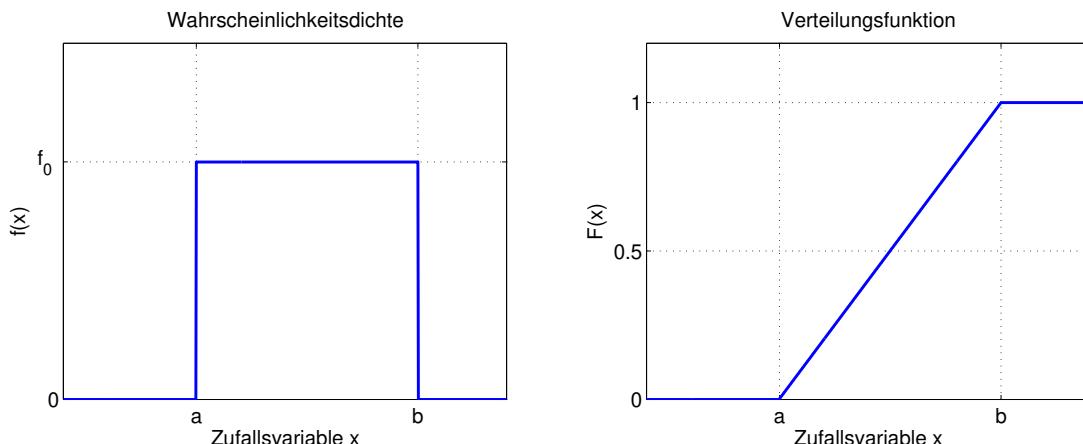


Bild 4.3: Grafische Darstellung der Wahrscheinlichkeitsdichte $f(x)$ und Verteilungsfunktion $F(x)$ für das Glücksrad-Experiment

Die Wahrscheinlichkeit, dass das Glücksrad mit einem Winkel in dem Bereich von $a < x \leq b$ stehen bleibt, ist

$$P(a < x \leq b) = \int_a^b \frac{1}{2\pi} dx = \frac{b-a}{2\pi} \quad (4.21)$$

Alle Ereignisse in dem definierten Intervall zwischen a und b sind gleich wahrscheinlich.

4.1.4 Nomenklatur für diskrete und stetige Zufallsvariablen

Diskrete und stetige Verteilungen werden mit den Funktionen $f(x)$ und $F(x)$ beschrieben. Die Funktion $F(x)$ wird in beiden Fällen als Verteilungsfunktion bezeichnet. Bei diskreten Zufallsvariablen ist $f(x)$ die Wahrscheinlichkeitsverteilung, bei stetigen Zufallsvariablen ist $f(x)$ die Wahrscheinlichkeitsdichte. Bei Stichproben wird von relativen Häufigkeitsverteilungen $h(x)$ und relativen Summenhäufigkeiten $H(x)$ gesprochen. Damit ergeben sich die in Tabelle 4.3 zusammengefassten Bezeichnungen.

Tabelle 4.3: Zusammenfassung der Nomenklatur für Stichproben sowie diskrete und stetige Zufallsvariablen

Stichprobe	Diskreter Zufallsprozess	Stetiger Zufallsprozess
Relative Häufigkeitsverteilung $h(x)$	Wahrscheinlichkeitsverteilung $f(x)$	Wahrscheinlichkeitsdichte $f(x)$
Relative Summenhäufigkeit $H(x)$	Verteilungsfunktion $F(x)$	Verteilungsfunktion $F(x)$

4.2 Erwartungswerte von Verteilungen

Der Erwartungswert $E(x)$ der Zufallsvariablen x entspricht dem Wert, der sich bei oftmaligem Wiederholen des zugrunde liegenden Zufallsexperiments im Mittel einstellt. Durch den Erwartungswert $E(x)$ wird später die Lage der vorliegenden Verteilung beschrieben. Der Erwartungswert kann aber nicht nur von der Zufallsvariablen x bestimmt werden, sondern auch von Funktionen von Zufallsvariablen $g(x)$. Der Erwartungswert eignet sich damit zur Bestimmung von Kenngrößen einer Verteilung. Dieser Zusammenhang wird in Abschnitten 4.4.3 gezeigt.

4.2.1 Definition des Erwartungswert-Operators

Für eine beliebige Zufallsvariable x und eine für alle Werte von x definierte reellwertige Funktion $y = g(x)$ der Zufallsvariablen x wird der Ausdruck

$$E(y) = E(g(x)) = \sum_{x_n=-\infty}^{\infty} (g(x_n) \cdot f(x_n)) \quad (4.22)$$

beziehungsweise

$$E(y) = E(g(x)) = \int_{-\infty}^{\infty} g(x) \cdot f(x) dx \quad (4.23)$$

als mathematischer Erwartungswert der Funktion $g(x)$ oder als Erwartung von $g(x)$ bezeichnet. Dabei wird die Konvergenz der Summe beziehungsweise des Integrals vorausgesetzt.

Beispiel: Glücksspiel mit Würfeln

Zwei Personen A und B spielen das folgende Spiel: A würfelt mit einem regelmäßigen Würfel und erhält von B

- 10 Cent für eine Eins oder Zwei,
- 20 Cent für eine Drei oder Vier,
- 40 Cent für eine Fünf und
- 80 Cent für eine Sechs.

Spieler A soll an Spieler B vor jedem Spiel einen Betrag von 50 Cent zahlen. Um zu überprüfen, ob Spieler A bei mehreren Spielen einen Gewinn erzielt, muss die durchschnittliche Gewinnerwartung pro Spiel berechnet werden. Die durchschnittliche Gewinnerwartung ergibt sich aus der Wahrscheinlichkeit, eine bestimmte Zahl zu würfeln, und dem zugeordneten Gewinn. Zum Beispiel beträgt die Wahrscheinlichkeit, eine Eins zu würfeln, $1/6$, und der zugehörige Gewinn beträgt 10 Cent.

Zur Berechnung wird eine Zufallsvariable x definiert, die die beim Wurf des Würfels erzielte Augenzahl darstellt. In Tabelle 4.4 ist jedem möglichen Wert der Zufallsvariablen x der Gewinn von A zugeordnet. Er bildet die Funktion $g(x)$.

Tabelle 4.4: Zuordnung des Gewinns $g(x)$ zur Zufallsvariablen x

x	1	2	3	4	5	6
(x)	10	10	20	20	40	80

Da die Werte von x vom Zufall abhängen, gilt dies auch für den Wert, den $g(x)$ bei einem Spiel jeweils annimmt. Die Funktion $g(x)$ der Zufallsvariable x ist also selbst eine Zufallsvariable $y = g(x)$. In diesem Beispiel beträgt die durchschnittliche Gewinnerwartung

$$E(y) = E(g(x)) = \sum_{x_n=1}^6 (g(x_n) \cdot f(x_n)) = 10 \cdot \frac{1}{6} + 10 \cdot \frac{1}{6} + 20 \cdot \frac{1}{6} + 20 \cdot \frac{1}{6} + 40 \cdot \frac{1}{6} + 80 \cdot \frac{1}{6} = 30 \quad (4.24)$$

Die durchschnittliche Gewinnerwartung pro Spiel ist mit 30 Cent geringer als der Spieleinsatz von 50 Cent. Spieler A wird bei mehreren Spieldurchgängen demnach Geld verlieren.

4.2.2 Eigenschaften des Erwartungswert-Operators

Es gibt einige Eigenschaften des Erwartungswert-Operators, die dazu verwendet werden können, den Erwartungswert von Funktionen von Zufallsvariablen zu bestimmen. Da der Erwartungswert für stetige Zufallsgrößen über ein Integral definiert ist, ergeben sich die Eigenschaften des Erwartungswert-Operators aus den Eigenschaften der Integralrechnung. Im Folgenden werden Rechenregeln für den Erwartungswert stetiger Zufallsvariablen hergeleitet. Dieselben Rechenregeln gelten auch für diskrete Zufallsvariablen.

Erwartungswert einer Konstanten

Der Erwartungswert einer konstanten Größe k besitzt den Wert k , denn es gilt:

$$E(k) = \int_{-\infty}^{\infty} k \cdot f(x) dx = k \cdot \int_{-\infty}^{\infty} f(x) dx = k \cdot 1 = k \quad (4.25)$$

Linearität des Erwartungswert-Operators

Der Erwartungswert ist ein linearer Operator. Für die Zufallsvariable

$$y = a \cdot h(x) + b \cdot g(x) \quad (4.26)$$

wird der Erwartungswert berechnet aus

$$\begin{aligned} E(y) &= E(a \cdot h(x) + b \cdot g(x)) = \int_{-\infty}^{\infty} (a \cdot h(x) + b \cdot g(x)) \cdot f(x) dx \\ &= a \cdot \int_{-\infty}^{\infty} h(x) \cdot f(x) dx + b \cdot \int_{-\infty}^{\infty} g(x) \cdot f(x) dx = a \cdot E(h(x)) + b \cdot E(g(x)) \end{aligned} \quad (4.27)$$

Ein Sonderfall ist die lineare Transformation der Form

$$y = a \cdot x + b \quad (4.28)$$

Aufgrund der Linearität des Erwartungswertes ergibt sich

$$E(y) = E(a \cdot x + b) = a \cdot E(x) + b \quad (4.29)$$

Ein Vergleich der Gleichungen (4.28) und (4.29) zeigt, dass sich der Erwartungswert analog zur Zufallsvariable verschiebt.

Erwartungswert symmetrischer Verteilungen

Für Ist eine zum Punkt c symmetrische Verteilung $f(x)$ symmetrisch zum Punkt c

$$f(c - x) = f(c + x) \quad (4.30)$$

gilt für den soll der Erwartungswert-Operator

$$E(x) = \int_{-\infty}^{\infty} x \cdot f(x) dx = \int_{-\infty}^c x \cdot f(x) dx + \int_c^{\infty} x \cdot f(x) dx \quad (4.31)$$

berechnet werden. Durch die Substitution $x = c - \xi$ beziehungsweise $x = c + \xi$ ergibt sich

$$E(x) = \int_0^{\infty} (c - \xi) \cdot f(c - \xi) d\xi + \int_0^{\infty} (c + \xi) \cdot f(c + \xi) d\xi \quad (4.32)$$

Wegen der identischen Integrationsgrenzen können die Integrale zusammengefasst werden.

$$\begin{aligned} E(x) &= \int_0^{\infty} c \cdot f(c - \xi) d\xi + \int_0^{\infty} c \cdot f(c + \xi) d\xi - \int_0^{\infty} \xi \cdot f(c - \xi) d\xi + \int_0^{\infty} \xi \cdot f(c + \xi) d\xi \\ &= \int_{-\infty}^{\infty} c \cdot f(c - \xi) d\xi - \int_0^{\infty} \xi \cdot (f(c - \xi) - f(c + \xi)) d\xi \end{aligned} \quad (4.33)$$

Der erste Summand ist der Erwartungswert einer konstanten Größe c . Der zweite Summand besteht aus einer Differenz von Wahrscheinlichkeitsdichten. Da die Wahrscheinlichkeitsdichte symmetrisch ist, ist die Differenz null. Damit ergibt sich der Erwartungswert einer symmetrischen Verteilung zu

$$E(x) = c \cdot \int_{-\infty}^{\infty} f(c - \xi) d\xi - 0 = c \quad (4.34)$$

Zusammenfassung

In Tabelle 4.5 werden die Eigenschaften des Erwartungswert-Operators zusammengefasst. Das Rechnen mit Erwartungswerten ist insbesondere bei der Herleitung von Gesetzmäßigkeiten für die Grundgesamtheit und die Berechnung der Erwartungstreue bei Schätzungen von Bedeutung.

Tabelle 4.5: Zusammenfassung der Eigenschaften des Erwartungswert-Operators

Rechenoperation	Eigenschaften des Erwartungswert-Operators
Definition Erwartungswert	$E(y) = E(g(x)) = \sum_{x_n=-\infty}^{\infty} (g(x_n) \cdot f(x_n))$ <p style="text-align: center;">beziehungsweise</p> $E(y) = E(g(x)) = \int_{-\infty}^{\infty} g(x) \cdot f(x) dx$
Erwartungswert einer Konstanten	$E(k) = k$
Linearität	$E(a \cdot h(x) + b \cdot g(x)) = a \cdot E(h(x)) + b \cdot E(g(x))$
Lineare Transformation	$E(y) = E(a \cdot x + b) = a \cdot E(x) + b$
Symmetrie $f(c - x) = f(x - c)$	$E(x) = c$

4.3 Kennwerte von Verteilungen

Bei der beschreibenden Statistik werden vorliegende Daten aus Stichproben analysiert. Dabei werden für die konkreten Daten Häufigkeitsverteilungen bestimmt sowie empirische Kenngrößen für die Lage, die Streuung und die Symmetrie berechnet. Diese Kenngrößen lassen sich auch für Verteilungen Zufallsvariablen mit einer bekannten Wahrscheinlichkeitsverteilung berechnen. Im Gegensatz zur deskriptiven Statistik werden bei der Wahrscheinlichkeitsrechnung alle theoretisch möglichen Werte der Zufallsvariablen ausgewertet. Deshalb werden die Kennwerte auch als theoretische Kennwerte bezeichnet.

4.3.1 Momente und Zentralmomente einer Verteilung

Kennwerte von Verteilungen können als Moment beziehungsweise als Zentralmoment der Ordnung k berechnet werden. Das k -te Moment einer Verteilung oder der Zufallsvariable x ist definiert als der Erwartungswert der Funktion

$$g(x) = x^k \quad (4.35)$$

Für diskrete Verteilungen ergibt sich

$$E(x^k) = \sum_{x_n=-\infty}^{\infty} (x_n^k \cdot f(x_n)) \quad (4.36)$$

und für stetige Verteilungen berechnet sich das k -te Moment zu

$$E(x^k) = \int_{-\infty}^{\infty} x^k \cdot f(x) dx \quad (4.37)$$

Der Erwartungswert der Funktion

$$g(x) = (x - \mu)^k \quad (4.38)$$

führt zu dem k -ten Zentralmoment. Dabei ist μ der arithmetische Mittelwert der Verteilung, auf ihn wird in Abschnitt 4.3.2 detailliert eingegangen. Für diskrete Verteilungen berechnet sich das k -te Zentralmoment aus

$$E((x - \mu)^k) = \sum_{x_n=-\infty}^{\infty} ((x_n - \mu)^k \cdot f(x_n)) \quad (4.39)$$

und für stetige Verteilungen aus

$$E((x - \mu)^k) = \int_{-\infty}^{\infty} (x - \mu)^k \cdot f(x) dx \quad (4.40)$$

Aufgrund der Linearität des Erwartungswertes kann die Berechnung des Zentralmomentes auf die Berechnung des Momentes zurückgeführt werden. Zum Beispiel ergibt sich für das zweite Zentralmoment

$$\begin{aligned} E((x - \mu)^2) &= E(x^2 - 2 \cdot \mu \cdot x + \mu^2) = E(x^2) - 2 \cdot \mu \cdot E(x) + \mu^2 \\ &= E(x^2) - 2 \cdot \mu^2 + \mu^2 = E(x^2) - \mu^2 \end{aligned} \quad (4.41)$$

Entsprechend können höhere Zentralmomente umgeformt werden. Für das dritte Zentralmoment ergibt sich mit den Rechenregeln des Erwartungswertes

$$E((x - \mu)^3) = E(x^3) - 3 \cdot \mu \cdot E(x^2) + 2 \cdot \mu^3 \quad (4.42)$$

4.3.2 Lagekennwerte einer Verteilung

Wie bei der deskriptiven Statistik können für Wahrscheinlichkeitsverteilungen Lagekennwerte angegeben werden. In Anlehnung an die beschreibende Statistik werden in diesem Abschnitt der arithmetische Mittelwert und der Median einer Verteilung eingeführt.

Arithmetischer Mittelwert einer Verteilung

Der theoretisch erwartete Mittelwert einer Verteilung wird mit μ bezeichnet. Er berechnet sich bei einer diskreten Verteilung zu

$$\mu = \sum_{x_n=-\infty}^{\infty} (x_n \cdot f(x_n)) \quad (4.43)$$

und bei einer stetigen Verteilung zu

$$\mu = \int_{-\infty}^{\infty} x \cdot f(x) dx \quad (4.44)$$

Der Mittelwert kann auch als Erwartungswert eines Momentes erster Ordnung dargestellt werden.

$$\mu = E(x^1) = E(x) \quad (4.45)$$

Der Mittelwert der Verteilung wird auch als Mittelwert der Zufallsvariable x oder Erwartungswert von x bezeichnet. Bei der Berechnung des Mittelwertes wird vorausgesetzt, dass die Summe (4.43) beziehungsweise das Integral (4.44) konvergieren, was aber in praktischen Fällen immer gegeben ist.

Median einer Verteilung

Der Median einer Verteilung ergibt sich aus der Bedingung

$$F(x_{MED}) = 0.5 \quad (4.46)$$

Im Fall einer stetigen Zufallsvariable x ist in praktischen Fällen auch die Verteilungsfunktion $F(x)$ stetig. Damit kann der Median analytisch oder zumindest numerisch bestimmt werden. Im Fall einer diskreten Zufallsvariable x muss der Median mit vergleichbaren Verfahren wie bei der deskriptiven Statistik bestimmt werden.

Beispiel: Würfeln mit zwei Würfeln

Bei dem Beispiel des Würfeln mit zwei Würfeln ergibt sich ein Mittelwert von

$$\mu = \sum_{x_n=-\infty}^{\infty} (x_n \cdot f(x_n)) = 7 \quad (4.47)$$

Der Median errechnet sich nach den Darstellungen zur deskriptiven Statistik aus

$$x_{MED} = c_{n-1} + \frac{0.5 - F(c_{n-1})}{f(c_n)} \cdot d = 6 + \frac{0.5 - \frac{15}{36}}{\frac{6}{36}} \cdot 1 = 6 + \frac{1}{2} = 6.5 \quad (4.48)$$

Beispiel: Glücksrad

Bei dem Beispiel des Glücksrades ergibt sich ein Mittelwert von

$$\mu = \int_0^{2\pi} x \cdot f(x) dx = \int_0^{2\pi} x \cdot \frac{1}{2\pi} dx = \frac{1}{2\pi} \cdot \left(\frac{x^2}{2} \right) \Big|_0^{2\pi} = \pi \quad (4.49)$$

Der Median ergibt sich aus Bild 4.3 zu

$$x_{MED} = \pi \quad (4.50)$$

4.3.3 Streuungskennwerte einer Verteilung

Analog zu den Lagekennwerten von Verteilungen können Streuungskennwerte angegeben werden. In diesem Abschnitt werden die Spannweite, die Varianz und die Quantile einer Verteilung eingeführt.

Spannweite einer Verteilung

Sind die Ereignisse der Verteilung $f(x)$ auf ein Intervall I begrenzt, kann die Spannweite der Verteilung angegeben werden als

$$\Delta x = \max(I) - \min(I) \quad (4.51)$$

Varianz und Standardabweichung einer Verteilung

Die Varianz einer Verteilung ist ein Maß für die theoretisch erwartete Streuung der Werte, die die Zufallsvariable x annehmen kann. Sie wird mit σ^2 bezeichnet und ist über den Erwartungswert des zweiten Zentralmomentes

$$\sigma^2 = E((x - \mu)^2) \quad (4.52)$$

definiert. Im diskreten Fall berechnet sie sich zu

$$\sigma^2 = \sum_{x_n=-\infty}^{\infty} ((x_n - \mu)^2 \cdot f(x_n)) \quad (4.53)$$

und im stetigen Fall zu

$$\sigma^2 = \int_{-\infty}^{\infty} (x - \mu)^2 \cdot f(x) dx \quad (4.54)$$

Die Varianz einer Verteilung wird auch als Varianz der Zufallsvariable x bezeichnet. Die Varianz der Verteilung kann mit den Rechenregeln des Erwartungswertes umgeformt werden zu

$$\sigma^2 = E((x - \mu)^2) = E(x^2 - 2 \cdot x \cdot \mu + \mu^2) = E(x^2) - 2 \cdot \mu \cdot E(x) + \mu^2 = E(x^2) - \mu^2 \quad (4.55)$$

Aus den Definitionsgleichungen ergibt sich, dass die Varianz immer größer oder gleich 0 ist. Die positive Wurzel der Varianz wird bei der beschreibenden Statistik als Standardabweichung eingeführt. Für diskrete Verteilungen berechnet sich die Standardabweichung σ aus

$$\sigma = \sqrt{\sum_{x_n=-\infty}^{\infty} ((x_n - \mu)^2 \cdot f(x_n))} \quad (4.56)$$

und für stetige Verteilungen ergibt sich

$$\sigma = \sqrt{\int_{-\infty}^{\infty} (x - \mu)^2 \cdot f(x) dx} \quad (4.57)$$

Quantilabstände einer Verteilung

Analog zum Median einer Verteilung ergibt sich aus der Bedingung

$$F(x_P) = P \quad (4.58)$$

das P-Quantil einer Verteilung. Im Fall einer stetigen Zufallsvariable x ist auch die Verteilungsfunktion $F(x)$ stetig. Damit kann das P-Quantil analytisch oder zumindest numerisch bestimmt werden. Im Fall einer diskreten Zufallsvariable muss das P-Quantil wieder mit vergleichbaren Verfahren bestimmt werden, wie bei der deskriptiven Statistik. Der Quantilabstand berechnet sich dann aus der Differenz zweier Quantile. Der Inter-Quartil-Range berechnet sich zu

$$IQR = x_{0.75} - x_{0.25} \quad (4.59)$$

Beispiel: Würfeln mit zwei Würfeln

Bei dem Beispiel des Würfels mit zwei Würfeln ergibt sich eine Spannweite von

$$\Delta x = \max(I) - \min(I) = 12 - 2 = 10 \quad (4.60)$$

Die Varianz von

$$\sigma^2 = \sum_{x_n=-\infty}^{\infty} ((x_n - \mu)^2 \cdot f(x_n)) = 5.83 \quad (4.61)$$

führt zu einer Standardabweichung

$$\sigma = \sqrt{\sum_{x_n=-\infty}^{\infty} ((x_n - \mu)^2 \cdot f(x_n))} = \sqrt{5.83} = 2.41 \quad (4.62)$$

Der Inter-Quartil-Range errechnet sich zu

$$IQR = x_{0.75} - x_{0.25} = 8.25 - 4.75 = 3.5 \quad (4.63)$$

Beispiel: Glücksrad

Bei dem Glücksrad erstreckt sich das Intervall I von 0 bis 2π , die Spannweite beträgt damit $\Delta x = 2\pi$. Die Varianz beträgt

$$\sigma^2 = \int_{-\infty}^{\infty} (x - \mu)^2 \cdot f(x) dx = \int_0^{2\pi} (x - \pi)^2 \cdot \frac{1}{2\pi} dx = \frac{1}{3} \cdot \pi^2 \quad (4.64)$$

Damit ergibt sich eine Standardabweichung von

$$\sigma = \sqrt{\int_{-\infty}^{\infty} (x - \mu)^2 \cdot f(x) dx} = \sqrt{\int_0^{2\pi} (x - \pi)^2 \cdot \frac{1}{2\pi} dx} = \frac{1}{\sqrt{3}} \cdot \pi \quad (4.65)$$

Aus Bild 4.3 kann der Inter-Quartil-Range bestimmt werden zu

$$IQR = x_{0.75} - x_{0.25} = \frac{3 \cdot \pi}{2} - \frac{\pi}{2} = \pi \quad (4.66)$$

4.3.4 Schiefe oder Symmetrie einer Verteilung

Nachdem die Begriffe des Momentes und Zentralmomentes einer Verteilung eingeführt sind, wird die Symmetrie einer Verteilung mit dem Momenten- und Quantilkoeffizienten der Schiefe beschrieben.

Momentenkoeffizient der Schiefe

Analog zum Momentenkoeffizient der Schiefe in der deskriptiven Statistik wird das dritte normierte zentrale Moment zur Definition der Schiefe γ einer Verteilung verwendet:

$$\gamma = \frac{1}{\sigma^3} \cdot E((x - \mu)^3) \quad (4.67)$$

Quantilkoeffizient der Schiefe

Alternativ kann die Symmetrie oder Schiefe einer Verteilung über eine Kenngröße charakterisiert werden, die die Symmetrie der Quantile der Verteilung bewertet. Dazu wird der Quantilkoeffizient der Schiefe wie bei der deskriptiven Statistik berechnet aus

$$g_P = \frac{(x_{1-P} - x_{MED}) - (x_{MED} - x_P)}{x_{1-P} - x_P} \quad (4.68)$$

Für $P = 25\%$ ergibt sich der Quartilkoeffizient zu

$$g_{0.25} = \frac{(x_{0.75} - x_{MED}) - (x_{MED} - x_{0.25})}{x_{0.75} - x_{0.25}} \quad (4.69)$$

Die Interpretation des Momenten- oder Quantilkoeffizienten der Schiefe erfolgt analog zu den Ausführungen bei der deskriptiven Statistik.

Tabelle 4.6: Bewertung des Momentenkoeffizienten der Schiefe

Kennwert	Symmetrieeigenschaft
$g_P > 0$	Rechtsschiefe Verteilung
$g_P = 0$	Symmetrische Verteilung
$g_P < 0$	Linksschiefe Verteilung

Beispiel: Schiefe einer Verteilung

Die Schiefe einer stetigen Verteilung wird an einem Beispiel verdeutlicht.

$$f(x) = \begin{cases} 0 & \text{für } x < 0 \\ x \cdot e^{-x} & \text{für } x \geq 0 \end{cases} \quad (4.70)$$

Bild 4.4 stellt die Verteilung aus Gleichung (4.70) dar.

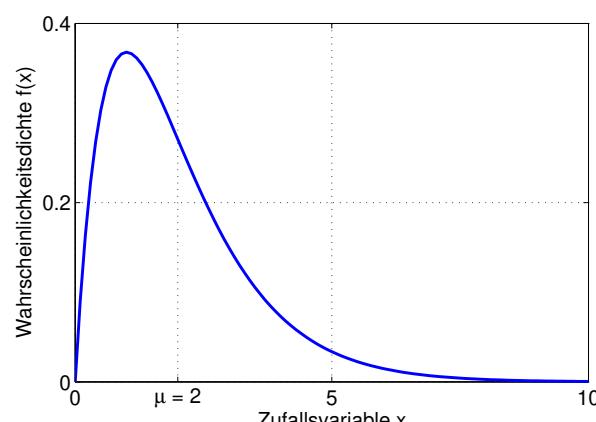


Bild 4.4: Grafische Darstellung der Wahrscheinlichkeitsverteilung aus Gleichung (4.70)

Durch Einsetzen in die Definitionsgleichungen ergibt sich der Mittelwert zu

$$\mu = E(x) = \int_0^{\infty} x^2 \cdot e^{-x} dx = 2 \quad (4.71)$$

das zweite Moment zu

$$E(x^2) = \int_0^{\infty} x^3 \cdot e^{-x} dx = 6 \quad (4.72)$$

und das dritte Moment zu

$$E(x^3) = \int_0^{\infty} x^4 \cdot e^{-x} dx = 24 \quad (4.73)$$

Mit den Gleichungen (4.55) und (4.67) sowie (4.72) und (4.73) können die Varianz und die Schiefe der Verteilung berechnet werden zu

$$\sigma^2 = E((x - \mu)^2) = E(x^2) - \mu^2 = 6 - 4 = 2 \quad (4.74)$$

$$\gamma = \frac{1}{\sigma^3} \cdot (E(x^3) - 3 \cdot \mu \cdot E(x^2) + 2 \cdot \mu^3) = \frac{24 - 3 \cdot 2 \cdot 6 + 2 \cdot 8}{2 \cdot \sqrt{2}} = \sqrt{2} \quad (4.75)$$

Es handelt sich um eine Funktion mit positiver Schiefe $\gamma = 1.4$, die Verteilung ist damit rechtsschief.

Lageregeln zur Interpretation der Symmetrie einer Stichprobe

Die Symmetrieeigenschaften der Verteilung einer Stichprobe können auch an der Lage von Median und Mittelwert abgelesen werden. Auch dazu wird auf die Ausführungen zur deskriptiven Statistik verwiesen.

Tabelle 4.7: Lageregeln von Median und arithmetischem Mittelwert zur Beschreibung der Symmetrie

Lagekennwerte	Symmetrieeigenschaft
$\mu > x_{MED}$	Rechtsschiefe Verteilung
$\mu = x_{MED}$	Symmetrische Verteilung
$\mu < x_{MED}$	Linksschiefe Verteilung

4.3.5 Zusammenfassung Kennwerte von Verteilungen

Tabelle 4.8 fasst die Kennwerte von Verteilungen zusammen. Die Definition der Parameter über Momente ist von größerem praktischen Nutzen und wird in den folgenden Kapiteln weiter benötigt und ausgebaut.

Tabelle 4.8: Zusammenfassung der Kennwerte von Verteilungen

Momente	Basis für Definition	
	Momente	Quantile
Lage	Mittelwert $\mu = E(x^1) = E(x)$	Median $F(x_{MED}) = 0.5$
Streuung	Varianz $\sigma^2 = E((x - \mu)^2)$	Inter-Quartil-Range $IQR = x_{0.75} - x_{0.25}$
Streuung	Varianz $\gamma = \frac{1}{\sigma^3} \cdot E((x - \mu)^3)$	Inter-Quartil-Range $g_{0.25} = \frac{(x_{0.75} - x_{0.5}) - (x_{0.5} - x_{0.25})}{x_{0.75} - x_{0.25}}$

4.4 Funktionen von Zufallsvariablen

Als Zufallsvariable wird allgemein eine Variable bezeichnet, die dem Ergebnis eines Zufallsexperiments eine reelle Zahl zuordnet. Die reelle Zahl kann über eine Funktion

$$y = g(x) \quad (4.76)$$

abgebildet werden. Es wird von einer Funktion der Zufallsvariable x gesprochen. Bild 4.5 verdeutlicht die Abbildung der Variable x auf die Variable y über die Funktion $g(x)$.

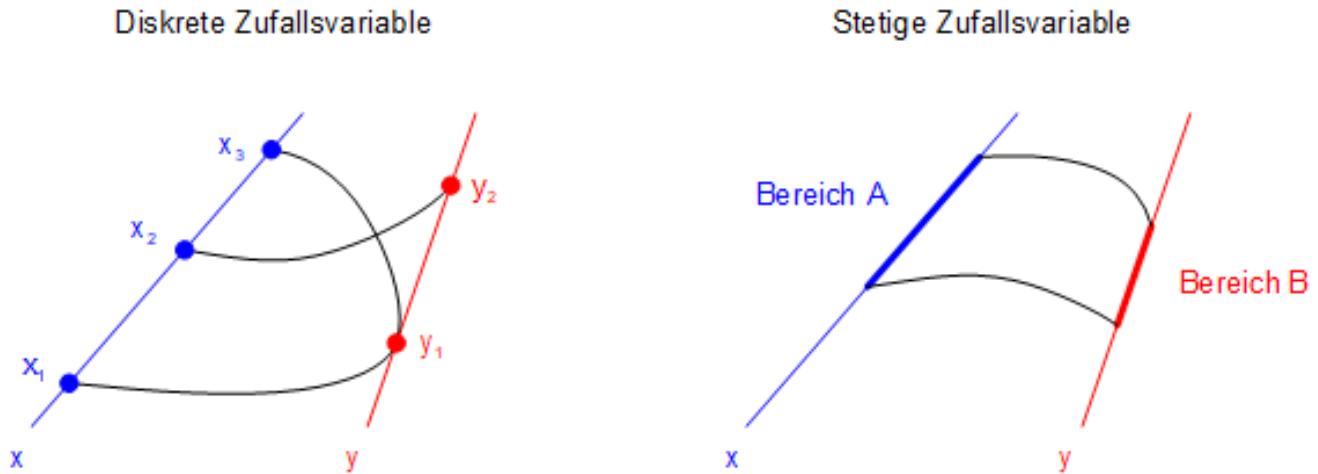


Bild 4.5: Grafische Veranschaulichung der Funktion von Zufallsvariable

Für die Berechnung praktischer Aufgabenstellungen zur Variable y ist es erforderlich, die Wahrscheinlichkeitsverteilung der Variable y zu kennen. Sie kann berechnet werden, wenn die Verteilungsfunktion der Variable x bekannt ist.

4.4.1 Funktion einer diskreten Zufallsvariable

Bei diskreten Zufallsvariablen wird einem Zufallereignis eine feste Zahl x_n zugeordnet. Durch eine Abbildung

$$y_n = g(x_n) \quad (4.77)$$

ändert sich an der Wahrscheinlichkeit des zugrunde liegenden Ereignisses nichts, sodass sich die Wahrscheinlichkeit von x_n auf y_n überträgt. Bild 4.5 macht deutlich, dass die Variable y_n möglicherweise durch unterschiedliche Variable x_n erreicht werden kann. Damit errechnet sich die Wahrscheinlichkeit der Zufallsvariable y_n ähnlich wie bei Ereignisbäumen aus der Summe

$$f_Y(y_n) = \sum_{y_n=g(x_n)} f_X(x_n) \quad (4.78)$$

Bei bekannter Wahrscheinlichkeitsverteilung $f_Y(y)$ errechnet sich die Verteilungsfunktion der Variable y definitionsgemäß zu

$$F_Y(y) = \sum_{y_n=-\infty}^y f_Y(y_n) \quad (4.79)$$

Beispiel: Abbildung einer diskreten Zufallsvariable

Gegeben ist eine Zufallsvariable x . Ihr Ereignisraum und ihre Wahrscheinlichkeitsverteilung sind in Tabelle 4.9 dargestellt. Die Zufallsvariable x wird über die Funktion

$$y = (x - 1)^2 \quad (4.80)$$

auf eine Zufallsvariable y abgebildet. Sie ist ebenfalls in Tabelle 4.9 aufgeführt.

Tabelle 4.9: Diskrete Zufallsvariable x und ihre Wahrscheinlichkeitsverteilung $f_X(x)$

x_n	-1	0	1	2	3
$f_X(x_n)$	0.1	0.2	0.4	0.2	0.1
y_n	4	1	0	1	4

Da zum Beispiel der Wert $y = 1$ über $x = 0$ und $x = 2$ erreicht werden kann, werden die entsprechenden Wahrscheinlichkeiten für das Ereignis $y = 1$ addiert.

$$f_Y(y = 1) = f_X(x = 0) + f_X(x = 2) = 0.2 + 0.2 = 0.4 \quad (4.81)$$

Es ergibt sich die in Tabelle 4.10 gezeigte Wahrscheinlichkeitsverteilung und Verteilungsfunktion der Zufallsvariable y .

Tabelle 4.10: Diskrete Zufallsvariable y und ihre Wahrscheinlichkeitsverteilung $f_Y(y)$

y_n	0	1	4
$f_Y(y_n)$	0.4	0.4	0.2
$F_Y(y_n)$	0.4	0.8	1

4.4.2 Funktion einer kontinuierlichen Zufallsvariable

Auch bei kontinuierlichen Zufallsvariablen wird einem Zufallsereignis eine Zahl x zugeordnet. Durch die Abbildung

$$y = g(x) \quad (4.82)$$

entsteht eine Zufallsvariable y , deren Verteilungsfunktion $F_Y(y)$ im Folgenden für streng monoton steigende Funktionen $g(x)$ hergeleitet wird. Die Verteilungsfunktion $F_Y(y)$ ist definiert als die Wahrscheinlichkeit, dass eine Variable ψ kleiner als der Wert y ist. Da die Zufallsvariable $\psi = g(\xi)$ definiert ist, gilt

$$yF_Y(y) = P(\psi \leq y) = P(g(\xi) \leq y) \quad (4.83)$$

Die Funktion $g(x)$ ist streng monoton steigend. Damit gilt:

$$F_Y(y) = P(\xi \leq g^{-1}(y)) = F_X(g^{-1}(y)) \quad (4.84)$$

Aus der Verteilungsfunktion $F_Y(y)$ errechnet sich die Wahrscheinlichkeitsdichte $f_Y(y)$ mit der Kettenregel der Differentiationsrechnung zu

$$f_Y(y) = \frac{dF_Y}{dy} = \frac{dF_X(g^{-1}(y))}{dg^{-1}(y)} \cdot \frac{dg^{-1}(y)}{dy} = f_X(g^{-1}(y)) \cdot \frac{dg^{-1}(y)}{dy} \quad (4.85)$$

Die Regel kann auf Funktionen $g(x)$ verallgemeinert werden, die streng monoton sind. Es ergibt sich

$$f_Y(y) = \frac{dF_Y}{dy} = \frac{dF_X(g^{-1}(y))}{dg^{-1}(y)} \cdot \frac{dg^{-1}(y)}{dy} = f_X(g^{-1}(y)) \cdot \left| \frac{dg^{-1}(y)}{dy} \right| \quad (4.86)$$

Beispiel: Abbildung einer kontinuierlichen Zufallsvariable

Gegeben ist eine Zufallsvariable x , die in dem Bereich $0 < x \leq 2$ eine Wahrscheinlichkeitsdichte

$$f_X(x) = \frac{1}{2} \cdot x \quad (4.87)$$

und Verteilungsfunktion

$$F_X(x) = \frac{1}{4} \cdot x^2 \quad (4.88)$$

aufweist. Beide Funktionen sind in Bild 4.6 dargestellt.

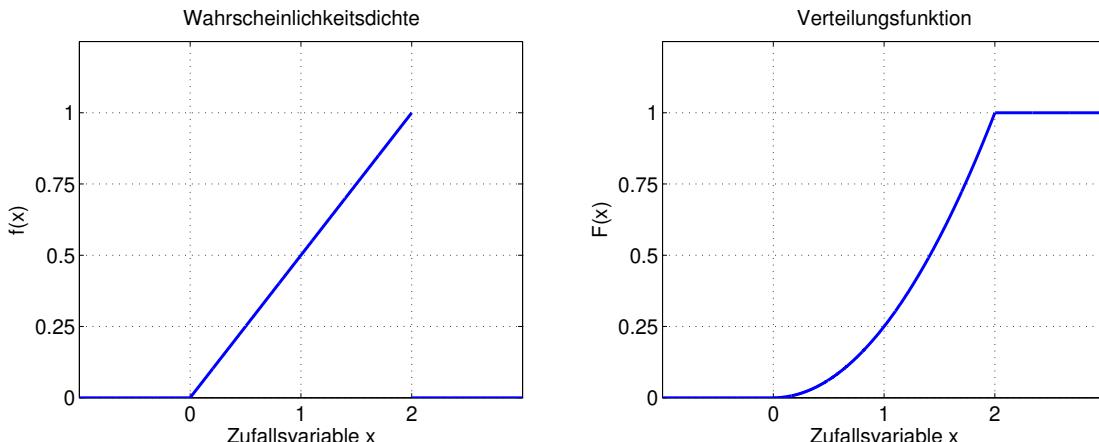


Bild 4.6: Wahrscheinlichkeitsdichte $f_X(x)$ und Verteilungsfunktion $F_X(x)$ der kontinuierlichen Zufallsvariable x

Die Zufallsvariable x wird mit der Funktion

$$y = g(x) = \frac{1}{2} \cdot x^2 \quad (4.89)$$

auf die Zufallsvariable y abgebildet. In dem Bereich von $0 < x \leq 2$ lautet die Umkehrfunktion

$$x = g^{-1}(y) = \sqrt{2 \cdot y} \quad (4.90)$$

Die Verteilungsfunktion der Variable y errechnet sich zu

$$F_Y(y) = F_X(g^{-1}(y)) = \frac{1}{4} \cdot x^2 \Big|_{x=\sqrt{2 \cdot y}} = \frac{1}{4} \cdot 2 \cdot y = \frac{1}{2} \cdot y \quad (4.91)$$

In dem relevanten Bereich von $0 < x \leq 2$ ist die Funktion $g(x)$ streng monoton steigend. Damit ergibt sich die Wahrscheinlichkeitsdichte

$$f_Y(y) = f_X(g^{-1}(y)) \cdot \left| \frac{d g^{-1}(y)}{dy} \right| = \frac{1}{2} \cdot x \Big|_{x=\sqrt{2 \cdot y}} \cdot \frac{d}{du} \sqrt{2 \cdot u} = \frac{1}{2} \cdot \sqrt{2 \cdot u} \cdot \sqrt{2} \cdot \frac{1}{2} \cdot \frac{1}{\sqrt{u}} = \frac{1}{2} \quad (4.92)$$

Wahrscheinlichkeitsdichte und Verteilungsfunktion der Zufallsvariable y sind in Bild 4.7 dargestellt.

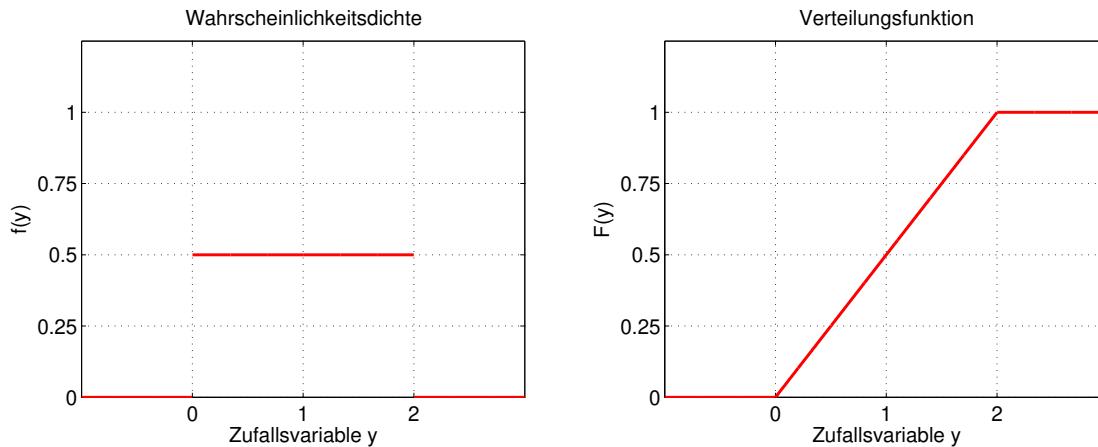


Bild 4.7: Wahrscheinlichkeitsdichte $f_Y(y)$ und Verteilungsfunktion $F_Y(y)$ der kontinuierlichen Zufallsvariable y

Die Abbildung von Zufallsvariablen hat zwei wichtige Anwendungen: die Standardisierung von Zufallsvariablen und die numerische Generierung von Zufallsvariablen mit einer definierten Verteilung.

4.4.3 Lineare Abbildung und Standardisierung einer Zufallsvariable

Eine Zufallsvariable y ist über die lineare Abbildung

$$y = a \cdot x + b \quad (4.93)$$

mit $a > 0$ definiert. Ihre Wahrscheinlichkeitsdichte errechnet sich nach Gleichung (4.86) zu

$$f_Y(y) = f_X\left(\frac{y-b}{a}\right) \cdot \left|\frac{1}{a}\right| \quad (4.94)$$

Die Zufallsvariable y besitzt einen Mittelwert von

$$\mu_y = E(y) = E(a \cdot x + b) = a \cdot E(x) + b = a \cdot \mu_x + b \quad (4.95)$$

und eine Varianz von

$$\begin{aligned} \sigma_y^2 &= E((y - \mu_y)^2) = E(y^2) - E(y)^2 = E((a \cdot x + b)^2) - E(a \cdot x + b)^2 \\ &= E(a^2 \cdot x^2 + 2 \cdot a \cdot b \cdot x + b^2) - (E(a \cdot x)^2 + 2 \cdot E(a \cdot x) \cdot E(b) + E(b)^2) \\ &= a^2 \cdot E(x^2) + 2 \cdot a \cdot b \cdot E(x) + b^2 - a^2 \cdot E(x)^2 - 2 \cdot a \cdot b \cdot E(x) - b^2 \\ &= a^2 \cdot (E(x^2) - E(x)^2) = a^2 \cdot \sigma_x^2 \end{aligned} \quad (4.96)$$

Die Standardabweichung der Zufallsvariablen y ergibt sich aus der positiven Wurzel der Varianz zu

$$\sigma_y = \sqrt{\sigma_y^2} = \sqrt{a^2 \cdot \sigma_x^2} = a \cdot \sigma_x \quad (4.97)$$

Ein Sonderfall der linearen Abbildung ist die Standardisierung von Zufallsvariablen. Standardisierte Zufallsvariablen weisen einen Mittelwert von $\mu_y = 0$ und eine Standardabweichung von $\sigma_y = 1$ auf. Aus diesen beiden Bedingungen ergeben sich zwei Gleichungen für die beiden unbekannten Variablen a und b :

$$\mu_y = a \cdot \mu_x + b = 0 \quad (4.98)$$

$$\sigma_y^2 = a^2 \cdot \sigma_x^2 = 1 \quad (4.99)$$

Aus den Bedingungen für a und b

$$a = \frac{1}{\sigma_x} \quad (4.100)$$

und

$$b = -\frac{\mu_x}{\sigma_x} \quad (4.101)$$

ergibt sich die standardisierte Zufallsvariable

$$y = a \cdot x + b = \frac{x}{\sigma_x} - \frac{\mu_x}{\sigma_x} = \frac{x - \mu_x}{\sigma_x} \quad (4.102)$$

Standardisierte Zufallsvariablen sind mittelwertfrei und weisen eine Varianz von 1 auf. Sie sind insbesondere im Zusammenhang von Testverteilungen von Bedeutung.

Beispiel: Glücksrad

Die Standardisierung von Zufallsvariablen wird an dem Beispiel des Glücksrads vertieft. Mit dem Mittelwert

$$\mu_x = \pi \quad (4.103)$$

und der Standardabweichung von

$$\sigma_x = \frac{1}{\sqrt{3}} \cdot \pi \quad (4.104)$$

ergibt sich die Abbildungsgleichung zu

$$y = \frac{x - \mu_x}{\sigma_x} = \frac{x - \pi}{\frac{1}{\sqrt{3}} \cdot \pi} = \sqrt{3} \cdot \frac{x - \pi}{\pi} \quad (4.105)$$

Durch die Transformation werden die Intervallgrenzen von 0 und 2π abgebildet auf

$$y_{\min} = \sqrt{3} \cdot \frac{0 - \pi}{\pi} = -\sqrt{3} \quad (4.106)$$

und

$$y_{\max} = \sqrt{3} \cdot \frac{2\pi - \pi}{\pi} = \sqrt{3} \quad (4.107)$$

Die Wahrscheinlichkeitsdichte ergibt sich in dem Zahlenbereich $y_{\min} < y \leq y_{\max}$ aus der Bedingung für das sichere Ereignis zu

$$f(y) = \frac{1}{2 \cdot \sqrt{3}} \quad (4.108)$$

Bild 4.8 stellt für das Glücksrad die Verteilung der Zufallsvariable x und der standardisierten Zufallsvariable y gegenüber.

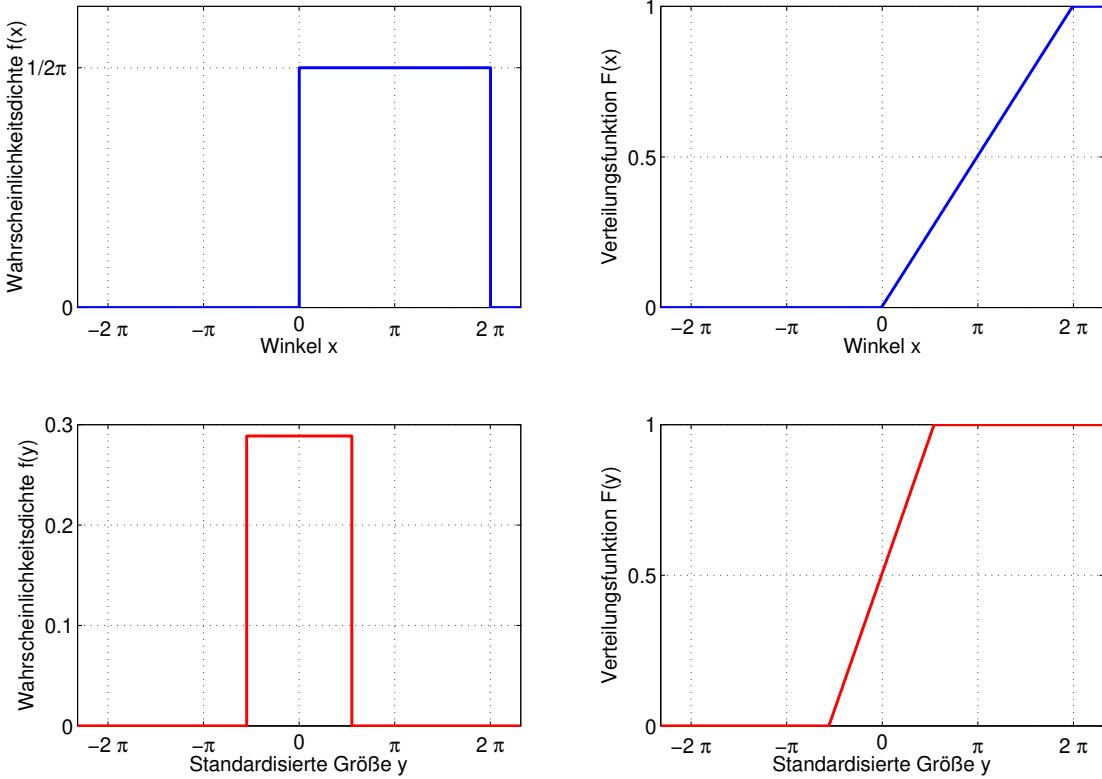


Bild 4.8: Grafische Darstellung der Wahrscheinlichkeitsverteilung für das Beispiel Glücksrad mit und ohne Standardisierung

Nach der Standardisierung der Zufallsvariable ist die Verteilung symmetrisch um den Mittelwert $\mu_y = 0$. Die Varianz ergibt sich erwartungsgemäß zu

$$\sigma_y^2 = \int_{-\infty}^{\infty} (y - \mu_y)^2 \cdot f(y) dy = \int_{-\sqrt{3}}^{\sqrt{3}} y^2 \cdot \frac{1}{2 \cdot \sqrt{3}} dy = \frac{1}{2 \cdot \sqrt{3}} \cdot \frac{1}{3} \cdot (3 \cdot \sqrt{3} + 3 \cdot \sqrt{3}) = 1 \quad (4.109)$$

4.4.4 Generierung von Zufallszahlen mit einer definierten Verteilung

In Programmen werden zur Generierung von Zufallszahlen Generatoren eingesetzt, die quasizufällig einen Wert $0 < x \leq 1$ erzeugen. Die Wahrscheinlichkeitsdichte dieser Zahlen ist typischerweise gleichverteilt.

$$f_X(x) = 1 \quad (4.110)$$

Die Zufallsvariable besitzt damit im Zahlenbereich $0 < x \leq 1$ die Verteilungsfunktion

$$F_X(x) = x \quad (4.111)$$

Um Zufallszahlen y mit einer beliebigen Verteilung $F_Y(y)$ zu generieren, wird die Funktion

$$y = F_Y^{-1}(x) \quad (4.112)$$

berechnet. Nach Gleichung (4.86) besitzt sie die Wahrscheinlichkeitsdichte

$$f_Y(g^{-1}(y)) \cdot \left| \frac{dg^{-1}(y)}{dy} \right| = 1 \cdot \left| \frac{dF(y)}{dy} \right| = f_Y(y) \quad (4.113)$$

4.5 Spezielle diskrete Verteilungen

Viele diskrete Fragestellungen der Wahrscheinlichkeitsrechnung lassen sich mit wenigen speziellen, diskreten Verteilungen beschreiben. Sie ergeben sich aus der Wahrscheinlichkeitstheorie und werden im Folgenden beschrieben.

4.5.1 Diskrete Gleichverteilung

Weisen bei einem Zufallsexperiment alle möglichen Werte x_n der Zufallsvariable x die gleiche Wahrscheinlichkeit p auf, ergibt sich aus der Bedingung für das sichere Ereignis

$$1 = \sum_{n=1}^N f(x_n) = \sum_{n=1}^N p = N \cdot p \quad (4.114)$$

Damit lautet die Wahrscheinlichkeitsverteilung

$$f(x) = p = \frac{1}{N} \quad (4.115)$$

und die Verteilungsfunktion berechnet sich zu

$$F(x) = \sum_{x_n=0}^x p = \sum_{x_n=0}^x \frac{1}{N} \quad (4.116)$$

Zum Beispiel sind bei einem Würfelexperiment mit einem regelmäßigen Würfel die Wahrscheinlichkeiten für eine beliebige Augenzahl mit

$$p = \frac{1}{6} \quad (4.117)$$

gleichverteilt. Die Wahrscheinlichkeitsverteilung und Verteilungsfunktion der Gleichverteilung sind in Bild 4.9 dargestellt. Die Wahrscheinlichkeitsverteilung $f(x)$ ist konstant, die Verteilungsfunktion $F(x)$ nimmt an jeder Stelle, an der ein möglicher Wert der Zufallsvariable steht, zu.

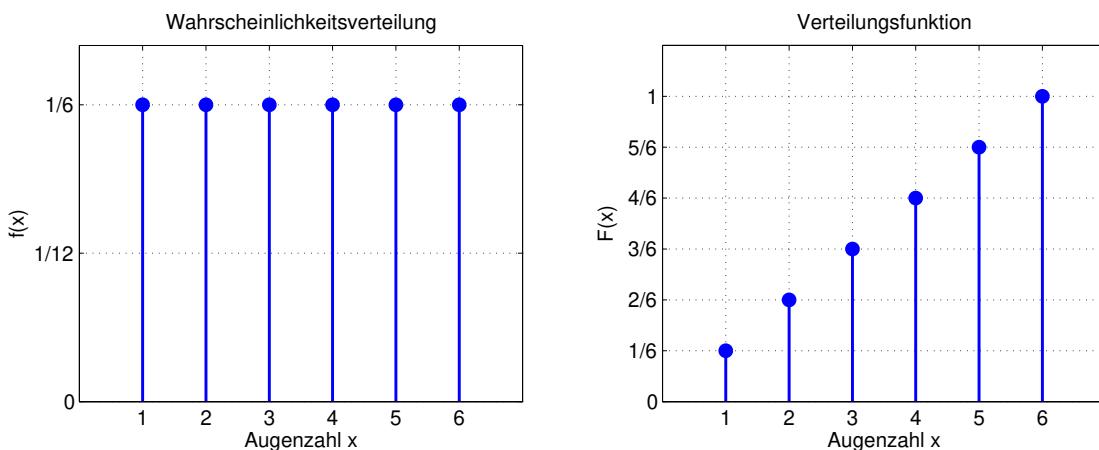


Bild 4.9: Grafische Darstellung der Wahrscheinlichkeitsverteilung $f(x)$ und Verteilungsfunktion $F(x)$ für das Würfeln einer Augenzahl x

Der Mittelwert der Gleichverteilung errechnet sich aus

$$\mu = E(x) = \frac{1}{N} \cdot \sum_{n=1}^N x_n \quad (4.118)$$

und die Varianz ergibt sich zu

$$\sigma^2 = E((x - \mu)^2) = \frac{1}{N} \cdot \sum_{n=1}^N (x_n - \mu)^2 \quad (4.119)$$

Beispiel: Bewertung zweier Anlagemöglichkeiten

Dem Kunden einer Bank werden zur Geldanlage zwei mögliche Anlagestrategien angeboten. Als konventionelle Variante steht eine Festgeldanlage mit einer festen Verzinsung von jährlich 3.33 % zur Verfügung. Als weitere Anlagevariante wird dem Kunden ein Modell auf Basis von Wertpapieren angeboten. Dabei wird in Abhängigkeit eines Wirtschaftsindex ein Zinssatz von 1.5, 2.5 oder 5 % ausgezahlt. Da über den Verlauf des Wirtschaftsindex für das Anlagejahr keine Vorhersagen getroffen werden können, haben alle drei möglichen Zinsbeträge eine gleich hohe Wahrscheinlichkeit, sie sind somit gleichverteilt. Bild 4.10 zeigt die resultierende Wahrscheinlichkeitsverteilung $f(x)$ und die Verteilungsfunktion $F(x)$.

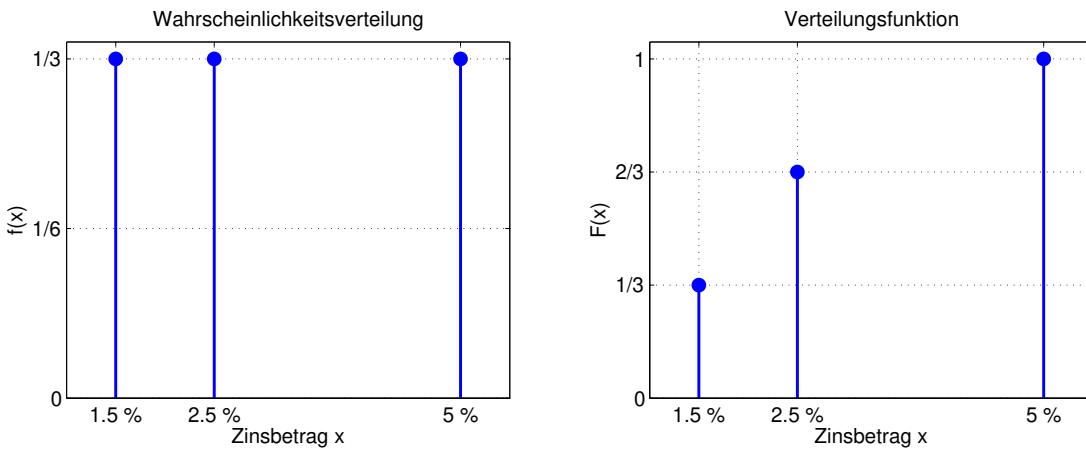


Bild 4.10: Grafische Darstellung der Wahrscheinlichkeitsverteilung $f(x)$ und Verteilungsfunktion $F(x)$ für die Verzinsung bei Wertpapieren

Um zu entscheiden, bei welcher Anlagevariante mit höheren Zinserträgen gerechnet werden kann, wird der Erwartungswert beziehungsweise der Mittelwert der zweiten Anlagevariante bestimmt. Dieser folgt zu

$$\mu = \frac{1}{N} \cdot \sum_{n=1}^N x_n = \frac{1}{3} \cdot (1.5 + 2.5 + 5) = 3\% \quad (4.120)$$

Im Mittel wird der Kunde bei der Anlagevariante auf Basis des Wirtschaftsindex eine jährliche Verzinsung von 3 % erhalten. Da die mittlere Zinserwartung bei höherem Risiko geringer ist als bei der Festgeld-Anlage, ist die Festgeld-Anlage zu bevorzugen.

Die Erstellung von Bild 4.10 und die Berechnung des Mittelwertes wurde mit MATLAB durchgeführt.

```

1 % Variablendefinition
2 x = [1.5 2.5 5];
3
4 % Berechnung der Wahrscheinlichkeitsverteilung und der Verteilungsfunktion
5 p = unidpdf(1:3,3);
6 P = unidcdf(1:3,3);
7
8 % Grafische Darstellung
9 figure(1);
10 subplot(1,2,1);
11 stem(x,p);
12 subplot(1,2,2);
13 stem(x,P);
14
15 % Berechnung des Mittelwertes
16 mu = mean(x);
```

Alternativ kann die Umsetzung in Python erfolgen.

```

1 Bibliotheken importieren
2 import numpy as np
3 import matplotlib.pyplot as plt
4 from scipy.stats import randint
5
6 Definition der Stützstellen
7 X = [1.5, 2.5, 5]
8 Xp = np.linspace(1,3,3)
9 low, high = 1, 4
10
11 Berechnung der Wahrscheinlichkeitsverteilung
12 p = randint.pmf(Xp, low, high)
13 P = randint.cdf(Xp, low, high)
14
15 Grafische Darstellung
16 fig = plt.figure(1, figsize=(12, 4))
17 ax1, ax2 = fig.subplots(1,2)
18
19 ax1.plot(X, p, 'bo', ms=8, label='randint pmf')
20 ax1.vlines(X, 0, p, colors='b', lw=1, alpha=1)
21 ax1.grid(True, which='both', axis='both', linestyle='--')
22 ax1.axis([1, 5.5, 0, 0.35])
23 ax1.set_xticks(X)
24 ax1.set_yticks([0, 1/3])
25 ax1.set_yticklabels(['0', '1/3'])
26 ax1.set_xlabel('Zinsbetrag x / %')
27 ax1.set_ylabel('f(x)')
28 ax1.set_title('Wahrscheinlichkeitsverteilung')
29
30 ax2.plot(X, P, 'bo', ms=8, label='randint pmf')
31 ax2.vlines(X, 0, P, colors='b', lw=1, alpha=1)
32 ax2.grid(True, which='both', axis='both', linestyle='--')
33 ax2.axis([1, 5.5, 0, 1.05])
34 ax2.set_xticks(X)
```

```
35 | ax2.set_yticks([1/3, 2/3, 1])
36 | ax2.set_yticklabels(['1/3', '2/3', '1'])
37 | ax2.set_xlabel('Zinsbetrag x / %')
38 | ax2.set_ylabel('F(x)')
39 | ax2.set_title('Verteilungsfunktion')
40 |
41 Berechnung Mittelwert
42 mu = np.mean(X)
43 print(' ')
44 print('Arithmetischer Mittelwert: ', mu)
```

4.5.2 Bernoulli-Verteilung

Ein Spezialfall diskreter Verteilungen ist die Bernoulli-Verteilung, die ein Bernoulli Experiment beschreibt. Dabei wird geprüft, ob ein Ereignis bei einfacher Ausführung des Experiments eingetreten ist oder nicht. Die Zufallsvariable x ist damit definiert über

$$x = \begin{cases} 1 & \text{für das günstige Ereignis A} \\ 0 & \text{für das ungünstige Ereignis A'} \end{cases} \quad (4.121)$$

Eine solche Variable wird als binäre oder Bernoulli-Variable bezeichnet. Ist die Wahrscheinlichkeit für das günstige Ereignis A

$$P(A) = f(1) = p \quad (4.122)$$

errechnet sich die Wahrscheinlichkeit für das ungünstige Ereignis A' zu

$$P(A') = f(0) = 1 - p = q \quad (4.123)$$

Die Wahrscheinlichkeitsverteilung hat nur die beiden Werte p und $q = 1 - p$. Der Mittelwert der Bernoulli-Verteilung errechnet sich zu

$$\mu = E(x) = \sum_{n=1}^2 x_n \cdot P(x_n) = 0 \cdot q + 1 \cdot p = p \quad (4.124)$$

und die Varianz beträgt

$$\sigma^2 = E((x - \mu)^2) = E(x^2) - \mu^2 = p - p^2 = p \cdot q \quad (4.125)$$

Beispiele für Bernoulli-Experimente sind die Funktion von Bauelementen oder das Erfüllen von Spezifikationsmerkmalen. Bernoulli-Variablen erlauben jedoch nur eine grobe Kategorisierung von Ereignissen, eine feine Einteilung zum Beispiel nach der Frage, wie weit ein Grenzwert überschritten wurde, können mit dem Bernoulli-Experiment nicht beantwortet werden.

Beispiel: Bit-Fehler bei der digitalen Signalübertragung

Als Anwendungsbeispiel für die Bernoulli-Verteilung wird die Übertragung eines stochastischen 1-Bit-Binärsignals betrachtet, das den Zustand 0 oder 1 annehmen kann. Bei der Übertragung des Bits wird mit einem Störsignal gerechnet, das eine Änderung des Zustandes mit einer Wahrscheinlichkeit von $p = 0.25$ bewirkt.

Zur Darstellung wird die Zufallsvariable x definiert, die den Wert 1 annimmt, wenn das Bit während der Übertragung verfälscht wurde und die den Wert 0 annimmt, wenn keine Verfälschung vorliegt. Damit ergibt sich die Wahrscheinlichkeitsfunktion $f(x)$ der Zufallsvariablen x zu

$$f(x) = \begin{cases} \frac{1}{4} & \text{für } x = 1 \\ \frac{3}{4} & \text{für } x = 0 \end{cases} \quad (4.126)$$

Die Wahrscheinlichkeitsverteilung $f(x)$ und die Verteilungsfunktion $F(x)$ sind in Bild 4.11 dargestellt.

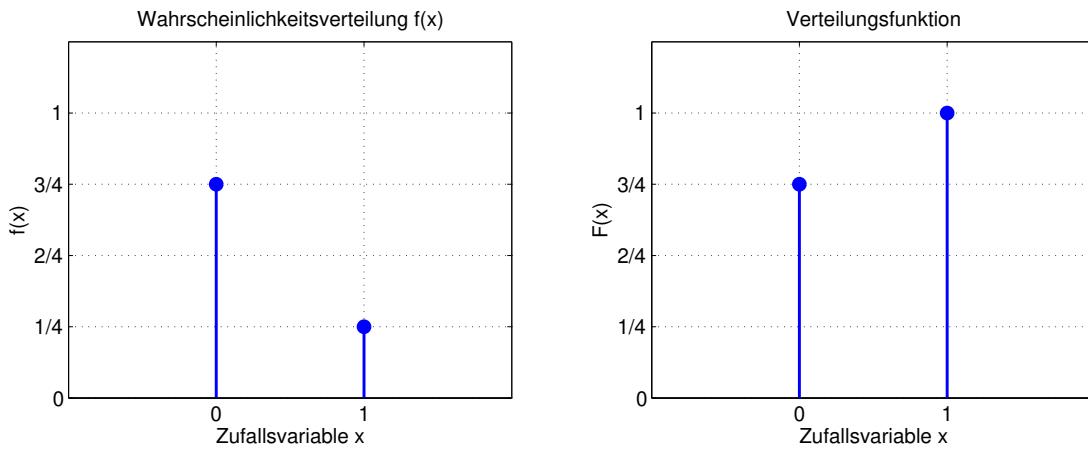


Bild 4.11: Grafische Darstellung der Wahrscheinlichkeitsverteilung $f(x)$ und Verteilungsfunktion $F(x)$ einer Bernoulli-Verteilung einer 1-Bit-Binärsignal-Übertragung

4.5.3 Binomial-Verteilung

Die Binomial-Verteilung ist eine Erweiterung der Bernoulli-Verteilung. Sie gibt an, wie viele günstige Ereignisse stattfinden, wenn ein Zufallsexperiment N -fach ausgeführt wird. Die Wahrscheinlichkeit für das Ereignis A ist bei jeder einzelnen Durchführung konstant

$$P(A) = p \quad (4.127)$$

Damit ergibt sich bei jeder einzelnen Durchführung des Zufallsexperiments für das inverse Ereignis A' die Wahrscheinlichkeit

$$P(A') = 1 - p = q \quad (4.128)$$

Bei einfacher Ausführung ($N = 1$) kann die Zufallsvariable x nur die Werte 0 oder 1 annehmen. In dem Fall ergibt sich die Wahrscheinlichkeitsverteilung

$$f(x) = p^x \cdot q^{1-x} \quad (4.129)$$

Wird das Zufallsexperiment N -fach ausgeführt, tritt das Ereignis A x -fach ein. Die Anzahl unterschiedlicher Anordnungen errechnet sich als Kombinationen ohne Wiederholungen zu (N über x). Die Wahrscheinlichkeitsverteilung bei N -facher Ausführung des Zufallsexperiments ergibt sich damit aus dem Produkt der Realisierungsmöglichkeiten mit der Wahrscheinlichkeit $f(x)$ des betreffenden Ereignisses. Sie folgt daher zu

$$f(x) = \binom{N}{x} \cdot p^x \cdot q^{N-x} \quad (4.130)$$

Gleichung (4.130) beschreibt die Wahrscheinlichkeit, dass ein Ereignis A bei N unabhängigen Ausführungen des Experiments x -mal eintritt, wenn das Ereignis A bei Einzelausführung die konstante Wahrscheinlichkeit p besitzt und die Wahrscheinlichkeit des nicht Eintreffens $q = 1 - p$ ist. Diese Verteilung wird als Binomial-Verteilung bezeichnet und ist in Bild 4.12 für $N = 16$ und unterschiedliche Wahrscheinlichkeiten p dargestellt. In Abhängigkeit von N und p besitzt die Wahrscheinlichkeitsverteilung $f(x)$ ein Maximum.

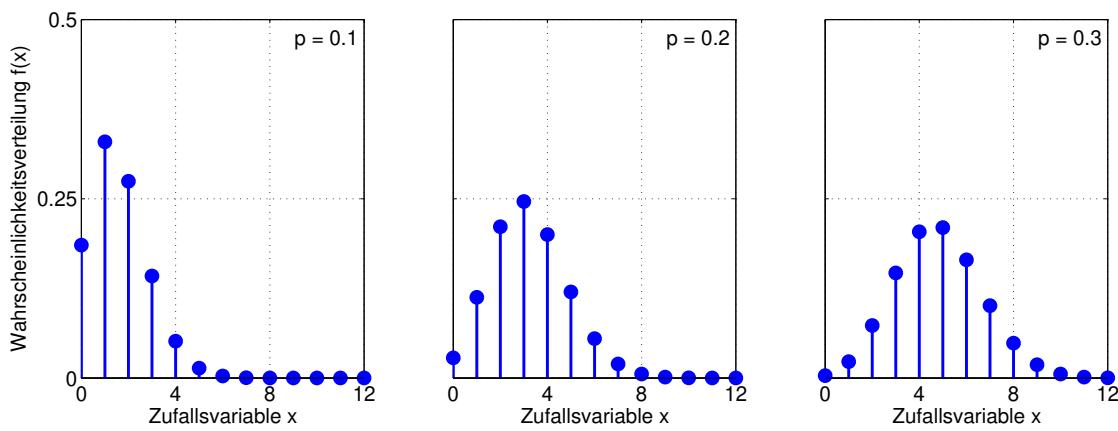


Bild 4.12: Wahrscheinlichkeitsverteilung der Binomial-Verteilung mit $N = 16$ für unterschiedliche Erfolgswahrscheinlichkeiten p

Die Verteilungsfunktion $F(x)$ ergibt sich aus der Summe über die Wahrscheinlichkeitsverteilung zu

$$F(x) = \sum_{x_n=0}^x \binom{N}{x_n} \cdot p^{x_n} \cdot q^{N-x_n} \quad (4.131)$$

Bild 4.13 stellt die Wahrscheinlichkeitsverteilung $f(x)$ und die Verteilungsfunktion $F(x)$ der Binomial-Verteilung für $N = 16$ und $p = 0.2$ dar.

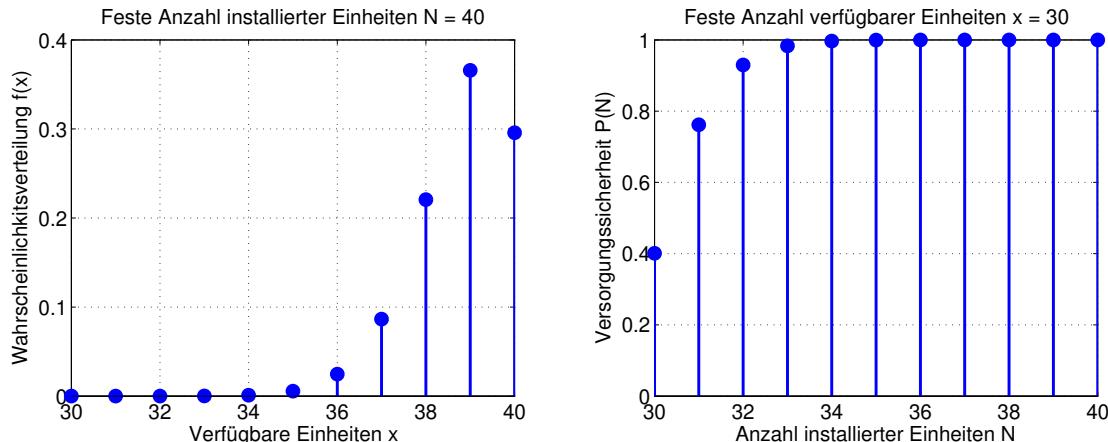


Bild 4.13: Grafische Darstellung der Wahrscheinlichkeitsverteilung $f(x)$ und Verteilungsfunktion $F(x)$ einer Binomial-Verteilung mit $p = 0.2$

Durch Auswerten der momenterzeugenden Funktion ergeben sich Mittelwert und Varianz der Binomial-Verteilung zu

$$\mu = N \cdot p \quad (4.132)$$

und

$$\sigma^2 = N \cdot p \cdot q \quad (4.133)$$

Mit der Einführung von Funktionen mehrerer Zufallsvariablen wird sich zeigen, dass sich Mittelwert und Varianz der Binomial-Verteilung aus der Summe von N Mittelwerten beziehungsweise N Varianzen der Bernoulli-Verteilung ergeben.

Beispiel: Versorgungssicherheit

Eine wichtige Aufgabe der Energieversorgungsunternehmen ist die Versorgung mit einer fiktiven Sicherheit von 99.9 % sicherzustellen. Um bei dem Ausfall einzelner Einspeiseeinheiten ausreichend elektrische Leistung zur Verfügung stellen zu können, werden Reserveeinheiten vorgehalten. Als Grundlage für die Berechnung dieses Beispiels wird ein fiktives Verbundnetz mit $N = 30$ erforderlichen Einspeiseeinheiten und einer unabhängigen Verfügbarkeitswahrscheinlichkeit jeder Einheit von $p = 97\%$ angenommen. Die Wahrscheinlichkeit für die Verfügbarkeit von x der N installierten Einheiten wird durch die Binomial-Verteilung beschrieben.

$$f(x) = \binom{N}{x} \cdot p^x \cdot (1-p)^{N-x} \quad (4.134)$$

Bild 4.14 zeigt links die Verfügbarkeitswahrscheinlichkeit von x Einheiten bei $N = 40$ installierten Einheiten. Es müssen mindestens 30 Einheiten verfügbar sein, es dürfen aber auch mehr sein. Damit ergibt sich die gesuchte Wahrscheinlichkeit zu

$$P(x \geq 30) = 1 - F(29) = 1 - \sum_{x_n=0}^{29} \binom{N}{x_n} \cdot 0.97^{x_n} \cdot 0.03^{N-x_n} \geq 0.999 \quad (4.135)$$

Die Anzahl installierter Einheiten N muss so groß sein, dass sich die Versorgungssicherheit von 99.9 % erreicht wird. Bild 4.14 zeigt im rechten Bildteil die Versorgungssicherheit als Funktion der installierten Einheiten N .

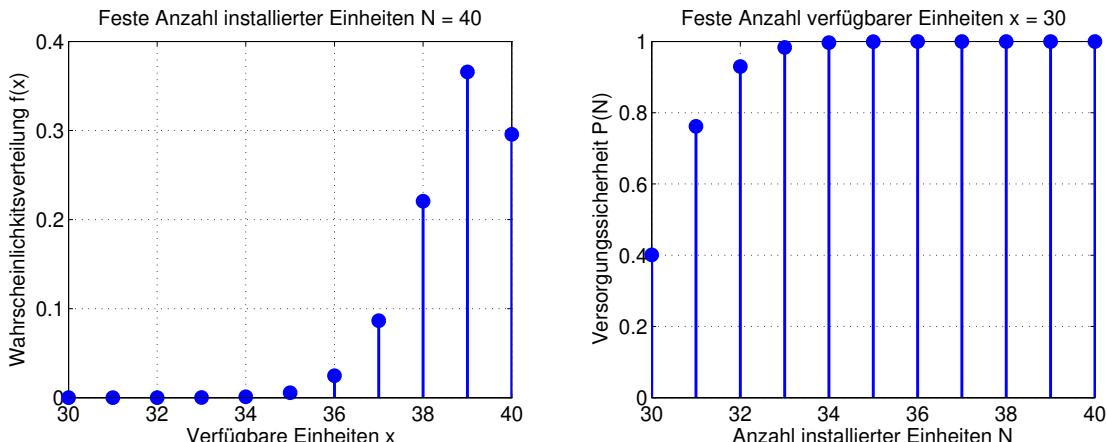


Bild 4.14: Grafische Darstellung der Verfügbarkeitswahrscheinlichkeit von x bei $N = 40$ installierten Einheiten und der Versorgungssicherheit als Funktion der installierten Einheiten N

Eine numerische Auswertung zeigt, dass mindestens 35 Einheiten installiert werden müssen, um die Versorgungssicherheit mit einer Wahrscheinlichkeit von 99.9 % sicherzustellen.

Mit MATLAB kann die Versorgungssicherheit für das definierte Verbundnetz berechnet werden.

```

1 % Variablendefinition
2 p = 0.97;
3 N = 30:40;
4
5 % Berechnung der Versorgungssicherheit
6 S = 1 - binocdf(30,N,p);
7
8 % Grafische Darstellung
9 stem(N,S);
```

Alternativ kann die Versorgungssicherheit in Python umgesetzt werden.

```

1 Bibliotheken importieren
2 import numpy as np
3 import matplotlib.pyplot as plt
4 from scipy.stats import binom
5
6 Variablendefinition
7 p = 0.97
8 N = np.linspace(30,40,11);
9 S = 1 - binom.cdf(29,N,p);
10
11 Grafische Darstellung
12 fig = plt.figure(1, figsize=(6, 4))
13 ax1 = fig.subplots(1,1)
14
15 ax.plot(N, S, 'bo', ms=8, label='randint pmf')
16 ax.vlines(N, 0, S, colors='b', lw=1, alpha=1)

```

Ein weiteres typisches Anwendungsbeispiel für die Binomial-Verteilung ist die Qualitätskontrolle. Sind zum Beispiel in einem Behälter M Schrauben und sind davon G Schrauben defekt, so ist die Wahrscheinlichkeit p , eine defekte Schraube zu ziehen

$$p = \frac{G}{M} \quad (4.136)$$

Nach dem Ziehen wird die Schraube wieder zurückgelegt. Damit ändert sich die Wahrscheinlichkeit p nicht und die Wahrscheinlichkeit, bei N Zügen mit Zurücklegen genau x defekte Schrauben zu ziehen, berechnet sich aus

$$f(x) = \binom{N}{x} \cdot \left(\frac{G}{M}\right)^x \cdot \left(1 - \frac{G}{M}\right)^{n-x} \quad (4.137)$$

Praktisch gesehen ist das Ziehen ohne Zurücklegen wichtiger als das Ziehen mit Zurücklegen. Die mathematische Beschreibung mithilfe der hypergeometrischen Verteilung ist jedoch anspruchsvoller, da sich die Erfolgswahrscheinlichkeit p während des Experimentes ändert.

4.5.4 Hypergeometrische Verteilung

In Abschnitt 4.5.3 wird die Binomial-Verteilung zur Beschreibung von Zufallsexperimenten mit gleichbleibender Wahrscheinlichkeit p vorgestellt. Werden zum Beispiel bei der Eingangskontrolle die überprüften Gegenstände nicht zurückgelegt, ändert sich die Wahrscheinlichkeit p im Laufe des Zufallsexperiments. Die Wahrscheinlichkeitsverteilung bei n -facher Ausführung des Zufallsexperiments kann in diesem Fall mit der hypergeometrischen Verteilung beschrieben werden.

Zur Herleitung der Wahrscheinlichkeitsverteilung wird davon ausgegangen, dass sich in einem Behälter M Elemente befinden, von denen G Elemente defekt sind. Es wird eine Stichprobe von N Elementen ausgewählt. Die Reihenfolge ist nicht relevant. Nach den Gleichungen zu Kombinationen aus Kapitel 2 kann aus einer Grundgesamtheit von M Elementen eine Teilmenge von N Elementen auf (M über N) Möglichkeiten gewählt werden. Entsprechend lassen sich aus G defekten Elementen x defekte Elemente auf (G über x) Möglichkeiten auswählen und $G - x$ brauchbare Elemente lassen sich aus $M - G$ brauchbaren Elementen auf (($M - G$) über ($G - x$)) Weisen ziehen.

Damit wird die Wahrscheinlichkeit, bei einer N-fachen Ziehung ohne Zurücklegen x defekte Elemente zu finden, mit der Wahrscheinlichkeitsverteilung

$$f(x) = \frac{\binom{G}{x} \cdot \binom{M-G}{N-x}}{\binom{M}{N}} \quad (4.138)$$

beschrieben. Die Verteilungsfunktion $F(x)$ der hypergeometrischen Verteilung ergibt sich durch Summation der Wahrscheinlichkeitsverteilung.

$$F(x) = \sum_{x_n=-\infty}^x \left(\frac{\binom{G}{x_n} \cdot \binom{M-G}{N-x_n}}{\binom{M}{N}} \right) \quad (4.139)$$

Der Mittelwert der hypergeometrischen Verteilung beträgt

$$\mu = N \cdot \frac{G}{M} \quad (4.140)$$

und entspricht dem Mittelwert der Binomial-Verteilung. Die Varianz der hypergeometrischen Verteilung ergibt sich aus

$$\sigma^2 = \frac{N \cdot G \cdot (M-G) \cdot (M-N)}{M^2 \cdot (M-1)} \quad (4.141)$$

Sowohl die Wahrscheinlichkeitsverteilung $f(x)$ als auch die Verteilungsfunktion $F(x)$ der Hypergeometrischen Verteilung mit den Parametern $G = 8$, $M = 16$ und $N = 8$ ist in Bild 4.15 abgebildet.

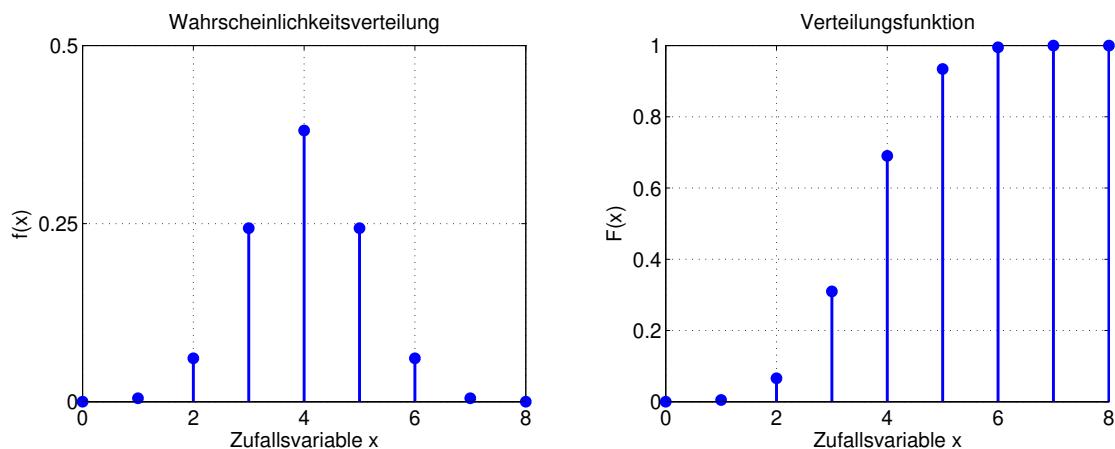


Bild 4.15: Darstellung der Wahrscheinlichkeitsverteilungen $f(x)$ und der Verteilungsfunktion $F(x)$ für die Hypergeometrische Verteilung mit $G = 8$, $M = 16$ und $N = 8$

Beispiel: Qualitätssicherung

Die hypergeometrische Verteilung wird mit einem Beispiel zur Qualitätssicherung vertieft. Eine Firma liefert Dichtungen in Packungen zu je 100 Stück. Eine Packung darf laut Liefervertrag höchstens 10 % Ausschuss enthalten. Jede Packung wird geprüft, indem 10 Stück zufällig und ohne Zurücklegen entnommen werden. Sind diese 10 Stück alle einwandfrei, wird die Packung angenommen. Andernfalls wird sie zurückgewiesen. Es soll die Frage untersucht werden, wie groß bei diesem Prüfverfahren die Wahrscheinlichkeit ungerechtfertigter Reklamationen ist, bei der eine Packung zurückgewiesen wird, obwohl sie den Lieferbedingungen entspricht.

Bei höchstens 10 % Ausschuss darf eine Packung mit $M = 100$ Dichtungen maximal $G = 10$ fehlerhafte Stücke enthalten, um gerade noch den Spezifikationen zu entsprechen. Zur Berechnung der Wahrscheinlichkeit einer ungerechtfertigten Zurückweisung wird zunächst die Wahrscheinlichkeit dafür bestimmt, dass sich unter $N = 10$ Ringen kein fehlerhaftes Stück befindet. Mit der Hypergeometrischen Verteilung ergibt sich diese zu

$$f(0) = \frac{\binom{10}{0} \cdot \binom{90}{10}}{\binom{100}{10}} = 0.33 \quad (4.142)$$

Die Wahrscheinlichkeit für eine ungerechtfertigte Reklamation beträgt

$$P = 1 - f(0) = 1 - 0.33 = 0.67 \quad (4.143)$$

Sie ist in diesem Fall mit 67 % sehr hoch und ist deshalb für eine qualifizierte Eingangskontrolle nicht zielführend. Die Berechnung der Wahrscheinlichkeit erfolgt mit MATLAB durch folgendes Programm.

```

1 % Variablendefinition
2 M = 100;
3 G = 10;
4 N = 10;
5
6 % Berechnung der Wahrscheinlichkeit
7 f0 = hygepdf(0,M,G,N);
8 P = 1 - f0;
```

Alternativ kann die Eingangskontrolle in Python berechnet werden.

```

1 Bibliotheken importieren
2 from scipy.stats import hypergeom
3
4 Variablendefinition
5 M = 100
6 G = 10
7 N = 10
8
9 Berechnung der Wahrscheinlichkeit
10 P = 1 - hypergeom.pmf(0,M,G,N)
11 print(' ')
12 print('Wahrscheinlichkeit für eine ungerechtfertigte Reklamation : ', P)
```

Die Hypergeometrische Verteilung kann unter bestimmten Bedingungen in die Binomial-Verteilung überführt werden. Bild 4.16 vergleicht die Binomial-Verteilung aus Abschnitt 4.5.3 mit der Hypergeometrischen Verteilung für unterschiedliche Verhältnisse der Anzahl G an Gutteilen zur Gesamtmenge M und dem Umfang der Stichprobe N . Die Approximation der hypergeometrischen Verteilung durch die Binomial-Verteilung verbessert sich mit sinkendem Verhältnis des Stichprobenumfangs N zur Grundgesamtheit M und sinkender Erfolgswahrscheinlichkeit $p = G/M$.

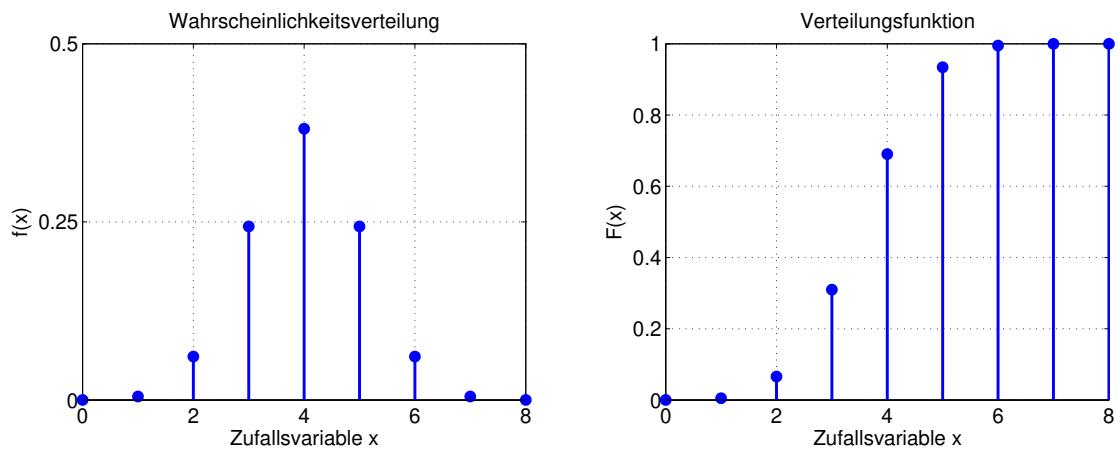


Bild 4.16: Grafischer Vergleich der Wahrscheinlichkeitsverteilungen für die Binomial- und die hypergeometrische Verteilung

4.5.5 Poisson-Verteilung

Ist die konstante Erfolgswahrscheinlichkeit p bei einem einzelnen Experiment klein und die Anzahl N der Ausführungen sehr groß, wird das Rechnen mit der Binomial-Verteilung wegen der Binomial-Koeffizienten aufwendig. Für den Fall $p \rightarrow 0$ kann die Binomial-Verteilung durch die Poisson-Verteilung approximiert werden. Dabei soll der Mittelwert

$$\mu = N \cdot p \quad (4.144)$$

erhalten bleiben. Für die Erfolgswahrscheinlichkeit p und die Wahrscheinlichkeit des Misserfolgs q gilt mit dieser Annahme

$$p = \frac{\mu}{N} \quad (4.145)$$

beziehungsweise

$$q = 1 - p = 1 - \frac{\mu}{N} \quad (4.146)$$

Durch Einsetzen dieser Beziehungen in die Definitionsgleichung der Binomial-Verteilung aus Gleichung (4.113) ergibt sich

$$f(x) = \binom{N}{x} \cdot p^x \cdot q^{N-x} = \binom{N}{x} \cdot \left(\frac{\mu}{N}\right)^x \cdot \left(1 - \frac{\mu}{N}\right)^{N-x} \quad (4.147)$$

Für den Grenzübergang $N \rightarrow \infty$ ergibt sich die Poisson-Verteilung [Krey91] mit der Wahrscheinlichkeitsverteilung

$$f(x) = \frac{\mu^x}{x!} \cdot e^{-\mu} \quad (4.148)$$

und der Verteilungsfunktion

$$F(x) = \sum_{x_n=0}^x \frac{\mu^{x_n}}{x_n!} \cdot e^{-\mu} \quad (4.149)$$

Bild 4.17 vergleicht die Binomial- und die Poisson-Verteilung für unterschiedliche Erfolgswahrscheinlichkeiten p und Stichprobenumfänge N .

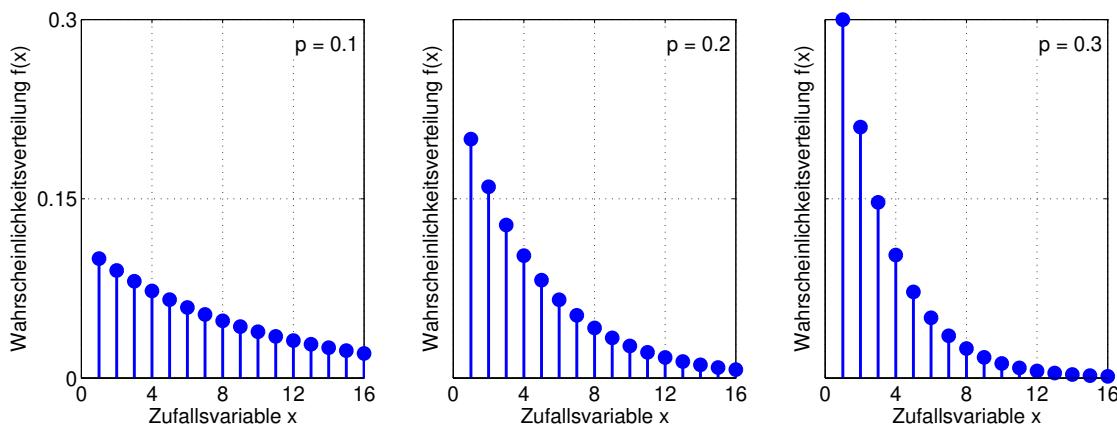


Bild 4.17: Grafischer Vergleich der Wahrscheinlichkeitsverteilung $f(x)$ für die Binomial- und die Poisson-Verteilung

Die Approximation der Binomial-Verteilung durch die Poisson-Verteilung verbessert sich mit steigendem Umfang N der Stichprobe und sinkender Erfolgswahrscheinlichkeit p . Der Mittelwert ist definiertengemäß identisch zu dem Mittelwert der Binomial-Verteilung

$$\mu = N \cdot p \quad (4.150)$$

Durch Auswerten der momenterzeugenden Funktion ergibt sich die Varianz der Poisson-Verteilung zu

$$\sigma^2 = N \cdot p = \mu \quad (4.151)$$

Beispiel: Fertigung von Widerständen

Als Beispiel wird die Fertigung von Widerständen untersucht. Die Widerstände mit einem Nennwert von 50Ω werden in Packungen zu je 100 Stück geliefert. Dabei wird die Garantie gegeben, dass alle Widerstände zwischen 45Ω und 55Ω liegen. Die Wahrscheinlichkeit, einen Widerstand zu produzieren, der nicht zwischen 45Ω und 55Ω liegt, beträgt erfahrungsgemäß nur 2. Es soll die Frage untersucht werden, wie groß die Wahrscheinlichkeit ist, dass eine bestimmte Packung diese Zusage erfüllt.

Die Wahrscheinlichkeit ergibt sich aus der Binomialverteilung mit

$$P(x=0) = \binom{N}{x} \cdot p^x \cdot q^{n-x} = \binom{100}{0} \cdot 0.002^0 \cdot 0.998^{100-0} = 0.8186 \quad (4.152)$$

Das Ergebnis kann auch mit der Poisson-Verteilung für

$$\mu = N \cdot p = 100 \cdot 0.002 = 0.2 \quad (4.153)$$

berechnet werden

$$P(x=0) = \frac{\mu^x}{x!} \cdot e^{-\mu} = \frac{0.2^0}{0!} \cdot e^{-0.2} = 0.8187 \quad (4.154)$$

Die Ergebnisse stimmen in guter Näherung überein, da das Experiment mit 100 Wiederholungen ausgeführt wird und die Wahrscheinlichkeit, ein Teil zu finden, das nicht innerhalb der Spezifikation liegt, mit 2. sehr gering ist.

Mit MATLAB berechnet sich die Wahrscheinlichkeit $P(x=0)$ durch

```

1 % Variablendefinition
2 x = 0;
3 N = 100;
4 p = 0.002;
5 mu = N*p;
```

```

6
7 % Berechnung der Wahrscheinlichkeit mit der Binomial-Verteilung
8 P_binomial = binocdf(x,N,p);
9
10 % Berechnung der Wahrscheinlichkeit mit der Poisson-Verteilung
11 P_poisson = poisscdf(x,mu);

```

In Python wird die Wahrscheinlichkeit mit folgendem Programmausschnitt berechnet.

```

1 Bibliotheken importieren
2 from scipy.stats import poisson
3
4 Variablendefinition
5 x = 0
6 N = 100
7 p = 0.002
8 mu = N*p
9
10 Berechnung der Wahrscheinlichkeit mit der Poisson-Verteilung
11 P = poisson.cdf(x,mu)
12 print(' ')
13 print('Wahrscheinlichkeit für keinen Ausfall: ', P)

```

4.5.6 Geometrische Verteilung

Wird ein Bernoulli-Experiment mehrfach ausgeführt, kann die Frage gestellt werden, wie oft das Experiment wiederholt werden muss, bis zum ersten Mal das günstige Ereignis eintritt. Die Wahrscheinlichkeit für ein Ereignis A ist wie beim Bernoulli-Experiment definiert als

$$P(A) = p \quad (4.155)$$

und die Wahrscheinlichkeit für das Ereignis A' als

$$P(A') = 1 - p = q \quad (4.156)$$

Der Parameter p entspricht der Wahrscheinlichkeit, direkt beim ersten Zug das Ereignis A zu bekommen. Die Wahrscheinlichkeit, erst bei der x-ten Ausführung des Experiments ein günstiges Ereignis zu erhalten, ergibt sich aus der Wahrscheinlichkeit, x - 1 ungünstige Ereignisse zu erhalten und ein günstiges Ereignis. Damit folgt die Wahrscheinlichkeitsverteilung f(x) zu

$$f(x) = p \cdot (1 - p)^{x-1} = p \cdot q^{x-1} \quad (4.157)$$

Die Wahrscheinlichkeitsverteilung f(x) aus Gleichung (4.157) wird als geometrische Verteilung bezeichnet. Sie hängt von der Wahrscheinlichkeit p für das günstige Ereignis A ab. Bild 4.18 zeigt die geometrische Verteilung für verschiedene Wahrscheinlichkeiten p. Dabei ist gut zu erkennen, dass die Wahrscheinlichkeitsverteilung f(x) exponentiell abnimmt.

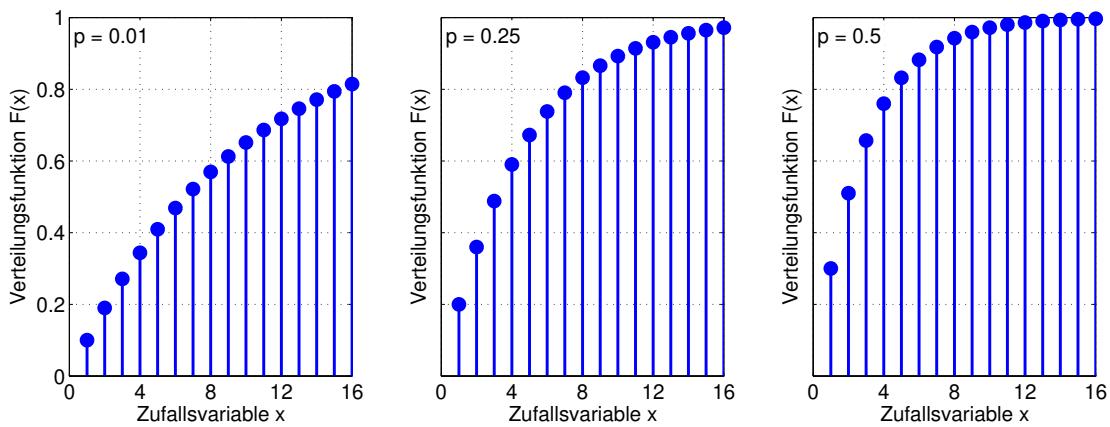


Bild 4.18: Grafische Darstellung der Wahrscheinlichkeitsverteilung $f(x)$ einer geometrischen Verteilung für unterschiedliche Erfolgswahrscheinlichkeiten p

Die Verteilungsfunktion $F(x)$ ergibt sich aus der Summe von $f(x)$. Da mindestens ein Ereignis stattfinden muss, startet die Summe bei eins.

$$F(x) = \sum_{x_n=1}^x p \cdot (1-p)^{x_n-1} = \sum_{x_n=1}^x p \cdot q^{x_n-1} \quad (4.158)$$

Im Fall der geometrischen Verteilung ist die Summe unendlich, da es theoretisch unendlich lange dauern kann, bis ein günstiges Ereignis stattfindet. Der Grenzwert von $F(x)$ ergibt sich für $x \rightarrow \infty$ zu

$$\lim_{x \rightarrow \infty} (F(x)) = \sum_{x_n=1}^{\infty} p \cdot q^{x_n-1} = \frac{p}{q} \cdot \sum_{x_n=1}^{\infty} q^{x_n} = \frac{p}{q} \cdot \left(\frac{1}{1-q} - 1 \right) = 1 \quad (4.159)$$

Das Ergebnis entspricht dem zweiten Axiom der Wahrscheinlichkeit. Die in Bild 4.19 dargestellte Verteilungsfunktion $F(x)$ für verschiedene Wahrscheinlichkeiten p zeigt das Verhalten nach Gleichung (4.159), sie nähert sich exponentiell der Asymptote 1 an.

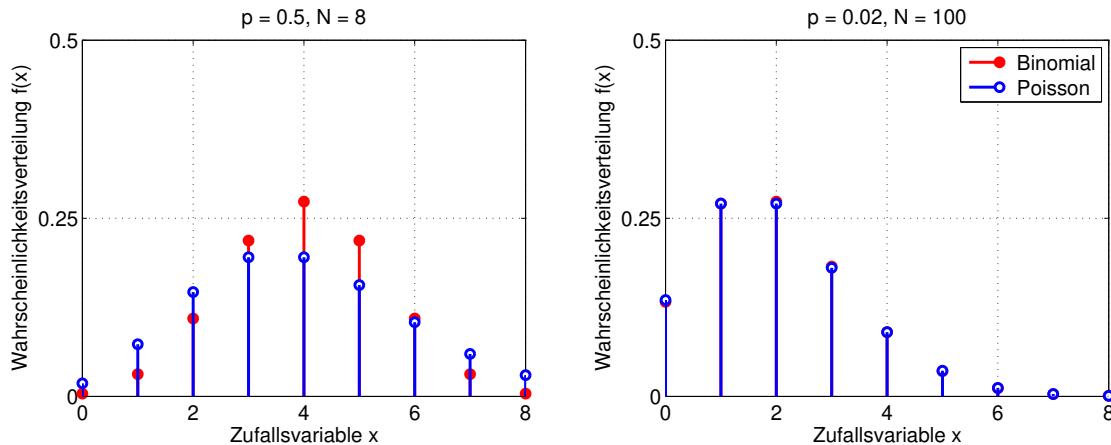


Bild 4.19: Grafische Darstellung der Wahrscheinlichkeitsverteilung $f(x)$ einer geometrischen Verteilung für unterschiedliche Erfolgswahrscheinlichkeiten p

Der Mittelwert der geometrischen Verteilung errechnet sich zu

$$\mu = \sum_{x_n=1}^{\infty} x_n \cdot (1-p)^{x_n-1} \cdot p = \frac{1}{p} \quad (4.160)$$

und die Varianz beträgt

$$\sigma^2 = \frac{1}{N-1} \cdot \sum_{n=1}^N (x_n - \mu)^2 \cdot (1-p)^{x_n-1} \cdot p = \frac{1-p}{p^2} \quad (4.161)$$

Geometrische Verteilungen werden angewendet, wenn die Wahrscheinlichkeit für Wartezeiten angegeben werden soll, bis zu der ein Ereignis eintritt. Ein wichtiges Anwendungsgebiet ist das Abschätzen von Wahrscheinlichkeiten für Ausfälle von Geräten und das Abschätzen von Fehlerraten bei der Datenübertragung.

Beispiel: Lebensdauer eines optischen Signalgebers

Ein optischer Signalgeber wird in einem festen Zeitabstand eingeschaltet. Die Wahrscheinlichkeit, dass der Signalgeber beim Einschalten ausfällt, beträgt 0.02 %. Die Wahrscheinlichkeit, dass der optische Signalgeber nicht während des Einschaltens ausfällt, kann vernachlässigt werden. Mithilfe der geometrischen Verteilung soll die Lebensdauer als Anzahl der Einschaltzyklen des Signalgebers bestimmt werden. Bild 4.20 zeigt die Wahrscheinlichkeitsverteilung $f(x)$ und die Verteilungsfunktion $F(x)$ der Anzahl von Schaltzyklen bis zum Ausfall des optischen Signalgebers.

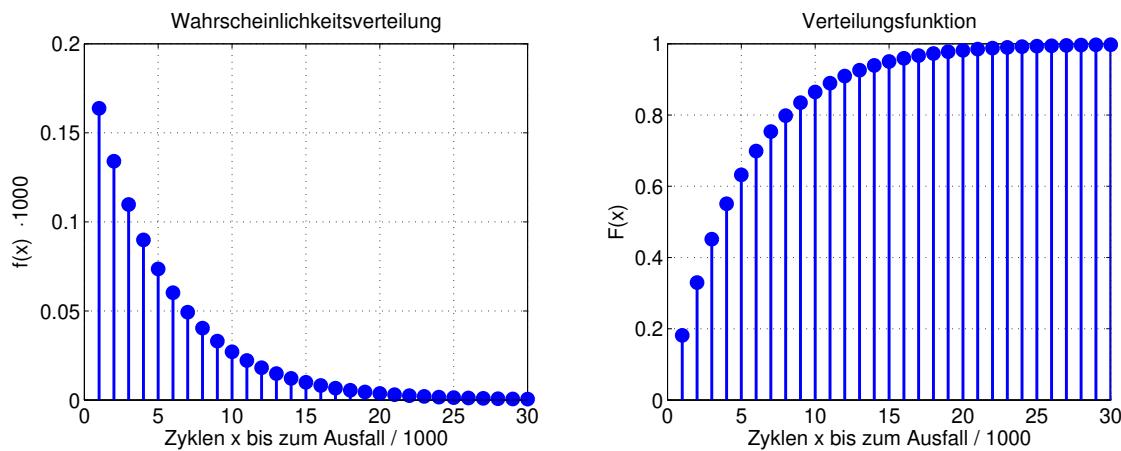


Bild 4.20: Grafische Darstellung der Wahrscheinlichkeitsverteilung $f(x)$ und Verteilungsfunktion $F(x)$ der Anzahl von Schaltzyklen bis zum Ausfall

Der Mittelwert der geometrischen Verteilung aus Bild 4.20Bild 4.20 berechnet sich zu

$$\mu = \frac{1}{p} = \frac{1}{0.0002} = 5000 \quad (4.162)$$

Die mittlere Lebensdauer des optischen Signalgebers liegt somit bei 5000 Einschaltzyklen. Die Wahrscheinlichkeit, dass der optische Signalgeber nicht mehr als 5000 Zyklen funktioniert, kann mit der Verteilungsfunktion berechnet werden zu

$$F(5000) = \sum_{x=1}^N p \cdot (1-p)^{x-1} = \sum_{x=1}^{5000} 0.0002 \cdot (1-0.0002)^{x-1} = 63.22\% \quad (4.163)$$

Die Berechnung erfolgte durch die im Folgenden dargestellte MATLAB-Sequenz.

```

1 % Variablendefinition
2 p0 = 0.0002;
3
4 % Berechnung des Mittelwertes und der Wahrscheinlichkeit F(Mittelwert)
5 mu = 1/p0;
6 P_5000 = geocdf(mu, p0);

```

In Python ergibt sich folgender Programmausschnitt.

```

1 Bibliotheken importieren
2 from scipy.stats import geom
3
4 Variablendefinition
5 p0 = 0.0002
6 mu = 1/p0
7
8 Berechnung der Wahrscheinlichkeit mit der Poisson-Verteilung
9 P = geom.cdf(mu, p0);
10 print(' ')
11 print('Wahrscheinlichkeit für defekt innerhalb 5000 Zyklen: ', P)

```

Zusammenfassung der diskreten Verteilungen

Im vorangegangenen Abschnitt werden spezielle diskrete Verteilungen vorgestellt und deren Anwendung an einem Beispiel verdeutlicht. Die Zusammenhänge zwischen der Bernoulli-, der Binomial- und der Poisson- sowie der Hypergeometrischen Verteilung sind nochmals in Bild 4.21 grafisch dargestellt.

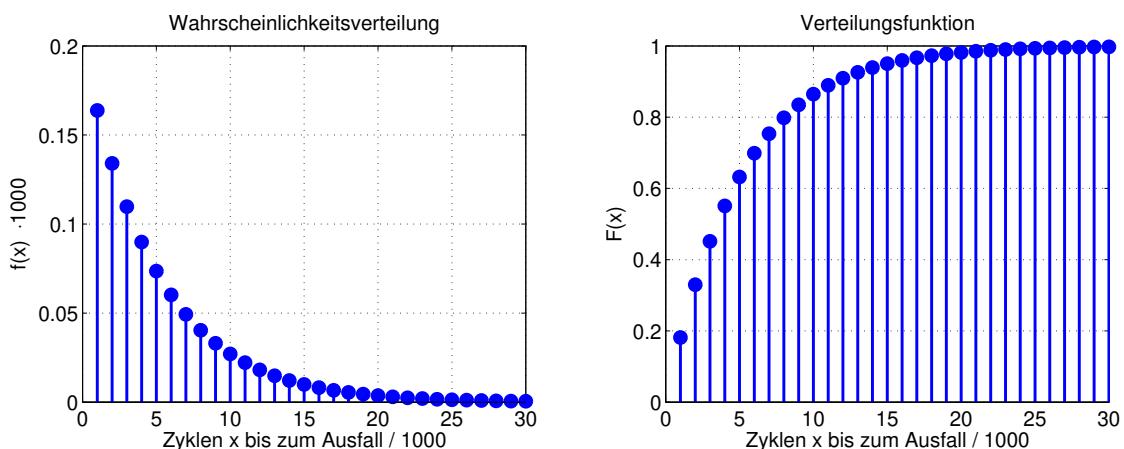


Bild 4.21: Zusammenhang diskreter Verteilungen für die Anzahl x für das Eintreffen eines Ereignisses A

Tabelle 4.11 gibt eine Übersicht über die diskutierten diskreten Wahrscheinlichkeitsverteilungen und ihren Anwendungen.

Tabelle 4.11: Übersicht über diskrete Wahrscheinlichkeitsverteilungen

Name und Anwendung	Wahrscheinlichkeitsverteilung	Kenngrößen μ und σ^2
Gleichverteilung: Gleiche Wahrscheinlichkeit für alle Ereignisse	$f(x) = p = \frac{1}{N}$	$\mu = \frac{1}{N} \cdot \sum_{n=1}^N x_n$ $\sigma^2 = \frac{1}{N} \cdot \sum_{n=1}^N (x_n - \mu)^2$
Bernoulli-Verteilung: Günstiges / ungünstiges Ereignis bei einfacher Ausführung des Experimentes	$f(x) = \begin{cases} p & \text{für } x = 1 \\ q = 1 - p & \text{für } x = 0 \end{cases}$	$\mu = p$ $\sigma^2 = p \cdot q$
Binomial-Verteilung: Anzahl günstiger Ereignisse bei N-facher Ausführung, konstante Erfolgswahrscheinlichkeit	$f(x) = \binom{N}{x} \cdot p^x \cdot q^{N-x}$	$\mu = N \cdot p$ $\sigma^2 = N \cdot p \cdot q$
Hypergeometrische Verteilung: Anzahl günstiger Ereignisse bei N-facher Ausführung, variable Erfolgswahrscheinlichkeit	$f(x) = \frac{\binom{N}{x} \cdot \binom{M-G}{G-x}}{\binom{M}{N}}$	$\mu = N \cdot \frac{G}{M}$ $\sigma^2 = \frac{N \cdot G \cdot (M-G) \cdot (M-N)}{M^2 \cdot (M-1)}$
Poisson-Verteilung: Approximation der Binomialverteilung	$f(x) = \frac{\mu^x}{x!} \cdot e^{-\mu}$	$\mu = N \cdot p$ $\sigma^2 = N \cdot p = \mu$
Geometrische Verteilung: Anzahl von Wiederholungen des Experimentes, bis das günstige Ereignis eintritt	$f(x) = p \cdot q^{x-1}$	$\mu = \frac{1}{p}$ $\sigma^2 = \frac{1-p}{p^2}$

Sowohl die Wahrscheinlichkeitsverteilung als auch die Verteilungsfunktionen der in Tabelle 4.11 aufgelisteten Verteilungen sind bei MATLAB in der Statistic Toolbox implementiert. Dabei werden die folgenden Endungen für die MATLAB-Funktionen eingesetzt:

- pdf Wahrscheinlichkeitsverteilung $f(x)$ (probability density function)
- cdf Verteilungsfunktion $F(x)$ (cumulative distribution function)

- `inv` Inverse Verteilungsfunktion $F^{-1}(x)$ (inverse cumulative distribution function)
- `rnd` Zufallszahlen-Generator einer Verteilung

Tabelle 4.12 gibt eine Übersicht über eine Auswahl von MATLAB-Funktionen zu diskreten Verteilungen.

Tabelle 4.12: Übersicht über diskrete Wahrscheinlichkeitsverteilungen in MATLAB

Verteilung	Wahrscheinlichkeitsverteilung $f(x)$	Verteilungsfunktion $F(x)$	inverse Verteilungsfunktion $F^{-1}(x)$	Zufallszahlen-generator
Gleichverteilung	unidpdf(x,N)	unidcdf(x,N)	unidinv(P,N)	unidrnd(N,m,n)
Binomial-Verteilung	binopdf(x,N,p)	binocdf(x,N,p)	binoinv(Y,N,p)	binornd(N,p,m,n)
Hypergeometrische Verteilung	hygepdf(x,M,G,N)	hygecdf(x,M,G,N)	hygeinv(P,M,G,N)	hygernd(M,G,N,m,n)
Poisson-Verteilung	poisspdf(x, μ)	poisscdf(x, μ)	poissinv(P, μ)	poissrnd(μ ,m,n)
Geometrische Verteilung	geopdf(x,p)	geocdf(x,p)	geoinv(P,p)	geornd(p,m,n)

Tabelle 4.13 gibt eine Übersicht über eine Auswahl von Python-Funktionen der Bibliothek `scipy.stats` zu diskreten Verteilungen.

Tabelle 4.13: Übersicht über diskrete Wahrscheinlichkeitsverteilungen der Python Bibliothek `scipy.stats`

Verteilung	Wahrscheinlichkeitsverteilung $f(x)$	Verteilungsfunktion $F(x)$	inverse Verteilungsfunktion $F^{-1}(x)$	Zufallszahlen-generator
Gleichverteilung	randint.pmf	randint.cdf	randint.ppf	randint.rvs
Binomial-Verteilung	binom.pmf	binom.cdf	binom.ppf	binom.rvs
Hypergeometrische Verteilung	hypergeom.pmf	hypergeom.cdf	hypergeom.ppf	hypergeom.rvs
Poisson-Verteilung	poisson.pmf	poisson.cdf	poisson.ppf	poisson.rvs
Geometrische Verteilung	geom.pmf	geom.cdf	geom.ppf	geom.rvs

4.6 Spezielle stetige Verteilungen

Die Beschreibung stetiger Zufallsgrößen erfolgt durch stetige Verteilungen. Dazu sind in den folgenden Abschnitten spezielle stetige Verteilungen aufgeführt. Einen besonderen Status nehmen Test- oder Prüfverteilungen ein, denen der Abschnitt 4.7 gewidmet ist.

4.6.1 Gleichverteilung

Analog zu den diskreten Verteilungen wird als einfaches Beispiel die Gleichverteilung definiert. Sie weist in einem Intervall von a bis b eine konstante Wahrscheinlichkeitsdichte $f(x) = k f_0$ auf, außerhalb dieses Intervalls ist die Wahrscheinlichkeitsdichte null. Die Wahrscheinlichkeit für das sichere Ereignis ist

$$F(\infty) = \int_{-\infty}^{\infty} f(x) dx = f_0 \cdot (b - a) = 1 \quad (4.164)$$

Damit ergibt sich für die Wahrscheinlichkeitsdichte $f(x)$

$$f(x) = \begin{cases} \frac{1}{b-a} & \text{für } a < x \leq b \\ 0 & \text{für alle anderen Werte} \end{cases} \quad (4.165)$$

Durch Integration ergibt sich eine Verteilungsfunktion von

$$F(x) = \int_a^x \frac{1}{b-a} d\xi = \frac{x-a}{b-a} \quad \text{für } a < x \leq b \quad (4.166)$$

Bild 4.22 stellt die Wahrscheinlichkeitsdichte $f(x)$ und Verteilungsfunktion $F(x)$ einer stetigen Gleichverteilung dar.

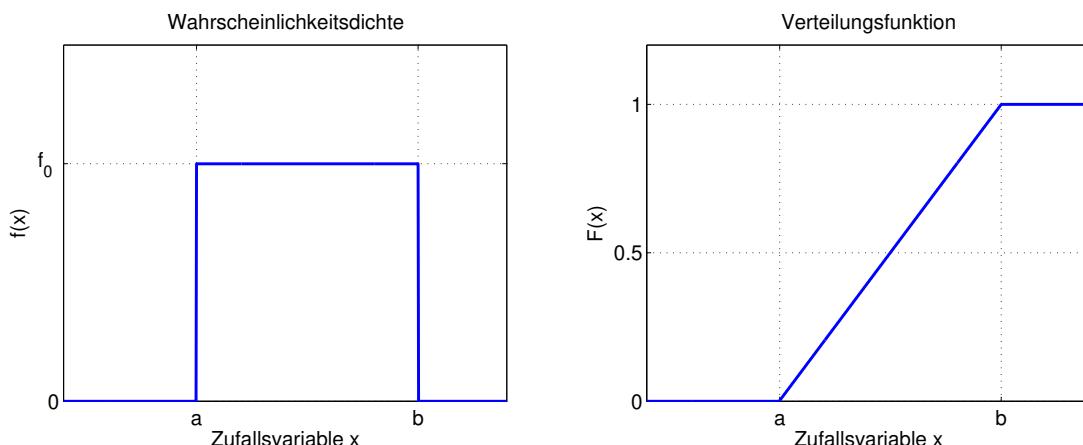


Bild 4.22: Grafische Darstellung der Wahrscheinlichkeitsdichte $f(x)$ und der Verteilungsfunktion $F(x)$ für eine Gleichverteilung

Der Mittelwert einer Gleichverteilung ergibt sich zu

$$\mu = E(x) = \int_{-\infty}^{\infty} x \cdot f(x) dx = \int_a^b x \cdot \frac{1}{b-a} dx = \frac{b^2 - a^2}{2 \cdot (b-a)} = \frac{a+b}{2} \quad (4.167)$$

Die Varianz einer Gleichverteilung berechnet sich zu

$$\sigma^2 = E((x - \mu)^2) = E(x^2) - E^2(x) = \frac{b^3 - a^3}{3 \cdot (b - a)} - \frac{(a + b)^2}{4} = \frac{(b - a)^2}{12} \quad (4.168)$$

Typische Anwendungsgebiete sind die Berechnung von durchschnittlichen Wartezeiten bei Transportprozessen oder Bedienungssystemen sowie die statistische Beschreibung von Diskretisierungsvorgängen.

Beispiel: Effektivwert des Quantisierungsrauschen eines A/D-Wandlers

Um ein analoges Eingangssignal U_E in ein digitales Ausgangssignal U_{ADC} zu wandeln, wird das zeitkontinuierliche Signal quantisiert. Die Quantisierung in diskrete Amplitudenwerte erfolgt dabei durch einen Analog-Digital-Wandler mit einer Auflösung von N Bit. Durch die damit festgelegte, endliche Anzahl von 2^N Quantisierungsstufen entsteht ein Fehler, der als Quantisierungsrauschen q aufgefasst werden kann. Bild 4.233 zeigt zwei Quantisierungsstufen und einen Eingangsspannungswert U_E .

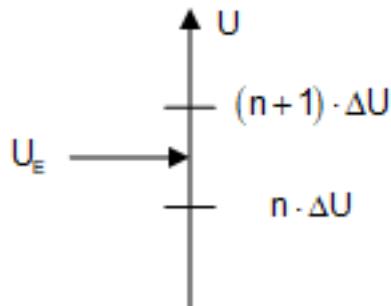


Bild 4.23: Quantisierung eines Eingangssignals U_E durch einen Analog-Digital-Wandler

Die Höhe einer Quantisierungsstufe ΔU wird bei einem Analog-Digital-Wandler durch die Anzahl der Quantisierungsstufen und den Messbereich U_{MAX} zu

$$\Delta U = \frac{U_{MAX}}{2^N} \quad (4.169)$$

definiert. Der Signalwert U_E wird durch den Analog-Digital-Wandler entweder auf den Wert $n \cdot \Delta U$ oder auf den Wert $(n + 1) \cdot \Delta U$ quantisiert. Dadurch ist der Fehler q in dem Intervall $-\Delta U/2 \leq q \leq \Delta U/2$ gleichverteilt mit der Wahrscheinlichkeitsdichte

$$f_0 = \frac{1}{\Delta U} \quad (4.170)$$

Zur Bewertung des Fehlers wird oft der Effektivwert des Quantisierungsrauschen herangezogen. Der Effektivwert einer Zufallsgröße ist allgemein als die Wurzel aus dem Erwartungswert des Quadrates der Zufallszahl definiert. Für das Quantisierungsrauschen folgt mit den Rechenregeln zum Erwartungswert

$$q_{EFF}^2 = E(q^2) = \sigma_q^2 + E(q)^2 = \sigma_q^2 + \mu_q^2 \quad (4.171)$$

Für die Gleichverteilung des Quantisierungsrauschen berechnet sich der Mittelwert zu

$$\mu_q = \frac{1}{2} \cdot \left(\frac{-\Delta U}{2} + \frac{\Delta U}{2} \right) = 0 \quad (4.172)$$

Die Varianz folgt zu

$$\sigma_q^2 = \frac{1}{12} \cdot \left(\frac{\Delta U}{2} - \frac{-\Delta U}{2} \right)^2 = \frac{\Delta U^2}{12} \quad (4.173)$$

Damit ergibt sich der Effektivwert des Quantisierungsrauschen eines Analog-Digital-Wandlers zu

$$q_{EFF} = \sqrt{\sigma_q^2} = \frac{\Delta U}{\sqrt{12}} \approx \frac{\Delta U}{3.46} \quad (4.174)$$

Der Effektivwert des Quantisierungsrauschen eines Analog-Digital-Wandlers kann damit aus den technischen Daten ermittelt werden.

4.6.2 Dreiecksverteilung

Liegen nur sehr wenige konkrete Daten zur Bestimmung der Verteilungsfunktion $f(x)$ einer Zufallsvariablen x vor, wird in der Praxis die Dreiecksverteilung als erste Schätzung verwendet. Die Dreiecksverteilung ist auch als Simpson-Verteilung bekannt. Ihre Wahrscheinlichkeitsverteilung ist in dem Intervall $a < x \leq b$ definiert durch

$$f(x) = \begin{cases} \frac{2}{(b-a) \cdot (c-a)} \cdot (x-a) & \text{für } a < x \leq c \\ \frac{2}{(b-a) \cdot (b-c)} \cdot (b-x) & \text{für } c < x \leq b \end{cases} \quad (4.175)$$

Außerhalb des Intervalls $a < x \leq b$ ist die Wahrscheinlichkeit null. In Gleichung (4.175) beschreibt der Parameter a den kleinsten und der Parameter b den größten Wert der Zufallsvariablen x , der Parameter c gibt die Lage des Maximums an. Durch Integration ergibt sich die Verteilungsfunktion $F(x)$ aus Gleichung (4.175) zu

$$F(x) = \begin{cases} \frac{(x-a)^2}{(b-a) \cdot (c-a)} & \text{für } a < x \leq c \\ 1 - \frac{(b-x)^2}{(b-a) \cdot (b-c)} & \text{für } c < x \leq b \end{cases} \quad (4.176)$$

Bild 4.24 stellt die Wahrscheinlichkeitsdichte $f(x)$ und Verteilungsfunktion $F(x)$ dar.

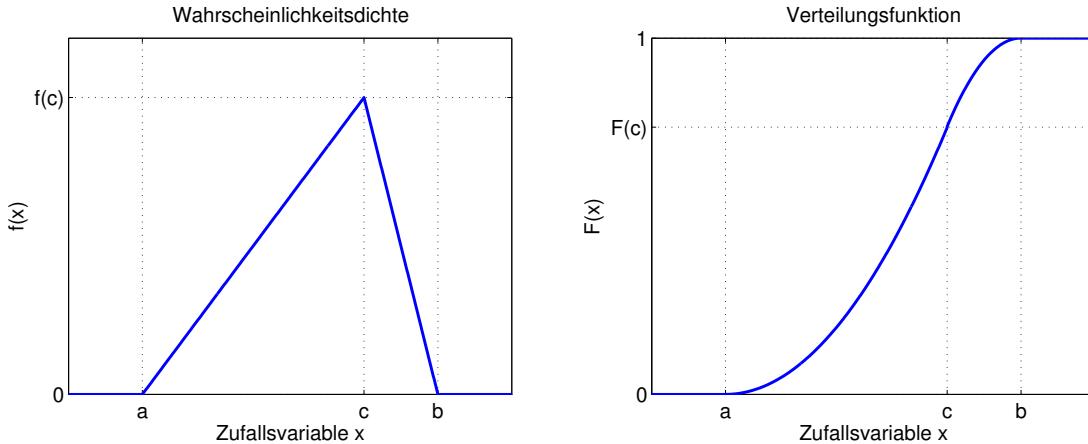


Bild 4.24: Wahrscheinlichkeitsdichte $f(x)$ und Verteilungsfunktion $F(x)$ für eine stetige Dreiecksverteilung

An der Stelle c besitzt die Wahrscheinlichkeitsverteilung $f(x)$ ihr Maximum, das sich aus

$$F(\infty) = \int_{-\infty}^{\infty} f(x) dx = \frac{b-a}{2} \cdot f(c) = 1 \quad (4.177)$$

ergibt zu

$$f(c) = \frac{2}{b-a} \quad (4.178)$$

Die Wahrscheinlichkeit $P(x \leq c)$ berechnet sich zu

$$P(x \leq c) = F(c) = \int_{-\infty}^c f(\xi) d\xi = \frac{1}{2} \cdot (c-a) \cdot \frac{2}{b-a} = \frac{c-a}{b-a} \quad (4.179)$$

Der Mittelwert der durch Gleichung (4.175) definierten stetigen Dreiecksverteilung liegt bei

$$\mu = E(x) = \frac{a+b+c}{3} \quad (4.180)$$

Die Varianz der dreiecksverteilten Zufallsvariable x ergibt sich aus

$$\sigma^2 = E(x^2) - \mu^2 = \frac{a^2 + b^2 + c^2 - a \cdot b - a \cdot c - b \cdot c}{18} = \frac{(a-b)^2 + (b-c)^2 + (a-c)^2}{36} \quad (4.181)$$

Bei Fragestellungen, bei denen keine detaillierten Daten vorliegen und die Annahme einer Gleichverteilung nicht gerechtfertigt ist, wird meist eine Drei-Punkt-Schätzung mithilfe der Dreiecksverteilung durchgeführt. Derartige Schätzungen werden unter anderen zur Berechnung des benötigten Budgets oder zur Bestimmung eines Auslieferungsdatums an Endkunden bei der Projektplanung benötigt. Dabei werden Parameter a , b und c aus der Erfahrung heraus bestimmt.

Beispiel: Drei-Punkt-Schätzung eines Projektaufwandes

Das Vorgehen einer Drei-Punkt-Schätzung wird an einem Beispiel zur Planung der Zeitspanne für die Programmierung einer Software verdeutlicht. Der Software-Entwickler schätzt, dass er für die Programmierung einer derartigen Software durchschnittlich 60 Stunden benötigt. Im günstigsten Fall schätzt er eine Bearbeitungsdauer von 45 Stunden, treten Probleme bei der Hardware auf, kann sich der Aufwand auf 150 Stunden vergrößern.

Es ergibt sich damit die in Bild 4.25 geschätzte Dreiecksverteilung mit einem Maximum von

$$f(c) = \frac{2}{b-a} = \frac{2}{150-45} = 0.019 \quad (4.182)$$

an der Stelle $c = 60$ Stunden.

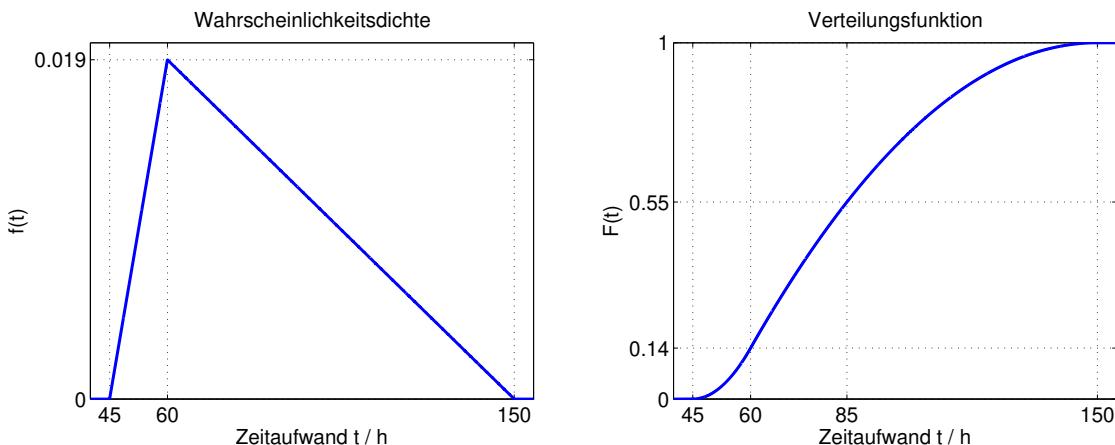


Bild 4.25: Grafische Darstellung der geschätzten Verteilung des Projektaufwandes

Der mittlere Aufwand für das Projekt berechnet sich mit Gleichung (4.180) zu

$$\mu = \frac{a+b+c}{3} = \frac{45+60+150}{3}h = 85h \quad (4.183)$$

Aus der Wurzel der in Gleichung (4.181) definierten Varianz folgt die Standardabweichung der Verteilung zu

$$\sigma = \sqrt{\frac{(a-b)^2 + (b-c)^2 + (a-c)^2}{36}} = \sqrt{\frac{(45-150)^2 + (150-60)^2 + (45-60)^2}{36}}h = 23.18h \quad (4.184)$$

Die mittlere Schätzung des Programmierers von 60 Stunden wird lediglich mit einer Wahrscheinlichkeit von

$$F(x=c) = \frac{c-a}{b-a} = \frac{60-45}{150-45} = 14.29\% \quad (4.185)$$

erreicht. Daher wird zum Beispiel zur Abschätzung der Kosten, die durch das Projekt entstehen, der mittlere Zeitaufwand von 85 Stunden herangezogen. Dieser wird bei der vorliegenden geschätzten Verteilung mit einer Wahrscheinlichkeit von 55.29 % eingehalten.

4.6.3 Weibull-Verteilung

Die Weibull-Verteilung wird unter anderem zur Modellierung von Lebensdauern in der Qualitätssicherung verwendet. Sie wird vor allem bei Fragestellungen wie der Materialermüdung von spröden Werkstoffen oder dem Ausfallen von elektronischen Bauteilen eingesetzt. Benannt ist sie nach dem Schweden Waloddi Weibull. Die Wahrscheinlichkeitsdichte der Weibull-Verteilung ist für $x < 0$ null. Für $x \geq 0$ ist sie definiert als

$$f(x) = \begin{cases} \frac{\beta}{\eta} \cdot \left(\frac{x}{\eta}\right)^{\beta-1} \cdot e^{-\left(\frac{x}{\eta}\right)^\beta} & \text{für } x \geq 0 \\ 0 & \text{für } x < 0 \end{cases} \quad (4.186)$$

Die Verteilungsfunktion $F(x)$ der Weibull-Verteilung lautet

$$F(x) = 1 - e^{-\left(\frac{x}{\eta}\right)^\beta} \quad (4.187)$$

Der Parameter β beschreibt, ob die Ausfallrate über der Lebensdauer zunimmt ($\beta > 1$), abnimmt ($0 < \beta < 1$) oder konstant bleibt ($\beta = 1$). Bild 4.26 stellt die Wahrscheinlichkeitsdichte $f(x)$ der Weibull-Verteilung für unterschiedliche Parametervarianten dar.

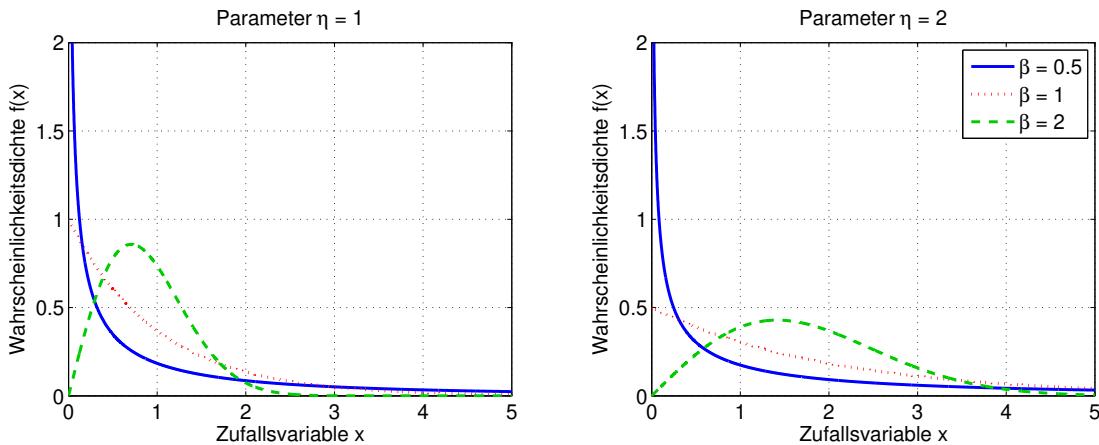


Bild 4.26: Wahrscheinlichkeitsdichte $f(x)$ der Weibullverteilung für $\beta = 0.5, 1$ und 2 und $\eta = 1, 2$

Mit wachsendem β geht die Dichtefunktion $f(x)$ der Weibull-Verteilung von einer rechtsschiefen in eine linksschiefe Verteilung über. Das entspricht der anschaulichen Vorstellung der variablen Ausfallrate. Für kleine Werte von $\beta < 1$ nimmt die Ausfallrate ab, die Verteilung wird deshalb mit wachsendem Wert der Zufallsvariablen x flacher. Mit einem Wert $\beta > 1$ nimmt die Ausfallrate mit wachsendem Wert der Zufallsvariablen x zu.

Der Parameter η beschreibt die Lage der Verteilung auf der x -Achse. Wird mit $F(x)$ die Verteilung von Lebensdauern beschrieben, stellt der Parameter η die charakteristische Lebensdauer dar. Daraus ergibt sich die Folgerung, dass $\eta > 0$ sein muss. In Anlehnung an die Zeitkonstante dynamischer Übertragungsglieder entspricht der Parameter η einer Lebensdauer mit einer Ausfallwahrscheinlichkeit von 63.2 %. Zur Analyse der Bedeutung des Parameters η stellt Bild 4.27 die Verteilungsfunktion $F(x)$ für unterschiedliche Parameter β und η dar.

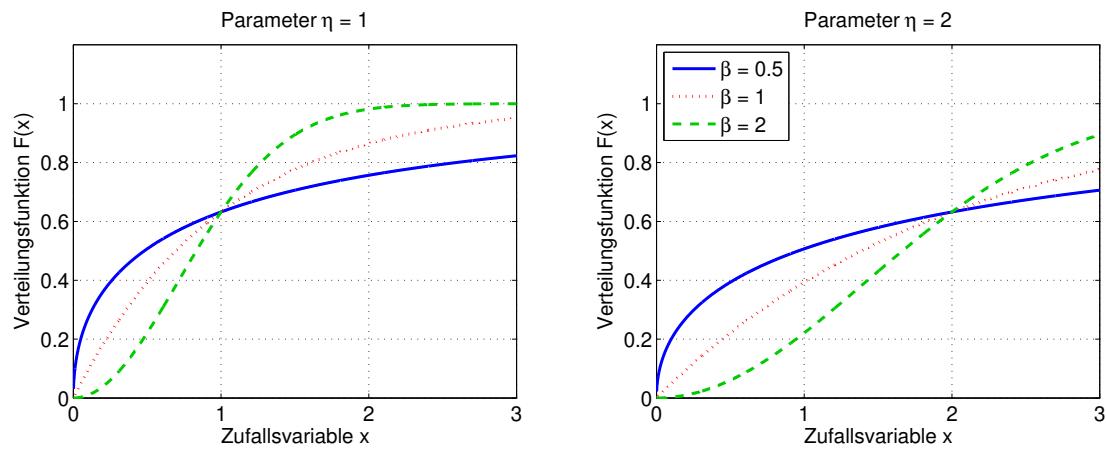


Bild 4.27: Verteilungsfunktion $F(x)$ der Weibullverteilung für Parameter $\beta = 0.5, 1$ und 2 und Parameter $\eta = 1, 2$

Der Mittelwert ergibt sich zu

$$\mu = \left(\frac{1}{\eta} \right)^{-\frac{1}{\beta}} \cdot \Gamma \left(\frac{1}{\beta} + 1 \right) \quad (4.188)$$

und die Varianz berechnet sich zu

$$\sigma^2 = \left(\frac{1}{\eta} \right)^{\frac{2}{\beta}} \cdot \left(\Gamma \left(\frac{2}{\beta} + 1 \right) - \Gamma^2 \left(\frac{1}{\beta} + 1 \right) \right) \quad (4.189)$$

wobei die Funktion

$$\Gamma(\alpha) = \int_0^\infty e^{-u} \cdot u^{\alpha-1} du \quad (4.190)$$

als Gamma-Funktion [Krey91] bezeichnet wird.

In der Praxis ist die Weibull-Verteilung neben der Exponential-Verteilung die am häufigsten verwendete Lebensdauerverteilung. Dabei wird die in Bild 4.28 gezeigte Darstellung im doppelt logarithmischen Maßstab eingesetzt, wodurch eine Approximation der Verteilung über zwei Geraden ermöglicht wird.

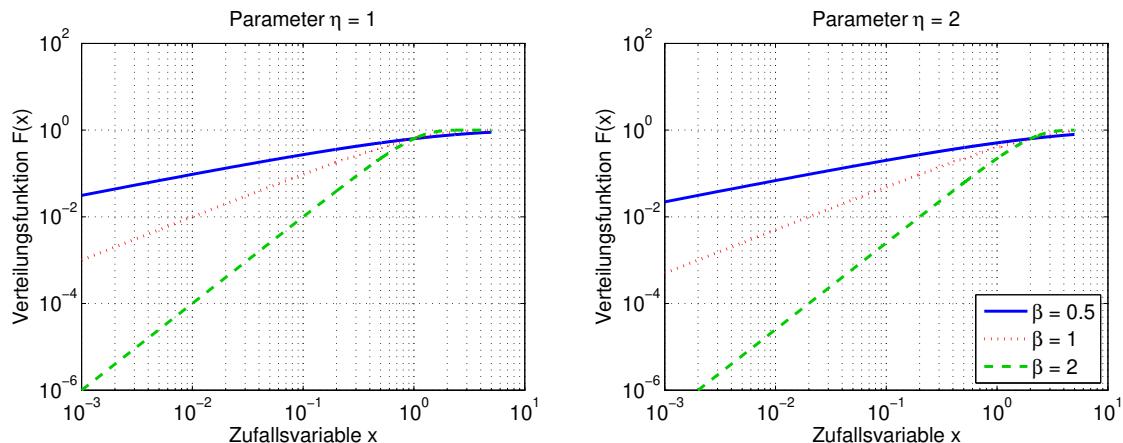


Bild 4.28: Verteilungsfunktion $F(x)$ der Weibullverteilung für $\beta = 0.5, 1$ und 2 und $\eta = 1, 2$ im doppelt logarithmischen Maßstab

Beispiel: Verteilung von Windgeschwindigkeiten

Statt der Anwendung zur Beschreibung von Lebensdauern wird hier ein Beispiel zur Berechnung von mittleren Leistungen bei Windkraftanlagen aufgegriffen. Der natürliche Wind schwankt in seiner Geschwindigkeit. Um die Energieerzeugung durch eine Windkraftanlage vorhersagen zu können, muss daher bekannt sein, welche Häufigkeitsverteilung der Wind an einem Standort besitzt. Üblicherweise werden die zeitlichen Häufigkeiten der verschiedenen Geschwindigkeiten durch die zweiparametrische Weibull-Verteilung mit den Parametern β und η beschrieben.

$$f(x) = \frac{\beta}{\eta} \cdot \left(\frac{x}{\eta} \right)^{\beta-1} \cdot e^{-\left(\frac{x}{\eta} \right)^\beta} \quad (4.191)$$

Der Parameter β ist der Weibull-Formfaktor und gibt die Form der Verteilung an, er nimmt einen Wert von $\beta = 1$ bis 3 an. Einen großen β -Wert gibt es für Winde mit geringen Schwankungen, wie zum Beispiel bei konstanten Passatwinden. In Europa ist ein β -Faktor von 2 üblich. Sehr variable Winde, wie zum Beispiel die Winde im Polargebiet werden durch ein kleines β beschrieben. Der Parameter β nimmt mit der Höhe leicht zu, da Turbulenzen und Schwankungen mit der Höhe sinken.

Der Parameter η ist der Weibull-Skalierungsfaktor in m/s. Er steht in einem bestimmten Verhältnis zum Mittelwert der Windgeschwindigkeit v der Verteilung und ist damit von dem Standort der Windkraftanlage abhängig. Der Parameter η beschreibt damit die Lage der Verteilung auf der Geschwindigkeitssachse.

Bild 4.29 vergleicht die Windgeschwindigkeitsverteilung von Europa ($\beta = 2$) mit der Windverteilung von Passatwinden ($\beta = 3$) für einen konstanten Skalierungsfaktor von $\eta = 10$.

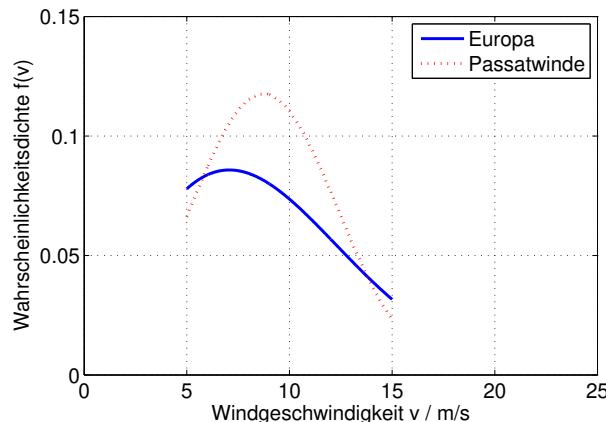


Bild 4.29: Vergleich der Windgeschwindigkeitsverteilungen von Europa ($\beta = 2$) mit der Windverteilung von Passatwinden ($\beta = 3$) jeweils für einen Skalierungsfaktor $\eta = 10$

Windkraftanlagen arbeiten in einem definierten Geschwindigkeitsintervall. Bei diesem Beispiel wird eine Windkraftanlage zugrunde gelegt, die bei Geschwindigkeiten zwischen 5 und 15 m/s arbeitet, einen Rotorradius von 5 m und einen als konstant angenommenen Leistungsbeiwert $c_p = 0.48$ besitzt. Die mittlere Leistung P der Windkraftanlage errechnet sich nach mit einer Dichte und $\rho = 1.2 \text{ kg/m}^3$ aus dem Erwartungswert

$$E(P) = \int_5^{15} c_p \cdot \rho \cdot A \cdot \frac{1}{2} \cdot v^3 \cdot f(v) dv \quad (4.192)$$

Dabei werden für $f(v)$ je nach Standort unterschiedliche Weibull-Verteilungen zugrunde gelegt. Für das Beispiel ergibt sich bei einem Standort in Europa ein Erwartungswert von 15.4 kW, bei einem Standort mit Passatwinden ein Erwartungswert von 19.1 kW. Aufgrund der engeren Verteilung bei Passatwinden befindet sich die Windkraftanlage oft in dem Betriebsbereich. Mit steigendem Wert β steigt deshalb die Ausbeute der Windenergie.

Die zur Berechnung erforderliche MATLAB-Sequenz zeigt sich wie folgt.

```

1 % Geschwindigkeitsvektor erzeugen
2 dv = 0.01;
3 v = 5:dv:15;
4
5 % Variablendefinition Leistungsberechnung
6 roh = 1.2;
7 r = 5;
8 A = pi*r^2;
9 cp = 0.48;
10
11 % Leistungsberechnung über Riemannsche Summe
12 P1 = sum(0.5 * cp * roh * A * v.^3.*wblpdf(v,10,2)*dv)
13 P2 = sum(0.5 * cp * roh * A * v.^3.*wblpdf(v,10,3)*dv)

```

Die Realisierung in Python führt zu demselben Ergebnis.

```

1 Bibliotheken importieren
2 import numpy as np
3 import matplotlib.pyplot as plt
4 from scipy.stats import weibull_min
5
6 Geschwindigkeitsvektor erzeugen
7 dv = 0.01
8 v = np.arange(5, 15+dv, dv)
9
10 Variablendefinition Leistungsberechnung
11 roh = 1.2
12 r = 5
13 A = np.pi*r**2
14 cp = 0.48
15 P = 0.5 * cp * roh * A * v**3
16
17 Variablendefinition Weibullverteilung
18 beta = [2, 3]
19 eta = 10
20 fw = np.empty((len(beta), len(v)))
21 for k in [0,1]:
22     fw[k,:] = weibull_min.pdf(v, beta[k], 0, eta)
23
24 Berechnung der Leistungen
25 P1 = np.sum(P*fw[0,:]*dv)
26 P2 = np.sum(P*fw[1,:]*dv)
27 print(' ')
28 print('Mittlere Leistung Europa: ', P1)
29 print('Mittlere Leistung Passatwinde: ', P2)

```

Die Weibull-Verteilung geht für eine konstante Ausfallrate ($\beta = 2$) in die Rayleigh-Verteilung über, bei der sich der Koeffizient b errechnet aus

$$b = \frac{\eta}{\sqrt{2}} \quad (4.193)$$

Die Rayleigh-Verteilung wird in Abschnitt 4.6.5 eingeführt und diskutiert.

Ein weiterer Sonderfall der Weibull-Verteilung stellt die Exponential-Verteilung dar. Sie ergibt sich aus der Weibull-Verteilung für eine konstante Ausfallrate ($\beta = 1$). Der Koeffizient λ der Exponential-Verteilung berechnet sich aus den Parametern der Weibull-Verteilung zu

$$\lambda = \frac{1}{\eta} \quad (4.194)$$

Die Exponential-Verteilung wird im folgenden Abschnitt 4.6.4 eingeführt.

4.6.4 Exponential-Verteilung

In Abschnitt 4.5.6 wird die geometrische Verteilung als Modell zur Abschätzung von Lebensdauern und Wartezeiten abgeleitet. Dabei war die Zufallsvariable x diskret. Die entsprechende Verteilung für stetige Zufallsvariablen ist die Exponential-Verteilung. Sie wird angewendet, um zum Beispiel die Lebensdauer von Produkten oder Zeiträume bis zu Schadensfällen zu berechnen. Voraussetzung für die Anwendung der Exponential-Verteilung ist, dass die noch zu erwartende Lebensdauer nicht von der bereits absolvierten Lebensdauer abhängig ist. Anschaulich bedeutet das, dass das Produkt keine Alterungserscheinungen aufweist. Diese Voraussetzung entspricht der Forderung nach statistischer Unabhängigkeit der Erfolgswahrscheinlichkeit bei der geometrischen Verteilung.

Die Exponential-Verteilung ist für $x > 0$ definiert als

$$f(x) = \lambda \cdot e^{-\lambda \cdot x} \quad (4.195)$$

Daraus folgt für $x > 0$ die Verteilungsfunktion $F(x)$

$$F(x) = 1 - e^{-\lambda \cdot x} \quad (4.196)$$

Der Parameter λ beschreibt, wie schnell sich die Wahrscheinlichkeitsdichte dem Wert null nähert. Bild 4.30 stellt für unterschiedliche Parameter λ die Wahrscheinlichkeitsdichte $f(x)$ und Verteilungsfunktion $F(x)$ der Exponential-Verteilung dar.

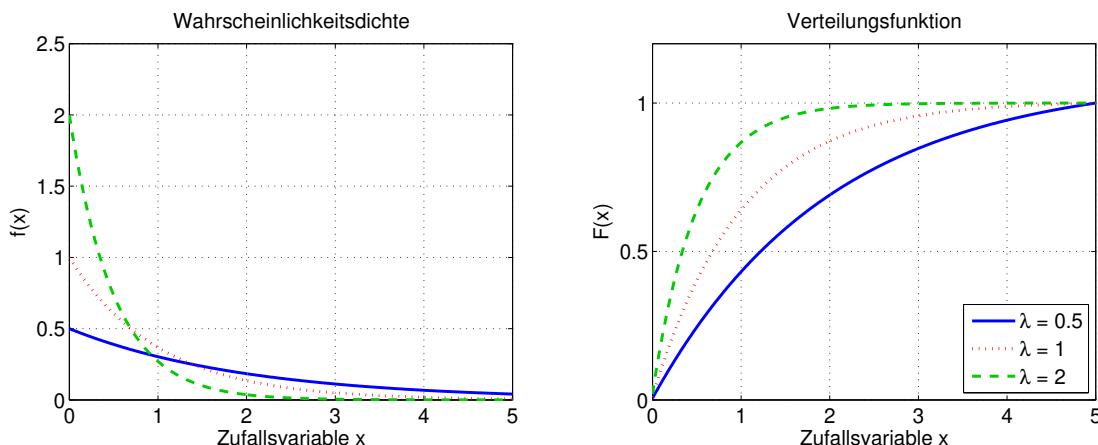


Bild 4.30: Wahrscheinlichkeitsdichte $f(x)$ und Verteilungsfunktion $F(x)$ der Exponentialverteilung für $\lambda = 0.5, 1, 2$

Die Erwartungswerte Mittelwert und Standardabweichung können wegen der Exponentialfunktion nicht mehr elementar bestimmt werden. Es kann gezeigt werden, dass sich der Mittelwert und die Varianz der Exponential-Verteilung berechnen zu

$$\mu = E(x) = \frac{1}{\lambda} \quad (4.197)$$

und

$$\sigma^2 = E((x - \mu)^2) = E(x^2) - E^2(x) = \frac{1}{\lambda^2} \quad (4.198)$$

Die Exponentialfunktion wird neben der Berechnung der noch zu erwartenden Lebensdauer von Bauelementen auch für die Abschätzung der mittleren Betriebsdauer zwischen Ausfällen von Produkten oder Einrichtungen, der Mean Time Between Failures (MTBF), verwendet. Dies wird im folgenden Beispiel aufgezeigt und diskutiert.

Beispiel: Mean Time Between Failures (MTBF) eines Lasersystems

Unter dem Begriff MTBF wird die mittlere Betriebsdauer zwischen zwei Ausfällen einer Einheit verstanden. Die Betriebsdauer gibt dabei an, wie lange eine instandgesetzte Einheit zwischen zwei aufeinanderfolgenden Ausfällen funktionsfähig ist. Die mittlere Betriebsdauer kann daher als Maß für die Zuverlässigkeit herangezogen werden und dient zur Abschätzung von Ausfällen in bestimmten Zeitintervallen. Ist die Betriebsdauer unter Berücksichtigung der oben genannten Einschränkungen exponentiellverteilt, ergibt sich die MTBF aus dem Kehrwert der konstanten Ausfallrate λ .

$$MTBF = \frac{1}{\lambda} \quad (4.199)$$

Für ein Lasersystem, dessen erreichbare mittlere Betriebsdauer von dem Hersteller mit 20 Monaten angegeben wird, ergibt sich die in Bild 4.31 dargestellte Exponential-Verteilung mit $\lambda = 0.05$.

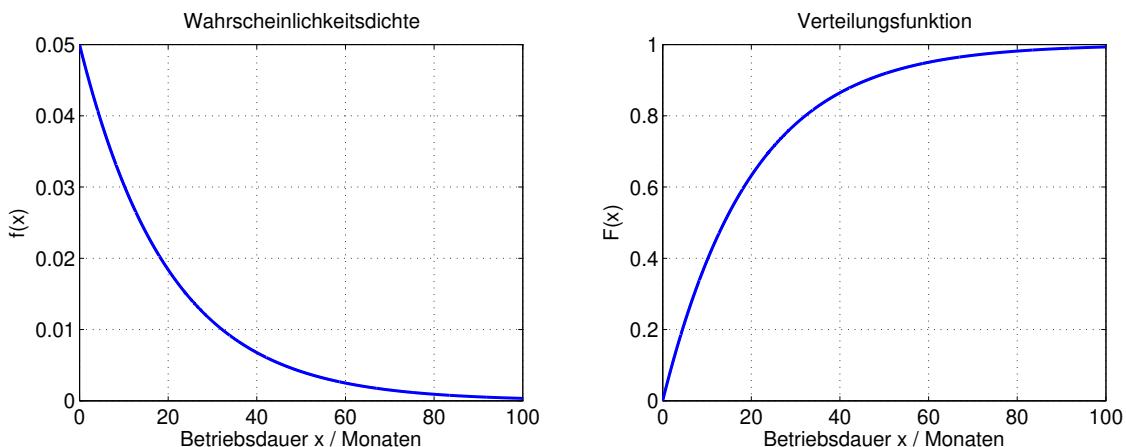


Bild 4.31: Wahrscheinlichkeitsdichte $f(x)$ und Verteilungsfunktion $F(x)$ eines Lasersystems

Bei der Interpretation des MTBF-Wertes des Lasersystems muss beachtet werden, dass der MTBF-Wert nicht aussagt, dass das Lasersystem im Mittel 20 Monate ohne Ausfall arbeitet. Aus der Verteilungsfunktion $F(x)$ aus Bild 4.31 berechnet sich die Wahrscheinlichkeit dafür, dass das Lasersystem bis zur prognostizierten mittleren Betriebsdauer ausfällt zu

$$F(MTBF) = 1 - e^{-0.05 \cdot 20} = 0.6321 \quad (4.200)$$

Lediglich 36.79 % der Lasersysteme wird somit 20 Monate zwischen zwei Ausfällen funktionieren. Eine vorbeugende Instandsetzung sollte daher stets vor der prognostizierten mittleren Betriebsdauer durchgeführt werden.

Die zur Berechnung erforderliche In MATLAB- ergibt sich folgende Programm-Sequenz zeigt sich wie folgt zur Berechnung.

```

1 % Variablendefinition
2 x = 0:0.01:100;
3 MTBF = 20;
4 lambda = 1/MTBF;
5
6 % Berechnung der Wahrscheinlichkeiten
7 p_MTBF_Ausfall = 1 - expcdf(20,1/lambda);
8 p_MTBF_keinAusfall = expcdf(MTBF,1/lambda);

```

Entsprechend ergibt sich in Python.

```
1 Bibliotheken importieren
2 from scipy.stats import expon
3
4 Variablendefinition und Berechnung der Wahrscheinlichkeit
5 lamb = 0.05
6 p_MTBF_Ausfall = expon.cdf(20, 0, 1/lamb)
7 p_MTBF_keinAusfall = 1 - p_MTBF_Ausfall
8 print(' ')
9 print('Wahrscheinlichkeit für Ausfall: ', p_MTBF_Ausfall)
10 print('Wahrscheinlichkeit für keinen Ausfall: ', p_MTBF_keinAusfall)
```

4.6.5 Rayleigh-Verteilung (Betragverteilung 2. Art)

Mithilfe der Rayleigh-Verteilung wird die Verteilung des Betrages zweier unabhängiger normalverteilter Zufallsgrößen mit der Standardabweichung $\sigma = b$ beschrieben. Zum Beispiel folgt der Wert des Radius eines Kreises, dessen x- und y-Koordinaten normalverteilt sind, einer Rayleigh-Verteilung. Ein anderer Anwendungsfall ist die Beschreibung der 10-Minuten-Mittelwerte von Windgeschwindigkeiten. Da die Rayleigh-Verteilung die Verteilung von Beträgen beschreibt, wird sie auch als Betragverteilung 2. Art bezeichnet. Sie wird zur Bewertung der Prozesssicherheit im Rahmen der statistischen Prozesskontrolle eingesetzt.

Die Wahrscheinlichkeitsdichte $f(x)$ der Rayleigh-Verteilung ist definiert als

$$f(x) = \frac{1}{b^2} \cdot x \cdot e^{-\frac{1}{2} \cdot \frac{x^2}{b^2}} \quad (4.201)$$

Daraus ergibt sich die Verteilungsfunktion $F(x)$ zu

$$F(x) = \frac{1}{b^2} \cdot \int_{-\infty}^x \xi \cdot e^{-\frac{1}{2} \cdot \frac{\xi^2}{b^2}} d\xi = 1 - e^{-\frac{1}{2} \cdot \frac{x^2}{b^2}} \quad (4.202)$$

Bild 4.32 zeigt die Wahrscheinlichkeitsdichte $f(x)$ und die Verteilungsfunktion $F(x)$ der Rayleigh-Verteilung für verschiedene Werte des Parameters b .

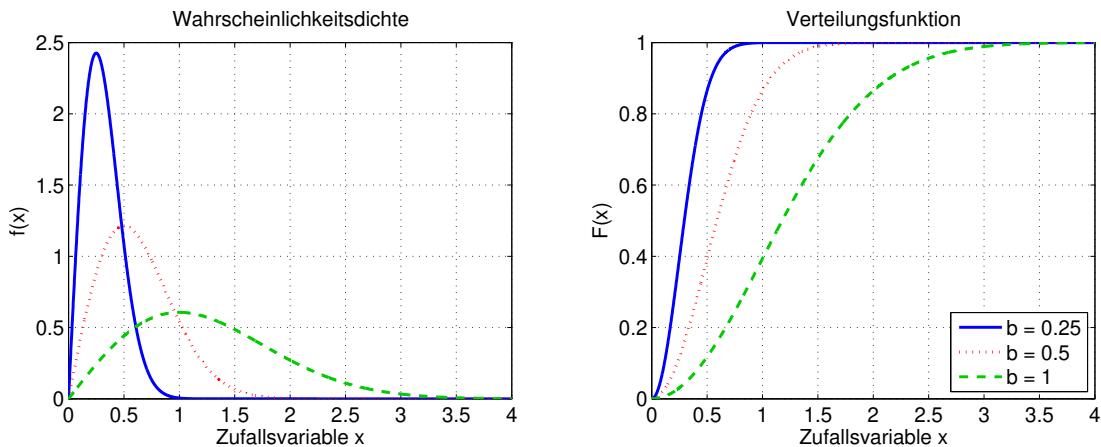


Bild 4.32: Wahrscheinlichkeitsdichte $f(x)$ und Verteilungsfunktion $F(x)$ der Rayleigh-Verteilung

Die Rayleigh-Verteilung besitzt einen Mittelwert von

$$\mu = b \cdot \sqrt{\frac{\pi}{2}} \quad (4.203)$$

Die Varianz der Rayleigh-Verteilung ergibt sich zu

$$\sigma^2 = \frac{4 - \pi}{2} \cdot b^2 \quad (4.204)$$

Damit weist bei der Rayleigh-Verteilung das Verhältnis des Mittelwertes μ zur Standardabweichung σ immer den festen Wert

$$\frac{\mu}{\sigma} = \sqrt{\frac{\pi}{4 - \pi}} \approx 1.91 \quad (4.205)$$

auf.

Beispiel: Abstand einer Bohrung zu ihrer Sollposition

Als Beispiel für die Anwendung der Rayleigh-Verteilung wird die Lage einer Bohrung betrachtet, die mittels einer Fertigungseinrichtung in eine Metallplatte gebohrt wird. Die Lage der Bohrungen unterliegt den Fertigungstoleranzen der Bohreinrichtung. Durch den Maschinenhersteller ist bekannt, dass die Abweichung der x - und y -Koordinaten vom Sollwert jeweils durch eine Normalverteilung mit einem Mittelwert $\mu_x = \mu_y = 0 \text{ } \mu\text{m}$ und einer Standardabweichung von $\sigma_x = \sigma_y = 0.25 \text{ } \mu\text{m}$ beschrieben werden kann.

Aus dem Satz von Pythagoras folgt, dass der Abstand Δz der Bohrung vom definierten Sollwert beschrieben werden kann durch

$$\Delta z = \sqrt{(\Delta x)^2 + (\Delta y)^2} \quad (4.206)$$

Die Verteilung der Größe Δz ist eine Rayleigh-Verteilung mit dem Parameter $b = 0.25$. Diese ist in Bild 4.33 zu sehen.

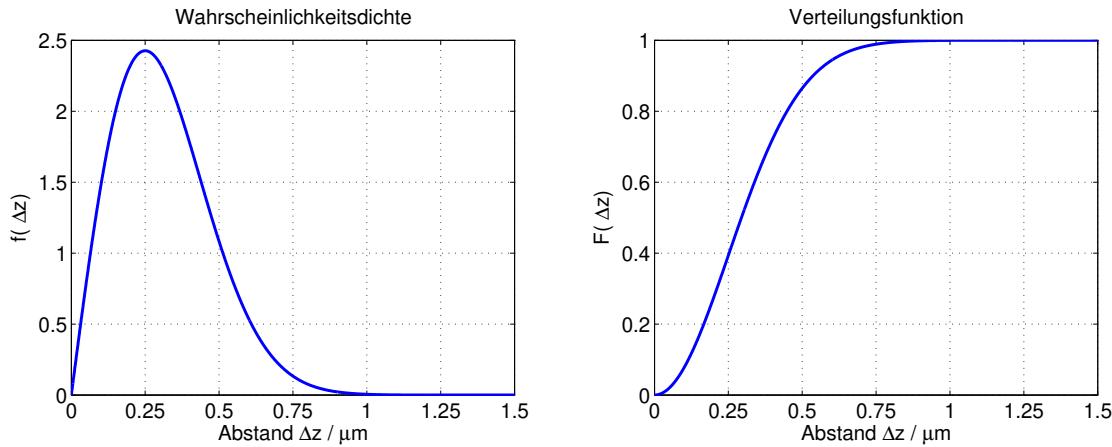


Bild 4.33: Verteilung des Abstandes Δz einer Bohrung zu ihrer Sollposition

4.6.6 Normalverteilung

Die Gauß- oder Normalverteilung wurde von Carl Friedrich Gauß im Zusammenhang mit dem Ausgleich von Messergebnissen gefunden. Sie ist die wichtigste stetige Verteilung, weil viele Messgrößen oder Beobachtungen normalverteilt sind. Außerdem lassen sich viele Verteilungen gut durch die Normalverteilung approximieren. Weiterhin kommen bei statistischen Prüfverfahren oft Größen vor, die entweder direkt normalverteilt sind oder sich bei Grenzübergängen als normalverteilt beschreiben lassen.

Allgemeine Definition der Normalverteilung

Die Wahrscheinlichkeitsdichte der Gauß- oder Normalverteilung ist für $-\infty < x < \infty$ definiert als

$$f(x) = \frac{1}{\sigma \cdot \sqrt{2 \cdot \pi}} \cdot e^{-\frac{1}{2} \cdot \left(\frac{x - \mu}{\sigma}\right)^2} \quad (4.207)$$

Eine Zufallsvariable x mit dieser Verteilung wird als normalverteilte Zufallsvariable bezeichnet. Sie besitzt die Parameter μ und σ . Bild 4.34 zeigt die Wahrscheinlichkeitsdichte der Normalverteilung mit einem Mittelwert von $\mu = 2$ und verschiedenen Werten der Standardabweichung σ .

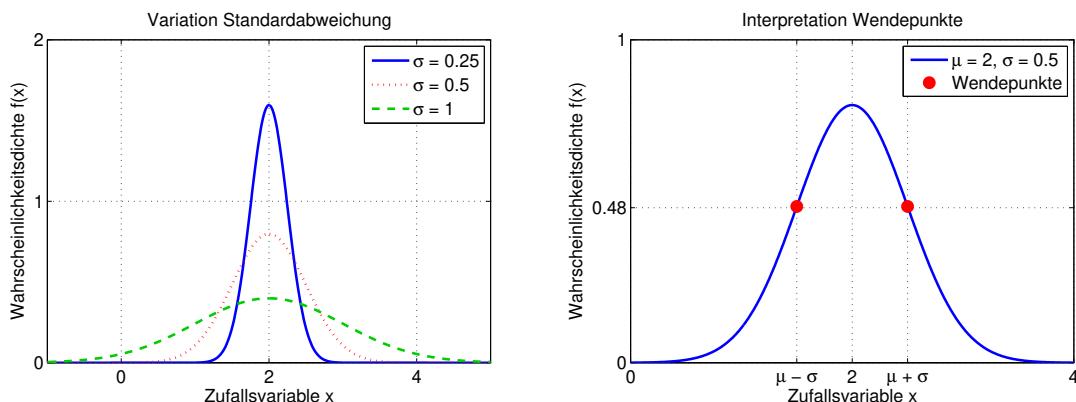


Bild 4.34: Wahrscheinlichkeitsdichte $f(x)$ für $\mu = 2$ und $\sigma = 0.25, 0.5$ und 1 und Lage der Wendepunkte

Aus der Symmetrie der Verteilung ergibt sich, dass das Maximum der Verteilung für den Mittelwert μ erreicht wird. Je kleiner σ ist, desto schmäler ist die Verteilung und desto ausgeprägter ist ihr Maximum. Je größer σ ist, desto breiter ist die Verteilung. Der Abstand der Wendepunkte der Wahrscheinlichkeitsdichte von dem Mittelwert μ entspricht der Standardabweichung σ .

Die Verteilungsfunktion $F(x)$ hat die Form

$$F(x) = \int_{-\infty}^x \frac{1}{\sigma \cdot \sqrt{2 \cdot \pi}} \cdot e^{-\frac{1}{2} \left(\frac{\xi - \mu}{\sigma} \right)^2} d\xi = \frac{1}{\sigma \cdot \sqrt{2 \cdot \pi}} \cdot \int_{-\infty}^x e^{-\frac{1}{2} \left(\frac{\xi - \mu}{\sigma} \right)^2} d\xi \quad (4.208)$$

Die Verteilungsfunktion $F(x)$ ist ein Integral, das sich nicht analytisch auswerten lässt. Die Berechnung wird numerisch ausgeführt, zum Beispiel durch die numerische Integration über die Wahrscheinlichkeitsdichte $f(x)$ oder mittels einer Approximation durch eine Taylor-Reihe.

Bild 4.35 zeigt die Verteilungsfunktion $F(x)$ der Normalverteilung für $\mu = 2$ und verschiedene Werte der Standardabweichung σ .

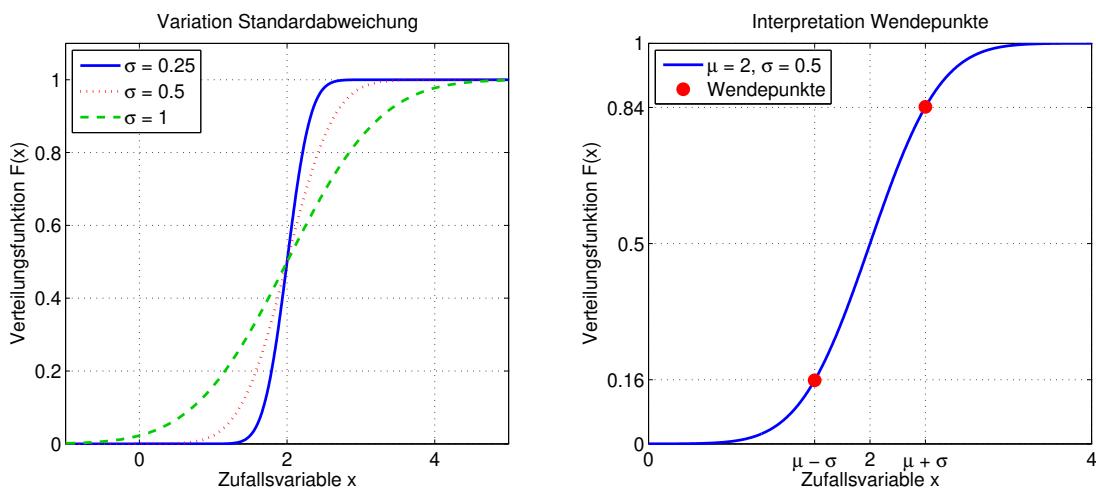


Bild 4.35: Verteilungsfunktion $F(x)$ für $\mu = 2$ und $\sigma = 0.25, 0.5$

Für eine normalverteilte Variable x berechnet sich die Wahrscheinlichkeit P , innerhalb des Intervalls $a < x \leq b$ zu sein, aus

$$P(a < x \leq b) = F(b) - F(a) = \frac{1}{\sigma \cdot \sqrt{2 \cdot \pi}} \cdot \int_a^b e^{-\frac{1}{2} \left(\frac{x - \mu}{\sigma} \right)^2} dx \quad (4.209)$$

Standardisierte Normalverteilung

Die Verteilungsfunktion der Normalverteilung kann nicht mehr analytisch berechnet werden. Durch eine Standardisierung der Zufallsvariablen werden alle Varianten der Normalverteilung auf eine standardisierte Form abgebildet, deren Wahrscheinlichkeitsdichte und Verteilungsfunktion in Form von Tabellen beschrieben ist. Bei der rechnerunterstützten Auswertung von Daten haben diese tabellierten Werte an Bedeutung verloren. Aus Gründen der übersichtlicheren Darstellung und zur Vorbereitung der Rechnung mit Testverteilungen wird im Folgenden dennoch die standardisierte Normalverteilung eingeführt und für die weiteren Berechnungen eingesetzt. Bei der Standardisierung einer Verteilung wird analog zu Abschnitt 4.4.3 eine Standardisierung der Zufallsvariablen durchgeführt, sodass sie einen Mittelwert von $\mu = 0$ und eine Standardabweichung von $\sigma = 1$ erreicht. Nach den Ausführungen in Abschnitt 4.4.3 führt die Standardisierung der normalverteilten Zufallsvariable x zu einer standardnormalverteilten Zufallsvariable z durch den Ausdruck

$$z = \frac{x - \mu_x}{\sigma_x} \quad (4.210)$$

Dabei geht die Normalverteilung aus Gleichung (4.207) über in die Standardnormalverteilung mit $\mu_z = 0$ und $\sigma_z = 1$. Die Dichtefunktion der standardisierten Normalverteilung vereinfacht sich damit gegenüber Gleichung (4.207) zu

$$f(z) = \frac{1}{\sqrt{2 \cdot \pi}} \cdot e^{-\frac{1}{2} \cdot z^2} \quad (4.211)$$

Die zugehörige Verteilungsfunktion wird beschrieben durch

$$F(z) = \int_{-\infty}^z \frac{1}{\sqrt{2 \cdot \pi}} \cdot e^{-\frac{1}{2} \cdot \zeta^2} d\zeta = \frac{1}{\sqrt{2 \cdot \pi}} \cdot \int_{-\infty}^z e^{-\frac{1}{2} \cdot \zeta^2} d\zeta \quad (4.212)$$

Die Wahrscheinlichkeitsdichte $f(z)$ und die Verteilungsfunktion $F(z)$ der Standardnormalverteilung ist in Bild 4.36 dargestellt.

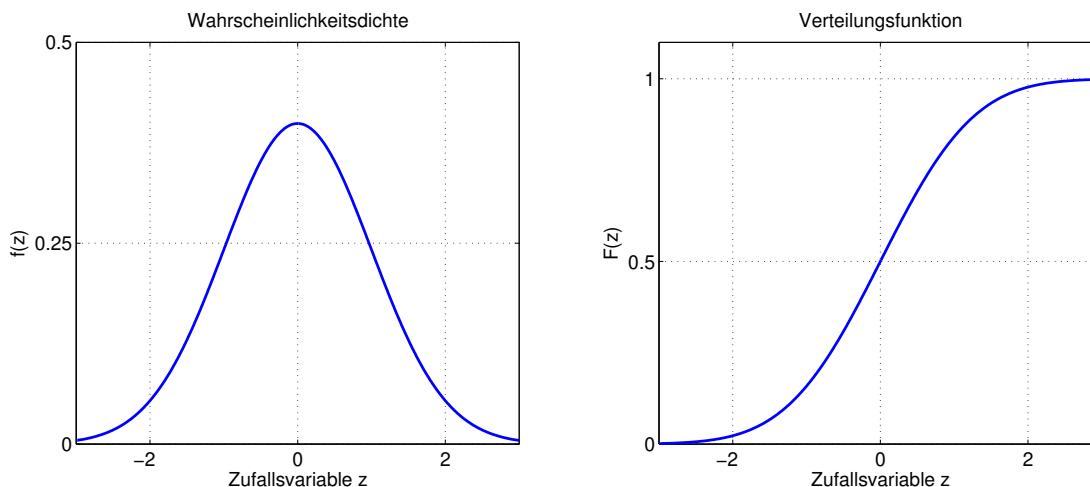


Bild 4.36: Wahrscheinlichkeitsdichte $f(z)$ und Verteilungsfunktion $F(z)$ der Standardnormalverteilung

Für eine standardnormalverteilte Variable z berechnet sich die Wahrscheinlichkeit P , innerhalb des Intervalls $a < z \leq b$ zu sein, aus

$$P(a < z \leq b) = F(b) - F(a) = \frac{1}{\sqrt{2 \cdot \pi}} \cdot \int_a^b e^{-\frac{1}{2} \cdot z^2} dz \quad (4.213)$$

Mit Gleichung (4.213) können die Aufenthaltswahrscheinlichkeiten für eine standardnormalverteilte Zufallsvariable z in symmetrischen Intervallen errechnet werden.

Tabelle 4.14: Aufenthaltswahrscheinlichkeiten für eine standardnormalverteilte Zufallsvariable z

Intervall	Aufenthaltswahrscheinlichkeit
$-1\sigma < z \leq 1\sigma$	$\approx 68.2689\%$
$-2\sigma < z \leq 2\sigma$	$\approx 95.4500\%$
$-3\sigma < z \leq 3\sigma$	$\approx 99.7300\%$

Eine Abweichung vom Mittelwert um mehr als eine Standardabweichung ist etwa in einem von drei Fällen zu erwarten, eine Abweichung um mehr als drei Standardabweichungen nur in einem von 370 Fällen.

Beispiel: Zusammenhang zwischen Toleranzziel und Ausschuss bei der Fertigung

Als Beispiel für die Normalverteilung soll der Zusammenhang zwischen Toleranzziel und Ausschuss bei der Fertigung von Widerständen betrachtet werden. Von einem Fertigungslos von 10000 Widerständen sei bekannt, dass die Widerstandsverteilung als Normalverteilung mit dem Mittelwert $\mu = 998 = ?$ und einer Standardabweichung von $\sigma = 5 = ?$ dargestellt werden kann. Es werden unterschiedliche Kunden beliefert. Kunde A fordert ein Toleranzziel von

$$R_A = 1000\Omega \pm 10\Omega \quad (4.214)$$

Kunde B fordert ein Toleranzziel von

$$R_B = 1000\Omega \pm 20\Omega \quad (4.215)$$

Für beide Kunden soll der Anteil von Ausschussteilen berechnet werden. Da die Widerstandswerte eine Normalverteilung mit $\mu = 998 = ?$ und $\sigma = 5 = ?$ besitzen, ergibt sich der Anteil von Ausschussteilen aus

$$A_A = F\left(z = \frac{990 - 998}{5}\right) + 1 - F\left(z = \frac{1010 - 998}{5}\right) = 6.3\% \quad (4.216)$$

Mit den Toleranzgrenzen von Kunde B ergibt sich

$$A_B = F\left(z = \frac{980 - 998}{5}\right) + 1 - F\left(z = \frac{1020 - 998}{5}\right) = 0.02\% \quad (4.217)$$

Für Kunde B ist der Ausschussanteil wegen des doppelt so großen Toleranzbereiches um mehr als zwei Größenordnungen geringer. Durch eine Zentrierung des Mittelwertes von $\mu = 998 = ?$ auf $\mu = 1000 = ?$ ließe sich der Ausschussanteil zudem auf 4.5 % beziehungsweise 63 ppm reduzieren.

Die Werte nach Gleichung (4.216) und (4.217) berechnen sich mit der folgenden MATLAB-Sequenz:

```

1 % Variablendefinition
2 N = 10000;
3 mu = 998;
4 sigma = 5;
5
6 % Berechnung des Ausschussanteils von Kunde A und B ohne Zentrierung
7 AA = normcdf((990-mu)/sigma) + 1 - normcdf((1010-mu)/sigma);
8 AA = normcdf((980-mu)/sigma) + 1 - normcdf((1020-mu)/sigma);
9
10 % Berechnung des Ausschussanteils von Kunde A und B nach Zentrierung
11 mu_neu = 1000;
12
13 AA = normcdf((990-mu_neu)/sigma) + 1 - normcdf((1010-mu_neu)/sigma);
14 AA = normcdf((980-mu_neu)/sigma) + 1 - normcdf((1020-mu_neu)/sigma);

```

Entsprechend ergibt sich in Python:

```

1 from scipy.stats import norm
2
3 Variablendefinition
4 N = 10000
5 mu = 998
6 sigma = 5
7
8 Berechnung des Ausschussanteils von Kunde A und B ohne Zentrierung
9 AA = norm.cdf((990-mu)/sigma) + 1 - norm.cdf((1010-mu)/sigma)
10 AB = norm.cdf((980-mu)/sigma) + 1 - norm.cdf((1020-mu)/sigma)
11 print(' ')
12 print('Ausschussanteil Kunde A: ', AA)
13 print('Ausschussanteil Kunde B: ', AB)
14
15 Berechnung des Ausschussanteils von Kunde A und B nach Zentrierung
16 mu_neu = 1000;
17 AA = norm.cdf((990-mu_neu)/sigma) + 1 - norm.cdf((1010-mu_neu)/sigma)
18 AB = norm.cdf((980-mu_neu)/sigma) + 1 - norm.cdf((1020-mu_neu)/sigma)
19 print(' ')
20 print('Ausschussanteil Kunde A nach Zentrierung: ', AA)
21 print('Ausschussanteil Kunde B nach Zentrierung: ', AB)

```

4.6.7 Logarithmische Normalverteilung

Verteilungen von nicht negativen Zufallsvariablen sind oft nicht symmetrisch. Verteilungen von Lebensdauern, Wartezeiten oder Einkommen sind rechtsschief und können deshalb nicht direkt mit der Normalverteilung beschrieben werden. In einigen Fällen kann die Verteilung durch die Exponential-, Rayleigh- oder Weibull-Verteilung erfolgen. Eine weitere Möglichkeit zur Beschreibung ist die logarithmische Normalverteilung. Die logarithmische Normalverteilung ist definiert durch die Zufallsvariable y , die sich aus der Exponentialfunktion der normalverteilten Zufallsvariable x ergibt.

$$y = g(x) = e^x \quad (4.218)$$

Die Wahrscheinlichkeitsdichte $f_Y(y)$ ergibt sich durch die Variablentransformation der in Gleichung (4.218) dargestellten Form durch

$$f_Y(y) = f_X(g^{-1}(y)) \cdot \left| \frac{dg^{-1}(y)}{dy} \right| \quad (4.219)$$

Mit der Umkehrfunktion

$$x = \ln(y) \quad (4.220)$$

ergibt sich die Wahrscheinlichkeitsdichte $f_Y(y)$

$$f_Y(y) = \frac{1}{\sigma_x \cdot \sqrt{2 \cdot \pi}} \cdot \frac{1}{y} \cdot e^{-\frac{1}{2} \left(\frac{\ln(y) - \mu_x}{\sigma_x} \right)^2} \quad (4.221)$$

und die Verteilungsfunktion $F_Y(y)$ lautet

$$F_Y(y) = \frac{1}{\sigma_x \cdot \sqrt{2 \cdot \pi}} \cdot \int_{-\infty}^y \frac{1}{\psi} \cdot e^{-\frac{1}{2} \left(\frac{\ln(\psi) - \mu_x}{\sigma_x} \right)^2} d\psi \quad (4.222)$$

Bild 4.37 stellt für Exponential-Verteilungen mit Mittelwert $\mu_x = 0$ und unterschiedlicher Standardabweichung σ_x die Wahrscheinlichkeitsdichte $f_Y(y)$ und Verteilungsfunktion $F_Y(y)$ dar.

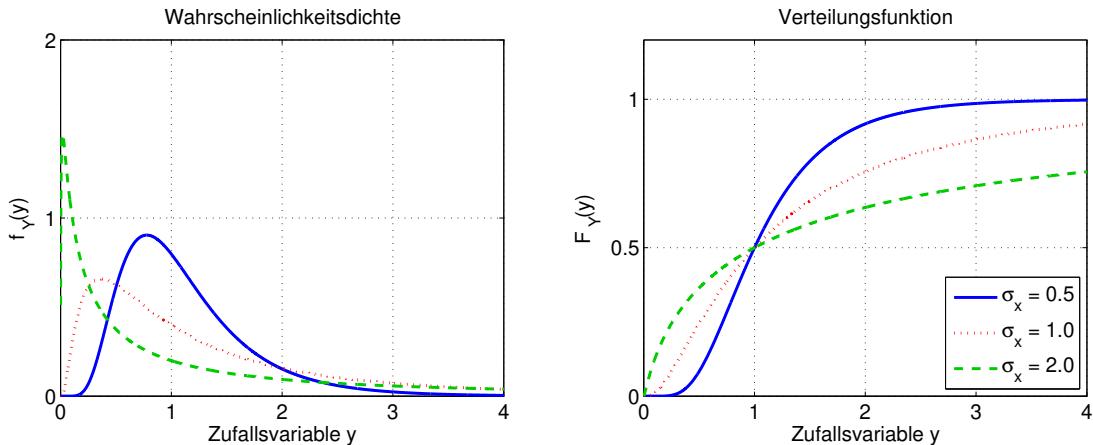


Bild 4.37: Wahrscheinlichkeitsdichte $f_Y(y)$ und Verteilungsfunktion $F_Y(y)$ der logarithmischen Normalverteilung für $\mu_X = 0$ und $\sigma_X = 0.5, 1.0$ und 2.0

Mittelwert μ_Y und Varianz σ_Y^2 können als Funktion des Mittelwertes μ_X und Varianz σ_X^2 der nicht logarithmierten Größen dargestellt werden als

$$\mu_Y = e^{\mu_X + \frac{\sigma_X^2}{2}} \quad (4.223)$$

und

$$\sigma_Y^2 = e^{2 \cdot \mu_X + \sigma_X^2} \cdot (e^{\sigma_X^2} - 1) \quad (4.224)$$

Die Verteilungsfunktion der logarithmischen Normalverteilung kann im doppelt logarithmisch geteilten Wahrscheinlichkeitspapier durch zwei Geraden angenähert werden. Die Approximation der Verteilungsfunktion durch Geradengleichungen wird mit steigendem Wert für σ_X besser, wie in Bild 4.38 zu sehen ist.

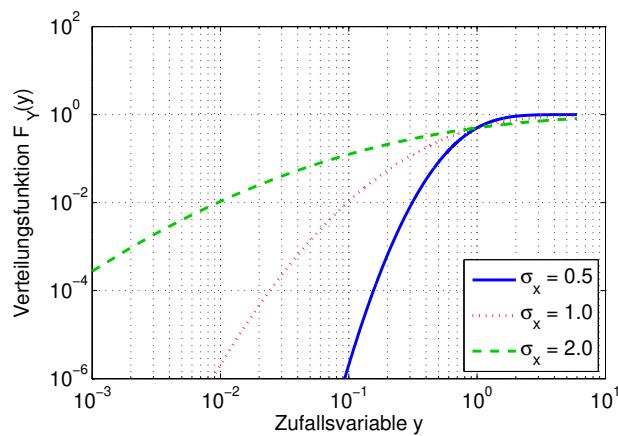


Bild 4.38: Verteilungsfunktion $F_Y(y)$ der logarithmischen Normalverteilung für $\mu_X = 0$ und $\sigma_X = 0.5, 1$ und 2 im doppelt logarithmischen Maßstab

Beispiel: Rauigkeit einer Oberfläche

Als Beispiel für die Anwendung der Logarithmischen Normalverteilung wird die Verteilung der Rauigkeit einer Oberfläche dargestellt. Die Untersuchung einer Oberfläche hat ergeben, dass die Verteilung der Rauigkeit durch eine logarithmische Normalverteilung mit einem Mittelwert von $\mu_R = 1.78$ nm und einer Standardabweichung von $\sigma_R = 0.65$ nm beschrieben werden kann. Bild 4.39 stellt die festgestellte Wahrscheinlichkeitsdichte $f(R)$ und die Verteilungsfunktion $F(R)$ der Rauheit R dar.

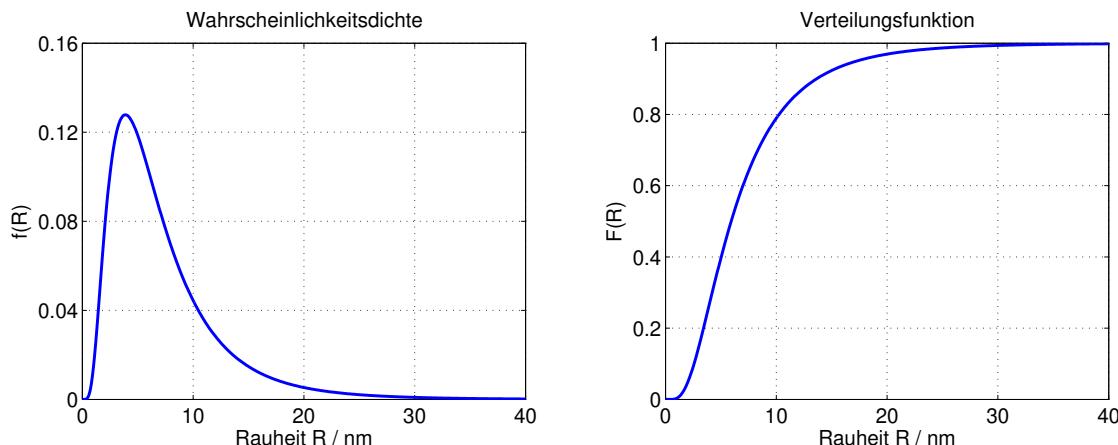


Bild 4.39: Darstellung der Verteilung der Rauheit R einer Oberfläche

Der Wert der Rauigkeit kann nicht kleiner als 0 werden. Dies führt zu einer rechtsschiefen Verteilung.

4.6.8 Betragsverteilung 1. Art

Wird die Normalverteilung an einem beliebigen Punkt unterhalb des Mittelwertes μ gefaltet, führt dies zur Betragsverteilung 1. Art. Durch die Faltung werden die Werte links des Faltungspunktes denen rechts vom Faltungspunkt additiv überlagert. Dabei entsteht die Wahrscheinlichkeitsdichte $f(x)$ der Betragsverteilung 1. Art zu

$$f(x) = \begin{cases} \frac{1}{\sigma \cdot \sqrt{2 \cdot \pi}} \cdot \left(e^{-\frac{1}{2} \cdot \left(\frac{x-\mu}{\sigma} \right)^2} + e^{-\frac{1}{2} \cdot \left(\frac{x+\mu}{\sigma} \right)^2} \right) & \text{für } x \geq 0 \\ 0 & \text{für } x < 0 \end{cases} \quad (4.225)$$

Die Verteilungsfunktion $F(x)$ folgt durch Integration zu

$$F(x) = \begin{cases} \frac{1}{\sigma \cdot \sqrt{2 \cdot \pi}} \cdot \int_0^x \left(e^{-\frac{1}{2} \cdot \left(\frac{\xi - \mu}{\sigma} \right)^2} + e^{-\frac{1}{2} \cdot \left(\frac{\xi + \mu}{\sigma} \right)^2} \right) d\xi & \text{für } x \geq 0 \\ 0 & \text{für } x < 0 \end{cases} \quad (4.226)$$

Die Wahrscheinlichkeitsdichte $f(x)$ der Betragsverteilung 1. Art ist in Bild 4.40 zu sehen. Die linke Grafik zeigt die Faltung an der Stelle $x = 0$ einer Normalverteilung mit einem Mittelwert $\mu = 0.75$ und einer Standardabweichung $\sigma = 0.5$. Ein Sonderfall stellt die Faltung an der Stelle $x = \mu$ dar. Dabei vereinfacht sich die Definitionsgleichung Gleichung (4.225) der Betragsverteilung 1. Art zu

$$f(x) = \begin{cases} \frac{2}{\sigma \cdot \sqrt{2 \cdot \pi}} \cdot e^{-\frac{1}{2} \cdot \left(\frac{x}{\sigma} \right)^2} & \text{für } x \geq 0 \\ 0 & \text{für } x < 0 \end{cases} \quad (4.227)$$

Dies entspricht der rechten Grafik in Bild 4.40. Hierbei wurde eine Normalverteilung mit einem Mittelwert $\mu = 1$ und einer Standardabweichung $\sigma = 0.5$ an der Stelle des Mittelwertes μ gefaltet.

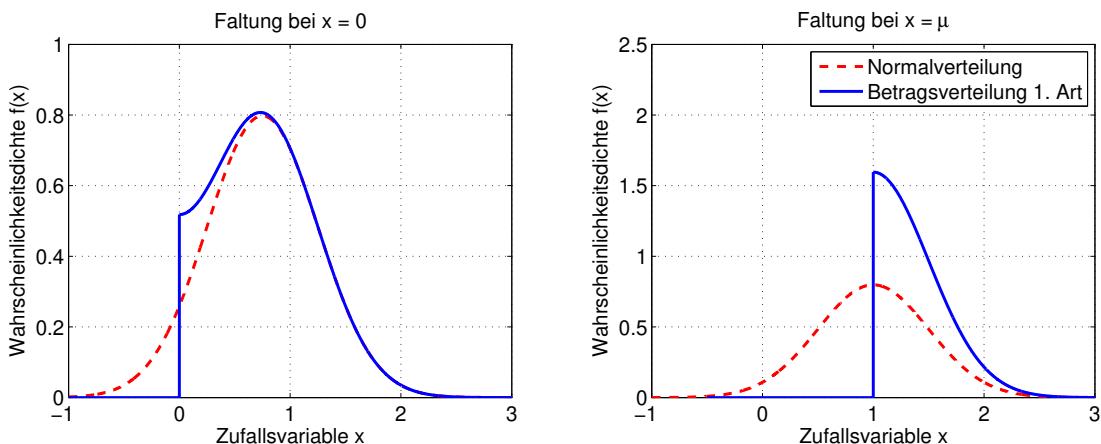


Bild 4.40: Wahrscheinlichkeitsdichte $f(x)$ der gefalteten Normalverteilung

Die Betragsverteilung 1. Art wird zur Bewertung der Prozesssicherheit im Rahmen der statistischen Prozesskontrolle eingesetzt.

Beispiel: Verteilung der Abweichung vom Sollwert bei der Widerstandsfertigung

Die Anwendung der Betragsverteilung 1. Art wird an einem Beispiel der Widerstandsfertigung verdeutlicht. Hierzu wird eine normalverteilte Widerstandsproduktion betrachtet, bei der die Widerstandswerte mit einer Standardabweichung von $\sigma = 0.75 \Omega$ um den Sollwert streuen. Der Betrag dieser Abweichungen vom Sollwert $|\Delta R|$ der Fertigung kann mithilfe der Betragsverteilung 1. Art beschrieben werden.

In der linken Grafik in Bild 4.41 ist die Wahrscheinlichkeitsdichte $f(|\Delta R|)$ der Wahrscheinlichkeitsdichte $f(|\Delta R|)$ gegenübergestellt. Die rechte Grafik zeigt die entsprechenden Verteilungsfunktionen $F(|\Delta R|)$ und $F(|\Delta R|)$.

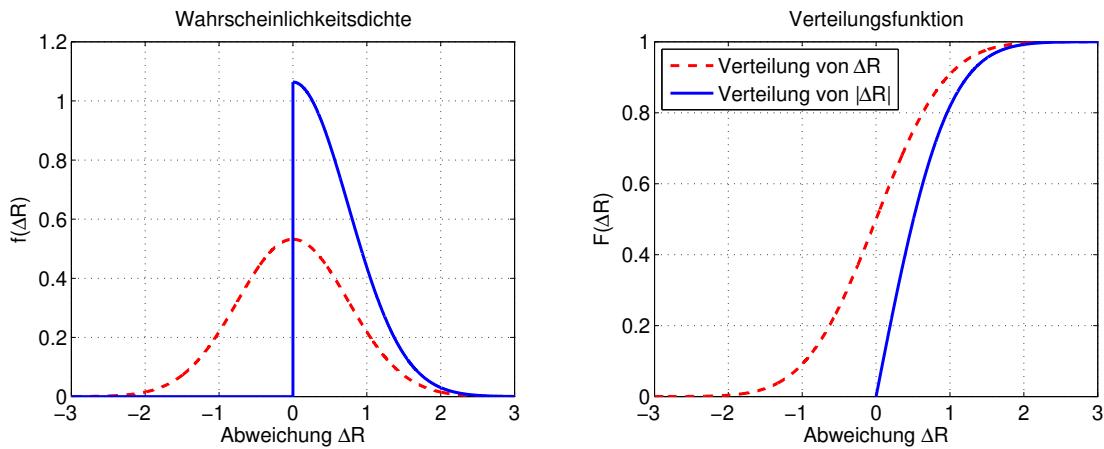


Bild 4.41: Wahrscheinlichkeitsdichte $f(\Delta R)$ beziehungsweise $f(|\Delta R|)$ und Verteilungsfunktion $F(\Delta R)$ beziehungsweise $F(|\Delta R|)$ der Widerstandsabweichung vom Sollwert

4.6.9 Zusammenfassung der stetigen Verteilungen

In diesem Abschnitt werden spezielle stetige Verteilungen vorgestellt und an Beispielen diskutiert. Dabei wird auch gezeigt, dass unter gewissen Randbedingungen die Verteilungen ineinander übergehen. Bild 4.42 stellt diese Zusammenhänge grafisch dar und gibt damit zusätzlich einen Überblick über die wichtigsten stetigen Verteilungen.

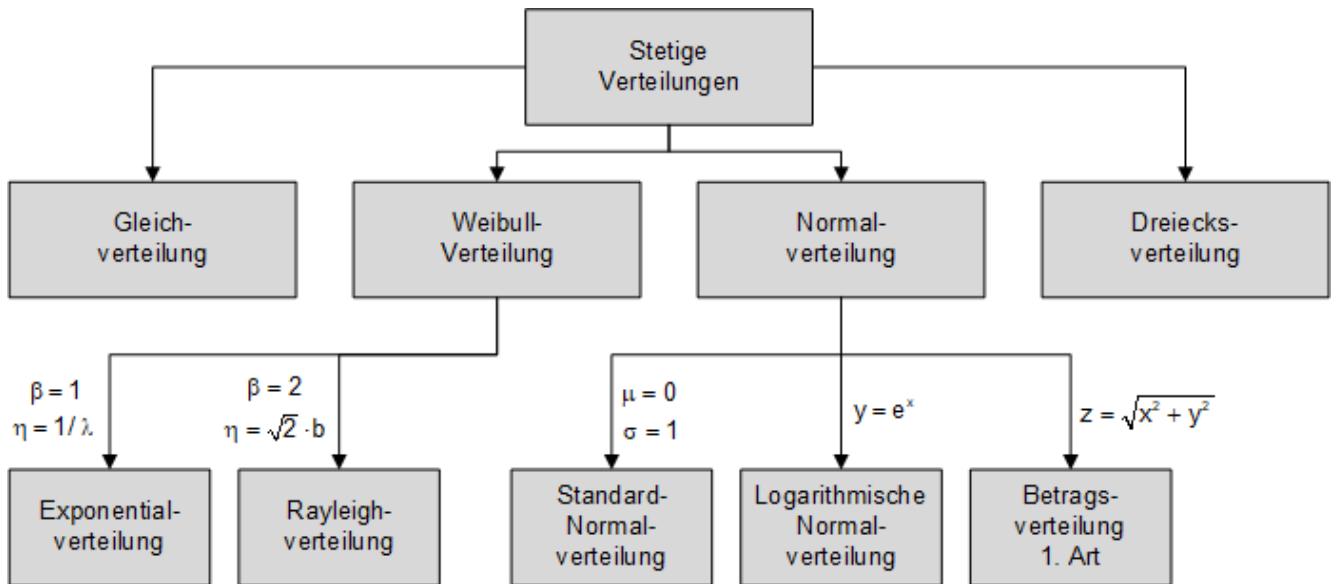


Bild 4.42: Zusammenhang der stetigen Verteilungen

Die diskutierten stetigen Wahrscheinlichkeitsverteilungen sind in Tabelle 4.15 zusammenfassend dargestellt.

Wie bereits bei den diskreten Verteilungen im vorigen Abschnitt werden auch für kontinuierliche Verteilungsfunktionen die entsprechenden MATLAB-Befehle der Statistic Toolbox tabellarisch vorgestellt. Dabei gelten die bereits eingeführten Endungen. Tabelle 4.16 zeigt eine Übersicht der MATLAB-Funktionen über die Auswahl der stetigen Verteilungen aus diesem Abschnitt.

Tabelle 4.15: Übersicht über diskrete Wahrscheinlichkeitsverteilungen

Name und Anwendung	Wahrscheinlichkeitsverteilung	Kenngrößen μ und σ^2
Gleichverteilung: Beschreibung von Wartezeiten und Diskretisierungsvorgängen	$f(x) = \frac{1}{b-a}$	$\mu = \frac{a+b}{2} \cdot \sum_{n=1}^N x_n$ $\sigma^2 = \frac{(b-a)^2}{12}$
Symmetrische Dreiecksverteilung: Toleranzverteilung bei Fertigungsprozessen	$f(x) = \begin{cases} \frac{4}{(b-a)^2} \cdot (x-a) & \text{für } a < x < \mu \\ \frac{-4}{(b-a)^2} \cdot (b-x) & \text{für } \mu < x < a \end{cases}$	$\mu = \frac{a+b}{2} \cdot \sum_{n=1}^N x_n$ $\sigma^2 = \frac{(b-a)^2}{24}$
Weibull-Verteilung: Lebensdauer von Produkten, Zeiträumen bis zum Schadensfall, Ausfall-wahrscheinlichkeit ändert sich über der Beobachtungs-zeit	$f(x) = \frac{\beta}{\mu} \cdot \left(\frac{x}{\mu}\right)^{\beta-1} \cdot e^{-\left(\frac{x}{\mu}\right)^\beta}$	siehe Abschnitt 4.6.3
Exponential-Verteilung: Lebensdauer von Produkten, Zeiträumen bis zum Schadensfall, Ausfallwahrscheinlichkeit ändert sich nicht über der Beobachtungszeit	$f(x) = \lambda \cdot e^{-\lambda \cdot x}$	$\mu = \frac{1}{\lambda}$ $\sigma^2 = \frac{1}{\lambda^2}$
Rayleigh-Verteilung: Rechtsschiefe Verteilung zur Beschreibung des Betrages zweier normalverteilter Zufallsgrößen	$f(x) = \frac{1}{b^2} \cdot x \cdot e^{-\frac{1}{2} \cdot \frac{x^2}{b^2}}$	$\mu = b \cdot \sqrt{\frac{\pi}{2}}$ $\sigma^2 = \frac{4-\pi}{2} \cdot b^2$
Normalverteilung: Approximation von Zufallsprozessen, insbesondere bei der Messwertverarbeitung und bei der Prozessregelung	$f(x) = \frac{1}{\sigma \sqrt{2\pi}} \cdot e^{-\frac{1}{2} \cdot \left(\frac{x-\mu}{\sigma}\right)^2}$	μ σ^2

<p>Logarithmische Normalverteilung: rechtsschiefe Verteilungen wie Lebensdauern, Wartezeiten oder Einkommen</p>	$f(y) = \frac{1}{\sigma_x \sqrt{2\pi}} \cdot \frac{1}{y} e^{-\frac{1}{2} \left(\frac{\ln(y) - \mu_x}{\sigma_x} \right)^2}$	$\mu_y = e^{\mu_x + \frac{\sigma_x^2}{2}}$ $\sigma_y^2 = e^{2\mu_x + \sigma_x^2} \cdot (e^{\sigma_x^2} - 1)$
<p>Betragsteilung 1. Art: Verteilungsfunktion des Betrages der Abweichungen um einen Sollwert</p>	$f(x) = \frac{1}{\sigma \sqrt{2\pi}} \cdot \left(e^{-\frac{1}{2} \left(\frac{x - \mu}{\sigma} \right)^2} + e^{-\frac{1}{2} \left(\frac{x + \mu}{\sigma} \right)^2} \right)$	

Tabelle 4.16: Übersicht über kontinuierliche stetige Wahrscheinlichkeitsverteilungen in MATLAB

Verteilung	Wahrscheinlichkeitsverteilung $f(x)$	Verteilungsfunktion $F(x)$	inverse Verteilungsfunktion $F^{-1}(x)$	Zufallszahlen-generator
Gleichverteilung	unifpdf(x,a,b)	unidcdf(x,a,b)	unidinv(P,a,b)	unifrnd(a,b)
Weibull Verteilung	wblpdf(x, η , β)	wblcdf(x, η , β)	wblinv(P, η , β)	wblrnd(η , β)
Exponentialverteilung	exppdf(x, μ)	expcdf(x, μ)	expinv(P, μ)	exprnd(μ)
Rayleigh-Verteilung	raylpdf(x,b)	raylcdf(x,b)	raylinv(P,b)	raylrnd(b)
Normalverteilung	normpdf(x, μ , σ)	normcdf(x, μ , σ)	norminv(P, μ , σ)	normrnd(μ , σ)
Logarithmische Normalverteilung	lognpdf(x, μ , σ)	logncdf(x, μ , σ)	loginv(P, μ , σ)	lognrnd(μ , σ)

Vergleichbare Python-Befehle bietet `scipy.stats`, sie sind in Tabelle 4.17 zusammengestellt.

Tabelle 4.17: Übersicht über stetige Wahrscheinlichkeitsverteilungen der Python Bibliothek `scipy.stats`

Verteilung	Wahrscheinlichkeitsverteilung $f(x)$	Verteilungsfunktion $F(x)$	inverse Verteilungsfunktion $F^{-1}(x)$	Zufallszahlen-generator
Gleichverteilung	<code>uniform.pdf(x,a,b)</code>	<code>uniform.cdf(x,a,b)</code>	<code>uniform.ppf(P,a,b)</code>	<code>uniform.rvs(a,b)</code>
Weibull Verteilung	<code>weibull_min.pdf(x,η,β)</code>	<code>weibul_min.cdf(x,η,β)</code>	<code>weibul_min.ppf(P,η,β)</code>	<code>weibul_min.rvs(η,β)</code>
Exponentialverteilung	<code>expon.pdf(x,μ)</code>	<code>expon.cdf(x,μ)</code>	<code>expon.ppf(P,μ)</code>	<code>expon.rvs(μ)</code>
Rayleigh-Verteilung	<code>raylpdf(x,b)</code>	<code>raylcdf(x,b)</code>	<code>raylinv(P,b)</code>	<code>raylrnd(b)</code>
Normalverteilung	<code>normpdf(x,μ,σ)</code>	<code>normcdf(x,μ,σ)</code>	<code>norminv(P,μ,σ)</code>	<code>normrnd(μ,σ)</code>
Logarithmische Normalverteilung	<code>lognpdf(x,μ,σ)</code>	<code>logncdf(x,μ,σ)</code>	<code>loginv(P,μ,σ)</code>	<code>lognrnd(μ,σ)</code>

4.7 Prüf- oder Testverteilungen

Viele statistische Aufgabenstellungen lassen sich mithilfe der Normalverteilung beschreiben. Oftmals müssen ihre Parameter μ und σ auf Basis von Stichproben geschätzt werden. Auf die Schätzung von Parametern wird in Kapitel 5 eingegangen. Die im Folgenden beschriebenen univariaten Prüf- oder Testverteilungen werden dazu genutzt, Aussagen zu Vertrauensbereichen der Parameter zu machen, die für die Verteilung auf Basis von Stichproben bestimmt wurden. Zu den univariaten Testverteilungen zählen die t-Verteilung von Student, die Chi-Quadrat-Verteilung und die F-Verteilung von Fisher.

Grundlage für die Berechnung von Parametern und ihren Vertrauensbereichen ist eine aus der Grundgesamtheit entnommene Stichprobe mit N unabhängigen Werten x_1, x_2, \dots, x_N . Die Auswahl der Stichprobe ist dabei zufällig, ihre Werte x_n werden bei identischer Grundgesamtheit von Mal zu Mal variieren. Jeder Stichprobenwert x_n kann deshalb als eine Realisierung der Zufallsvariablen x aufgefasst werden. Prüf- und Testverteilung sind damit Funktionen von N unabhängigen Zufallsvariablen. Ihre Herleitungen sind deshalb eigentlich Bestandteil der multivariaten Statistik. Da sie zum Lösen eindimensionaler Fragestellungen der folgenden Kapitel erforderlich sind, werden sie an dieser Stelle ohne Herleitung eingeführt. Die mathematischen Herleitungen sind im Anhang A.1 dargestellt, Voraussetzung für das Verständnis dieser Herleitungen ist das Wissen zu multivariaten Wahrscheinlichkeitsverteilungen aus Kapitel 7 und Kapitel 8.

4.7.1 Chi-Quadrat-Verteilung

Die Chi-Quadrat-Verteilung ist eine Verteilung, die bei der Schätzung von Verteilungsparametern, beispielsweise der Varianz, Anwendung findet. Ausgangspunkt für die Chi-Quadrat-Verteilung ist eine Stichprobe von N Werten x_1, x_2, \dots, x_N . Die Werte stammen aus einer normalverteilten Grundgesamtheit mit einem Mittelwert μ und einer Standardabweichung σ . Dann weist die Größe

$$\chi = x_1^2 + x_2^2 + \dots + x_N^2 \quad (4.228)$$

eine Chi-Quadrat-Verteilung mit $\nu = N$ Freiheitsgraden auf. Die Abhängigkeit von den Freiheitsgraden wird in einigen Fällen mit der Schreibweise

$$\chi = \chi(\nu) \quad (4.229)$$

verdeutlicht. Die Chi-Quadrat-Verteilung besitzt die Wahrscheinlichkeitsdichte

$$f(\chi) = \begin{cases} K_\nu \cdot \chi^{\frac{\nu-2}{2}} \cdot e^{-\frac{\chi}{2}} & \text{für } \chi > 0 \\ 0 & \text{sonst} \end{cases} \quad (4.230)$$

Durch Integration ergibt sich die Verteilungsfunktion

$$F(\chi) = K_\nu \cdot \int_0^\chi \xi^{\frac{\nu-2}{2}} \cdot e^{-\frac{\xi}{2}} d\xi \quad (4.231)$$

Die Konstante K_ν ergibt sich aus der Normierung der Wahrscheinlichkeitsdichte $f(\chi)$. Nach den Axiomen der Wahrscheinlichkeitstheorie muss für sie gelten

$$F(\infty) = K_\nu \cdot \int_0^\infty \chi^{\frac{\nu-2}{2}} \cdot e^{-\frac{\chi}{2}} d\chi = 1 \quad (4.232)$$

Diese Beziehung führt zu der Konstante K_ν von

$$K_\nu = \frac{1}{2^{\frac{\nu}{2}} \cdot \Gamma\left(\frac{\nu}{2}\right)} \quad (4.233)$$

Dabei ist der Ausdruck $\Gamma(\alpha)$ die in Kapitel 4.6.3 eingeführte Gamma-Funktion. Die Chi-Quadrat-Verteilung hat den Mittelwert

$$\mu = \nu = N \quad (4.234)$$

und die Varianz

$$\sigma^2 = 2 \cdot \nu = 2 \cdot N \quad (4.235)$$

Durch den Parameter ν können die Form und Lage der Chi-Quadrat-Verteilung beeinflusst werden. Bild 4.43 stellt die Wahrscheinlichkeitsdichte $f(\chi)$ und die Verteilungsfunktion $F(\chi)$ der Chi-Quadrat-Verteilung für unterschiedliche Freiheitsgrade ν dar.

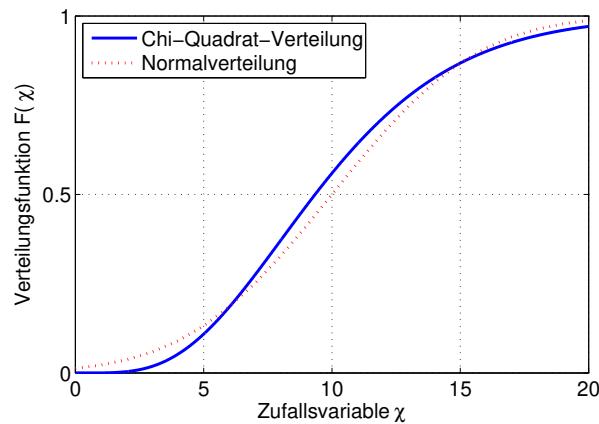


Bild 4.43: Darstellung der Chi-Quadrat-Verteilung für unterschiedliche Freiheitsgrade ν

Auch die Chi-Quadrat-Funktion lässt sich durch die Normalverteilung annähern. Die Zufallsvariable χ ist asymptotisch normalverteilt mit dem Mittelwert μ aus Gleichung (4.234) und der Varianz σ^2 aus Gleichung (4.235). Bild 4.44 vergleicht grafisch die Chi-Quadrat-Verteilung mit der Normalverteilung.

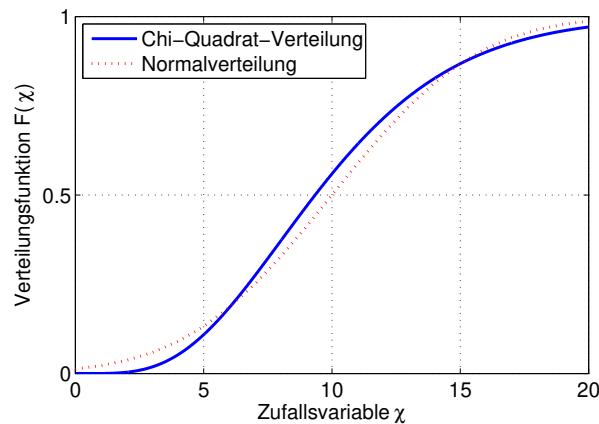


Bild 4.44: Näherung der Chi-Quadrat-Verteilung mit $N = 10$ durch eine Normalverteilung mit $\mu = 10$ und $\sigma^2 = 20$

Es kann gezeigt werden, dass die Varianz einer Stichprobe von N Werten

$$s^2 = \frac{1}{N-1} \cdot \sum_{n=1}^N (x_n - \bar{x})^2 \quad (4.236)$$

aus einer normalverteilten Grundgesamtheit mit einem Mittelwert μ und einer Varianz σ^2 eine Chi-Quadrat-Verteilung mit $v = N - 1$ Freiheitsgraden besitzt, wenn sie vorher mit dem Faktor v multipliziert und durch die Varianz σ^2 dividiert wurde.

$$\chi = (N-1) \cdot \frac{s^2}{\sigma^2} = v \cdot \frac{s^2}{\sigma^2} \quad (4.237)$$

Anwendung findet die Chi-Quadrat-Verteilung bei statistischen Aussagen zu der Varianz einer Grundgesamtheit, die aus einer Stichprobe bestimmt wird. Das ist zum Beispiel bei der Berechnung von Vertrauensintervallen für Varianzen und bei Hypothesentests zu Varianzen der Fall.

4.7.2 t-Verteilung von Student

Die Student t-Verteilung ist eine Wahrscheinlichkeitsverteilung, die von William Sealey Gosset entwickelt wurde. Er hatte festgestellt, dass der standardisierte Mittelwert normalverteilter Daten nicht mehr normalverteilt ist, wenn die Varianz des Merkmals unbekannt ist und mit der Stichprobenvarianz geschätzt werden muss. Die Herleitung wurde erstmals 1908 veröffentlicht, während Gosset in einer Guinness-Brauerei arbeitete. Da sein Arbeitgeber die Veröffentlichung nicht gestattete, veröffentlichte Gosset sie unter dem Pseudonym Student. Die zugehörige Theorie wurde erst durch die Arbeiten von R. A. Fisher belegt, der die Verteilung Students distribution (Students Verteilung) nannte.

Zur Erklärung des Hintergrundes dieser Verteilung wird die eingangs diskutierte Stichprobe von N Werten x_1, x_2, \dots, x_N angenommen. Die Werte stammen aus einer Grundgesamtheit mit Standard-Normalverteilung, die einen Mittelwert von $\mu = 0$ und eine unbekannte Standardabweichung σ besitzt. Aus der zufällig ausgewählten Stichprobe ergibt sich der Stichproben-Mittelwert

$$\bar{x} = \frac{1}{N} \cdot \sum_{n=1}^N x_n \quad (4.238)$$

und die Stichproben-Standardabweichung

$$s = \sqrt{\frac{1}{N-1} \cdot \sum_{n=1}^N (x_n - \bar{x})^2} \quad (4.239)$$

Gosset hat gezeigt, dass der Quotient

$$t = \frac{z}{\sqrt{\frac{\chi}{(N-1)}}} \quad (4.240)$$

einer standardnormalverteilten Größe z und einer Größe χ mit Chi-Quadrat-Verteilung eine t-Verteilung mit $v = N - 1$ Freiheitsgraden aufweist. Damit weist zum Beispiel die Größe

$$t = \frac{\frac{x - \mu}{\sigma}}{\sqrt{\frac{1}{(N-1)} \cdot (N-1) \cdot \frac{s^2}{\sigma^2}}} = \frac{\frac{x - \mu}{\sigma}}{\sqrt{\frac{s^2}{\sigma^2}}} = \frac{\frac{x - \mu}{\sigma}}{\frac{s}{\sigma}} = \frac{x - \mu}{s} \quad (4.241)$$

eine t-Verteilung mit $\nu = N - 1$ Freiheitsgraden auf. Die t-Verteilung hat die Wahrscheinlichkeitsdichte $f(t)$ von

$$f(t) = \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\sqrt{\nu \cdot \pi} \cdot \Gamma\left(\frac{\nu}{2}\right) \cdot \left(1 + \frac{t^2}{\nu}\right)^{\frac{\nu+1}{2}}} \quad (4.242)$$

und die Verteilungsfunktion $F(t)$

$$F(t) = \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\sqrt{\nu \cdot \pi} \cdot \Gamma\left(\frac{\nu}{2}\right)} \cdot \int_{-\infty}^t \frac{1}{\left(1 + \frac{\tau^2}{\nu}\right)^{\frac{\nu+1}{2}}} d\tau \quad (4.243)$$

Dabei ist der Ausdruck $\Gamma(\alpha)$ die in Kapitel 4.6.3 eingeführte Gamma-Funktion. Aus Gleichung (4.242) und Gleichung (4.243) wird ersichtlich, dass die t-Verteilung nur von dem Parameter ν , der Anzahl von Freiheitsgraden abhängig ist. Für $\nu = 1$ ergibt sich eine Verteilung, die keinen Mittelwert besitzt [Fahr06]. Für $\nu > 1$ hat die t-Verteilung wegen der Achsensymmetrie den Mittelwert 0. Für $\nu \leq 2$ hat die t-Verteilung keine Varianz, für $\nu > 2$ ergibt sich die Varianz

$$\sigma^2 = \frac{\nu}{\nu - 2} \quad (4.244)$$

Mit wachsender Anzahl an Freiheitsgraden ν strebt die Wahrscheinlichkeitsdichte der t-Verteilung gegen die Wahrscheinlichkeitsdichte der standardisierten Normalverteilung, da gilt

$$\lim_{\nu \rightarrow \infty} \sigma^2 = \lim_{\nu \rightarrow \infty} \left(\frac{\nu}{\nu - 2} \right) = 1 \quad (4.245)$$

Bild 4.45 stellt die t-Verteilung für $\nu = 2, 5$ und 25 im Vergleich zur Standard-Normalverteilung mit $\mu = 0$ und $\sigma^2 = 1$ dar.

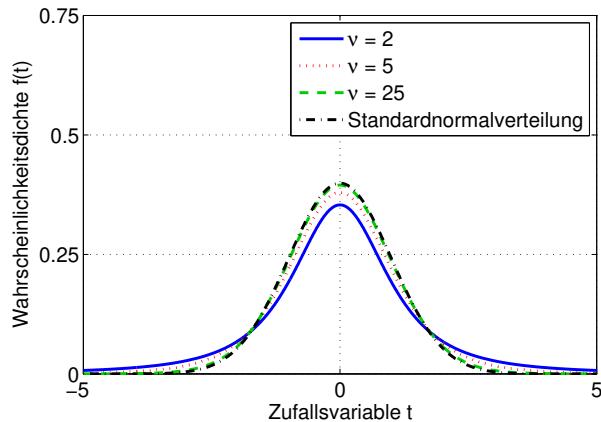


Bild 4.45: t-Verteilung mit $\nu = 2, 5$ und 25 im Vergleich zur standardisierten Normalverteilung

Die t-Verteilung ist breiter als die Normalverteilung mit einer Standardabweichung von $\sigma = s$. Dieser Zusammenhang ergibt sich daraus, dass die Standardabweichung aufgrund einer Stichprobe geschätzt wird und damit eine größere Unsicherheit besteht als bei bekannter Standardabweichung.

Die t-Verteilung wird immer verwendet, wenn der Mittelwert einer Stichprobe mit unbekannter Varianz bewertet werden soll. Zum Beispiel wird sie bei Hypothesentests dazu verwendet, um abzuschätzen, ob ein Stichprobenwert zufällig von dem Mittelwert abweicht oder ob die Abweichung systematisch ist. Außerdem wird die t-Verteilung dazu verwendet, einen Konfidenzbereich für einen Mittelwert einer Grundgesamtheit mit unbekannter Varianz anzugeben.

4.7.3 f-Verteilung von Fisher

Die f-Verteilung wurde von Ronald Aylmer Fisher entwickelt. Sie beschreibt die Verteilung einer Zufallsvariable f , die sich aus dem Quotienten zweier chi-quadrat-verteilter Zufallsvariablen ergibt, die auf ihre entsprechende Anzahl Freiheitsgrade normiert sind.

$$f = \frac{\chi(\nu_1)/\nu_1}{\chi(\nu_2)/\nu_2} \quad (4.246)$$

Die Zufallsvariable f aus Gleichung (4.246) besitzt eine f-Verteilung mit ν_1 und ν_2 Freiheitsgraden. Die Wahrscheinlichkeitsdichte der f-Verteilung ist definiert zu

$$f(f) = \frac{\Gamma\left(\frac{\nu_1 + \nu_2}{2}\right)}{\Gamma\left(\frac{\nu_1}{2}\right) \cdot \Gamma\left(\frac{\nu_2}{2}\right)} \cdot \left(\frac{\nu_1}{\nu_2}\right)^{\frac{\nu_1}{2}} \cdot \frac{f^{\frac{\nu_1}{2}-1}}{\left(\frac{\nu_1 \cdot f}{\nu_2} + 1\right)^{\frac{\nu_1 + \nu_2}{2}}} \quad (4.247)$$

Die Verteilungsfunktion $F(f)$ kann nicht analytisch berechnet werden, sondern muss numerisch bestimmt werden. Für $\nu_2 \leq 2$ besitzt die Verteilung keinen Mittelwert. Für $\nu_2 > 2$ hat die f-Verteilung den Mittelwert

$$\mu = \frac{\nu_2}{\nu_2 - 2} \quad (4.248)$$

Für $\nu_2 \leq 4$ hat die f-Verteilung keine Varianz, für $\nu_2 > 4$ ergibt sich die Varianz zu

$$\sigma^2 = \frac{2 \cdot \nu_2^2 \cdot (\nu_2 + \nu_1 - 2)}{\nu_1 \cdot (\nu_2 - 4) \cdot (\nu_2 - 2)^2} \quad (4.249)$$

Bild 4.46 stellt die Wahrscheinlichkeitsdichte $f(f)$ und die Verteilungsfunktion $F(f)$ der f-Verteilung für Wertekombinationen $(\nu_1, \nu_2) = (4, 4), (4, 20), (20, 4)$ und $(20, 20)$ dar.

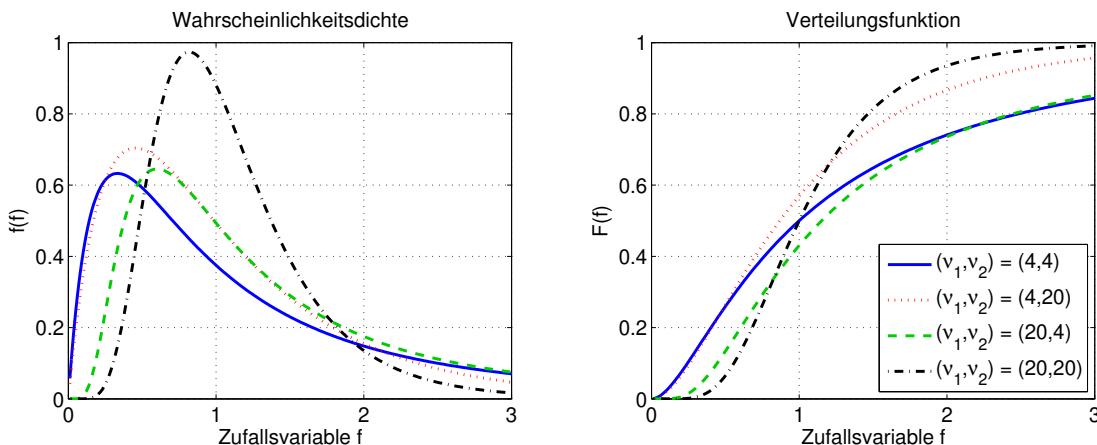


Bild 4.46: F-Verteilung für Wertekombinationen $(\nu_1, \nu_2) = (4, 4), (4, 20), (20, 4)$ und $(20, 20)$

Dieser theoretischen Definition der f-Verteilung liegt die Aufgabe zugrunde, eine Verteilungsfunktion für das Verhältnis von zwei Varianzen auf Basis von zwei Stichproben abzuschätzen. Ausgangspunkt sind zwei Stichproben x_1, x_2, \dots, x_N und y_1, y_2, \dots, y_M . Die Stichproben stammen aus unterschiedlichen normalverteilten Grundgesamtheiten, die die Varianz σ_X^2 beziehungsweise σ_Y^2 besitzen.

Mit den Erkenntnissen aus Gleichung (4.247) zur Beschreibung einer chi-quadrat-verteilten Größe und der allgemeinen Definition einer f-verteilten Zufallsvariablen f aus Gleichung (4.246) ergibt das Verhältnis der empirischen Varianzen

$$f = \frac{\chi(\nu_1)/\nu_1}{\chi(\nu_2)/\nu_2} = \frac{\frac{\nu_X \cdot s_X^2}{\sigma_X^2} \cdot \frac{1}{\nu_X}}{\frac{\nu_Y \cdot s_Y^2}{\sigma_Y^2} \cdot \frac{1}{\nu_Y}} = \frac{s_X^2}{s_Y^2} \cdot \frac{\sigma_Y^2}{\sigma_X^2} \quad (4.250)$$

eine f-Verteilung mit den Freiheitsgraden v_1 und v_2 .

Die f-Verteilung kann daher zum Beispiel bei der Varianzanalyse verwendet werden, um festzustellen, ob die Grundgesamtheiten zweier Stichproben die gleiche Varianz haben. Darüber hinaus wird sie bei Regressions- und Varianzanalysen eingesetzt, um zu testen, ob die jeweiligen Einflussgrößen signifikant sind oder nicht.

4.7.4 Zusammenfassung der Prüf- oder Testverteilungen

Die t-Verteilung und die Chi-Quadrat-Verteilung leiten sich aus der Normalverteilung ab, die f-Verteilung kann durch die Logarithmische Normalverteilung angenähert werden. Diese Zusammenhänge sind in Bild 4.47 grafisch dargestellt.

Es wird sich zeigen, dass diese Verteilungen für das Lösen von Aufgabenstellungen anwenden lassen, bei den Parameter geschätzt werden und eine Aussage über die Genauigkeit der Schätzung gemacht werden soll. Das dabei verwendete Grundprinzip ist immer dasselbe. Auf Basis einer Stichprobe wird ein Parameter einer Verteilung geschätzt. Da es sich um eine zufällige Stichprobe handelt, ist auch der geschätzte Parameter eine Zufallsgröße. Mithilfe der Testverteilungen wird die Sicherheit bewertet, mit der der Parameter bestimmt wurde. Ihre Wahrscheinlichkeitsdichten, Mittelwerte und Standardabweichungen sowie ihre Anwendungsbereiche sind in Tabelle 4.18 wiedergegeben.

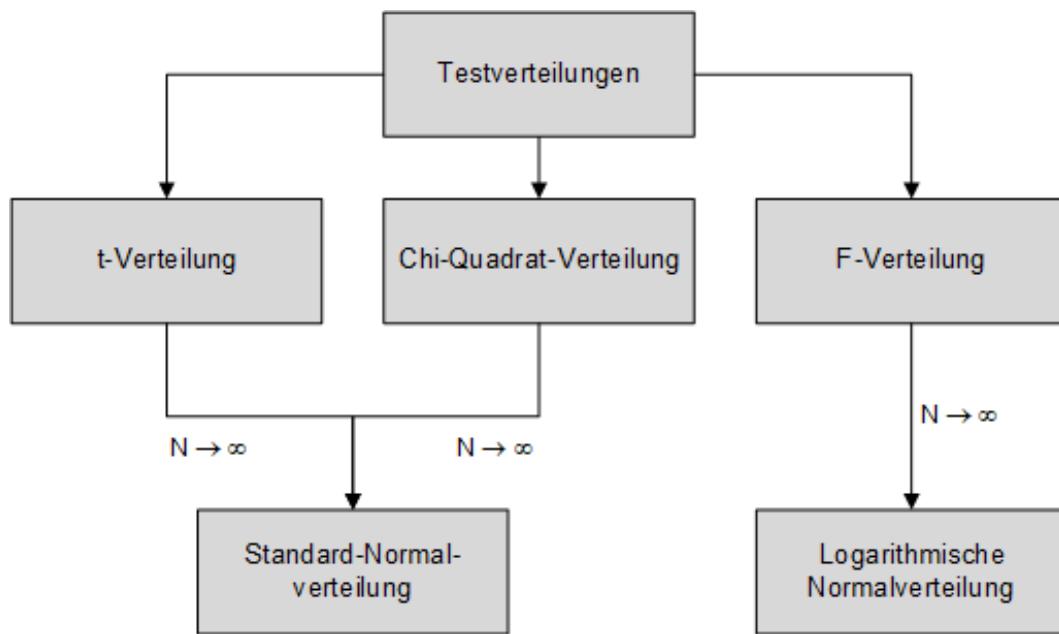


Bild 4.47: Zusammenhang der eingeführten Testverteilungen

Tabelle 4.18: Übersicht über die Testverteilungen und ihre Anwendungen

Name und Anwendung	Wahrscheinlichkeitsverteilung	Kenngrößen μ und σ^2
t-Verteilung mit ν Freiheitsgraden Statistische Bewertung von Mittelwerten auf Basis von Stichproben bei unbekannter Varianz	$f(t) = \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\sqrt{\nu \cdot \pi} \cdot \Gamma\left(\frac{\nu}{2}\right) \cdot \left(1 + \frac{t^2}{\nu}\right)^{\frac{\nu+1}{2}}$	$\mu = 0$ $\sigma^2 = \frac{\nu}{\nu-2}$
Chi ² -Verteilung mit ν Freiheitsgraden Statistische Bewertung der Varianz einer Grundgesamtheit auf Basis von Stichproben	$f(\chi) = K_\nu \cdot \chi^{\frac{\nu-2}{2}} \cdot e^{-\frac{\chi}{2}}$	$\mu = \nu$ $\sigma^2 = 2 \cdot \nu$
f-Verteilung mit ν_1, ν_2 Freiheitsgraden Statistische Vergleich der Varianz zweier Grundgesamtheiten auf Basis von Stichproben	$f(f) = \frac{\Gamma\left(\frac{\nu_1 + \nu_2}{2}\right)}{\Gamma\left(\frac{\nu_1}{2}\right) \cdot \Gamma\left(\frac{\nu_2}{2}\right)} \cdot \left(\frac{\nu_1}{\nu_2}\right)^{\frac{\nu_1}{2}} \cdot \frac{f^{\frac{\nu_1}{2}-1}}{\left(\frac{\nu_1 \cdot f}{\nu_2} + 1\right)^{\frac{\nu_1 + \nu_2}{2}}}$	$\mu = \frac{\nu_2}{\nu_2 - 2}$ $\sigma^2 = \frac{2 \cdot \nu_2^2 \cdot (\nu_2 + \nu_1 - 2)}{\nu_1 \cdot (\nu_2 - 4) \cdot (\nu_2 - 2)^2}$

In der Statistic-Toolbox von MATLAB sind die Prüf- und Testverteilungen implementiert. Tabelle 4.19 gibt einen Überblick über die entsprechenden Funktionen.

Tabelle 4.19: Übersicht über die Testverteilungen in MATLAB

Verteilung	Dichtefunktion $f(x)$	Wahrscheinlichkeitsfunktion $F(x)$	inverse Wahrscheinlichkeitsfunktion $F^{-1}(x)$
t-Verteilung	$tpdf(x,\nu)$	$tcdf(x,\nu)$	$tinv(P,\nu)$
Chi-Quadrat-Verteilung	$chi2pdf(x,\nu)$	$chi2cdf(x,\nu)$	$chi2inv(P,\nu)$
f-Verteilung	$fpdf(x,\nu_1,\nu_2)$	$fcdf(x,\nu_1,\nu_2)$	$finv(P,\nu_1,\nu_2)$

Auch in Python sind entsprechende Funktionen implementiert, Tabelle 4.20 gibt einen Überblick über diese Funktionen.

Tabelle 4.20: Übersicht über die Testverteilungen in der Python Bibliothek `scipy.stats`

Verteilung	Dichtefunktion $f(x)$	Wahrscheinlichkeitsfunktion $F(x)$	inverse Wahrscheinlichkeitsfunktion $F^{-1}(x)$
t-Verteilung	<code>t.pdf(x,nu)</code>	<code>t.cdf(x,nu)</code>	<code>t.ppf(P,nu)</code>
Chi-Quadrat-Verteilung	<code>chi2pdf(x,nu)</code>	<code>chi2cdf(x,nu)</code>	<code>chi2inv(P,nu)</code>
f-Verteilung	<code>fpdf(x,nu_1,nu_2)</code>	<code>fcdf(x,nu_1,nu_2)</code>	<code>finv(P,nu_1,nu_2)</code>

4.8 Literatur

- [Krey91] Kreyszig, Erwin: Statistische Methoden und ihre Anwendungen
4., unveränderter Nachdruck der 7. Auflage
Vandenhoeck & Ruprecht, Göttingen, 1991
- [Fahr06] Fahrmeir, Ludwig; Künstler, Rita; Pigeot, Iris; Tutz, Gerhard: Der Weg zur Datenanalyse
6. Auflage
Springer Berlin Heidelberg New York, 2006
- [Ross06] Ross, M. Sheldon: Statistik für Ingenieure und Naturwissenschaftler
3. Auflage
Spektrum Akademischer Verlag, München, 2006

5 Schätzung von unbekannten Parametern einer Verteilung

In Kapitel 3 werden die grafische Darstellung von Datensätzen und die zusammenfassende Beschreibung der Daten durch Lage- und Streuungskennwerte eingeführt. Diese Daten können als Stichprobe einer Grundgesamtheit verstanden werden. In diesem Kapitel werden Schlüsse aus einer Stichprobe für die zugehörige Grundgesamtheit gezogen. Insbesondere werden auf Basis einer Stichprobe Parameter der Grundgesamtheit geschätzt und der Bereich für zukünftige Werte prognostiziert.

5.1 Zielsetzung und Problematik der Parameterschätzung

Im Rahmen der Design For Six Sigma Methoden werden Mittelwert μ und Standardabweichung σ einer Verteilung auf Basis des Mittelwertes \bar{x} und der Standardabweichung s der Stichprobe geschätzt. Der Stichprobenmittelwert ist der Schätzwert für den Mittelwert der Grundgesamtheit.

$$\mu \approx \bar{x} \quad (5.1)$$

In gleicher Weise wird die Stichprobenvarianz als Schätzwert für die Varianz der Grundgesamtheit verwendet.

$$\sigma^2 \approx s^2 \quad (5.2)$$

Tabelle 5.1: Schätzung der Parameter einer Grundgesamtheit über eine Stichprobe

Charakteristik	Stichprobe	Grundgesamtheit
Mittelwert	\bar{x}	ν
	Stichproben-Mittelwert \bar{x} schätzt den Mittelwert der Grundgesamtheit ν	
Varianz	s^2	σ^2
	Stichproben-Varianz s^2 schätzt die Varianz der Grundgesamtheit σ^2	

Tabelle 5.1 stellt den Zusammenhang zwischen Grundgesamtheit und Stichprobe tabellarisch zusammen.

Mit den geschätzten Parametern ergibt sich eine Verteilung der Grundgesamtheit. Bild 5.1 verbindet die als Stabdiagramm dargestellte Stichprobe und die geschätzte Wahrscheinlichkeitsdichte der Grundgesamtheit.

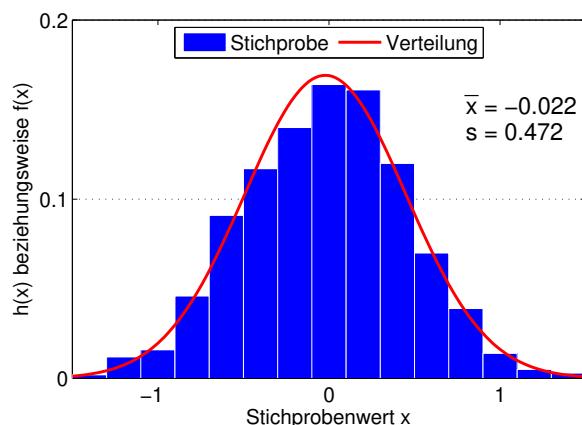


Bild 5.1: Stichprobe und auf Basis der Stichprobe geschätzte Wahrscheinlichkeitsdichte der Grundgesamtheit

Um die Problematik der beurteilenden Statistik zu verdeutlichen, wird eine Stichprobe aus einem Datensatz analysiert, der eine normalverteilte Grundgesamtheit mit einem Mittelwert von $\mu = 0$ und einer Standardabweichung von $\sigma = 0.5$ aufweist. Bild 5.2 zeigt die relativen Häufigkeiten zweier Stichproben mit einem Umfang von jeweils 10 Werten.

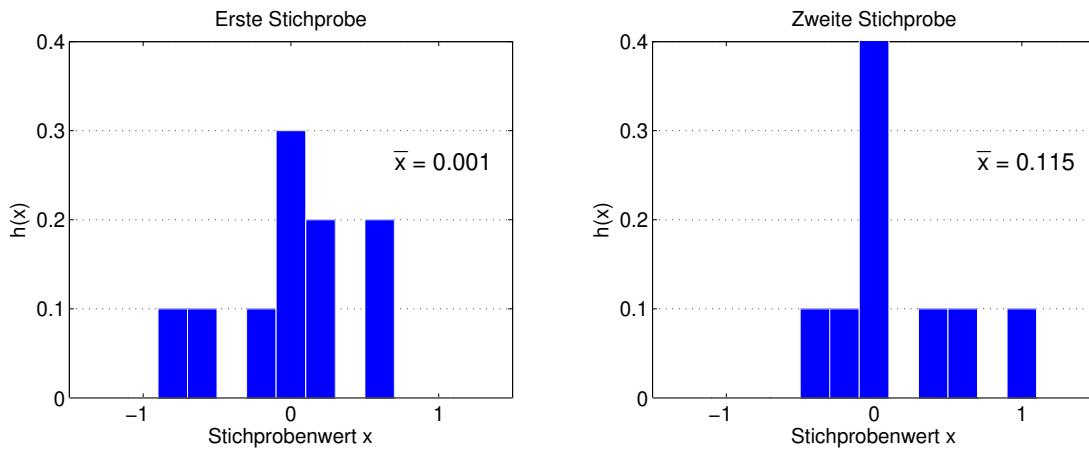


Bild 5.2: Häufigkeitsverteilungen zweier unterschiedlicher Stichproben derselben Grundgesamtheit mit einem Stichprobenumfang von $N = 10$

Obwohl die beiden Stichproben mit dem Umfang von 10 Teilen aus derselben Grundgesamtheit stammen, weichen ihre Mittelwerte \bar{x} stark voneinander ab. Das Beispiel zeigt, dass der Mittelwert der Grundgesamtheit auf Basis einer Stichprobe nur geschätzt werden kann. Er hängt von der Auswahl der Stichprobenwerte ab. Der über eine Stichprobe geschätzte Mittelwert ist damit selber eine Zufallsgröße.

Um den Einfluss des Stichprobenumfangs zu hinterfragen, wird auf derselben Datenbasis der Stichprobenumfang in Schritten von $N = 10, 100$ und 1000 Werte erweitert. Das Ergebnis ist in Bild 5.3 dargestellt.

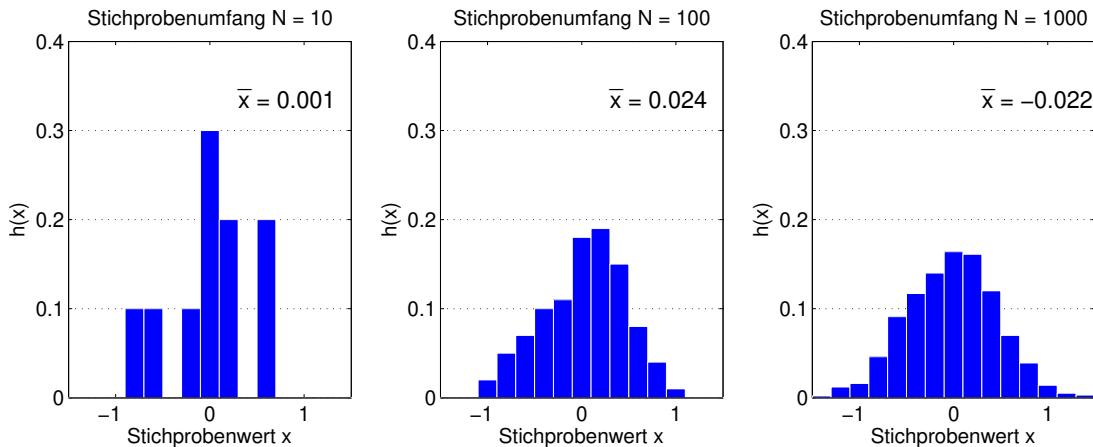


Bild 5.3: Häufigkeitsverteilungen von Stichproben derselben Grundgesamtheit mit einem Stichprobenumfang von $N = 10, 100, 1000$

Mit steigendem Stichprobenumfang nähert sich der Mittelwert der Stichprobe \bar{x} dem wahren Mittelwert $\mu = 0$ an. Die Schätzung des Mittelwertes wird also mit wachsendem Stichprobenumfang genauer. Aus Genauigkeitsgründen erscheint es deshalb erstrebenswert, möglichst viele Stichprobenwerte zu analysieren. Allerdings sprechen finanzielle, zeitliche oder prinzipielle Gründe für einen geringen Stichprobenumfang. Damit stellt sich die Frage, wie groß der Stichprobenumfang für eine bestimmte Aufgabe sein muss. Diese Frage wird mit der Bestimmung von Konfidenzintervallen beantwortet.

Die Darstellung der Stichprobe mit wachsendem Stichprobenumfang in Bild 5.3 verdeutlicht aber noch eine zweite Fragestellung. Bei einem geringen Stichprobenumfang ist die zugrunde liegende Verteilung nicht erkennbar. Sie ist erst bei größeren Stichprobenumfängen zu erkennen. Eine weitere Aufgabenstellung widmet sich deshalb der Frage, wie sicher es sich bei der vorliegenden Stichprobe um eine bestimmte Verteilung handelt. Diese Frage wird in Kapitel 12 mithilfe des sogenannten Wahrscheinlichkeitsnetzes beantwortet. Die Darstellungen in diesem Kapitel beschränken sich auf normalverteilte Zufallsvariable.

In diesem Kapitel wird die Theorie der Stichprobenentnahme aus einer unendlichen Grundgesamtheit vorgestellt. Dabei wird davon ausgegangen, dass die Wahrscheinlichkeit eines Stichprobenwertes nicht von den anderen Stichprobenwerten beeinflusst wird. Das ist insbesondere für eine unendlich große Grundgesamtheit der Fall, woraus sich die Bezeichnung der Theorie ergibt. Praktisch gesehen wird diese Annahme aber auch dann erfüllt, wenn die Grundgesamtheit sehr viel größer ist als die Anzahl der Stichproben, sodass die Annahme in nahezu allen praktischen Fällen berechtigt ist.

Weiterhin werden bei den Herleitungen Rechenregeln für mehrere unabhängige Zufallsvariablen benötigt, die in Kapitel 8 ausführlich dargestellt sind. Um Fragestellungen für univariate Verteilungen abschließen zu können, wird die Berechnung von Konfidenzbereichen vorgezogen.

5.2 Erwartungstreue der Parameterschätzung

Die Grundgesamtheit des zu analysierenden Wahrscheinlichkeitsprozesses weist einen Mittelwert μ und eine Varianz σ^2 auf. Beide Kenngrößen werden auf Basis einer Stichprobe mit den Werten x_1, x_2, \dots, x_N geschätzt. Dabei stellt sich die Frage, ob die geschätzten Parameter bei einer großen Anzahl von Stichprobenwerten mit dem erwarteten Parameter der Grundgesamtheit übereinstimmen.

5.2.1 Schätzen des Mittelwertes einer Grundgesamtheit

Der Mittelwert aus einer Stichprobe x_1, x_2, \dots, x_N wird als Schätzwert des Mittelwertes μ der Grundgesamtheit angesehen. Die Schätzung ergibt sich aus der Beziehung

$$\mu \approx \bar{x} = \frac{1}{N} \cdot (x_1 + x_2 + \dots + x_N) = \frac{1}{N} \cdot \sum_{n=1}^N x_n \quad (5.3)$$

Da die Stichprobenwerte zufällig ausgewählt wurden, ist der sich ergebende Mittelwert der Stichprobe \bar{x} ebenfalls eine Zufallsvariable. Deshalb wird die Erwartungstreue der Schätzung untersucht. Eine Schätzung wird als erwartungstreu bezeichnet, wenn der Schätzwert der Stichprobe und der entsprechende Wert der Grundgesamtheit für große Stichprobenumfänge übereinstimmen. Da jede einzelne Zufallsvariable x_n nach den Ausführungen in Kapitel 4.3.2 den Mittelwert

$$E(x_n) = \mu \quad (5.4)$$

besitzt, hat die Summe der Zufallsvariablen x_1, x_2, \dots, x_N den Erwartungswert $N \cdot \mu$. Für den Stichprobenmittelwert \bar{x} ergibt sich damit der Erwartungswert

$$E(\bar{x}) = E\left(\frac{1}{N} \cdot (x_1 + x_2 + \dots + x_N)\right) = \frac{1}{N} \cdot N \cdot E(x_n) = \mu \quad (5.5)$$

Die Erwartungswerte des Stichprobenmittelwertes und der Grundgesamtheit stimmen demnach überein. Die Schätzung wird als erwartungstreu bezeichnet.

Nach den Rechenregeln für mehrere unabhängige Zufallsvariablen, die in Kapitel 8 ausführlich dargestellt sind, errechnet sich die Varianz einer Summe von unabhängigen Zufallszahlen

$$y = x_1 + x_2 + \dots + x_n \quad (5.6)$$

aus der Summe der einzelnen Varianzen.

$$\sigma_y^2 = \sigma_{x_1}^2 + \sigma_{x_2}^2 + \dots + \sigma_{x_N}^2 = \sum_{n=1}^N \sigma_{x_n}^2 \quad (5.7)$$

Weiterhin ergibt sich führt nach diesen Regeln ein Faktor bei einer Skalierung von Zufallsvariablen der Form

$$y = \frac{1}{N} \cdot x \quad (5.8)$$

für die Varianzen zu der die Beziehung

$$\sigma_y^2 = \frac{1}{N^2} \cdot \sigma_x^2 \quad (5.9)$$

Damit errechnet sich die Varianz des Stichprobenmittelwertes zu

$$\sigma_{\bar{x}}^2 = E((\bar{x} - \mu)^2) = \frac{1}{N^2} \cdot \sum_{n=1}^N E((x_n - \mu)^2) = \frac{N \cdot \sigma^2}{N^2} = \frac{\sigma^2}{N} \quad (5.10)$$

Die Varianz des Stichprobenmittelwertes ergibt sich aus dem 1/N-fachen der Varianz der Grundgesamtheit. Die Streuung des Mittelwertes nimmt demnach mit steigendem Stichprobenumfang N ab. Zur grafischen Darstellung wird in Bild 5.4 eine Stichprobe aus einem Datensatz analysiert, der eine normalverteilte Grundgesamtheit mit einem Mittelwert von $\mu = 0$ und einer Standardabweichung von $\sigma = 0.5$ aufweist.

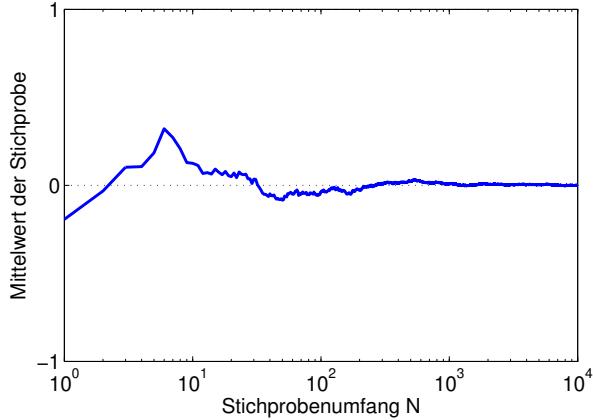


Bild 5.4: Mittelwert einer Stichproben von einem Zufallsprozess mit $\mu = 0$ und $\sigma = 0.5$ als Funktion des Stichprobenumfangs N

Je größer der Stichprobenumfang ist, desto sicherer und genauer ist die Schätzung des Mittelwertes. Da die Schätzung erwartungstreu ist, stimmen der Mittelwert der Grundgesamtheit $\mu = 0$ und der Mittelwert der Stichprobe \bar{x} für große Stichprobenumfänge überein.

5.2.2 Schätzen der Varianz einer Grundgesamtheit

In Abschnitt 5.2.1 wird der Mittelwert einer Grundgesamtheit auf Basis einer Stichprobe geschätzt. Dabei werden die Stichprobenwerte als Zufallsvariablen x_1, \dots, x_N aufgefasst, die den Mittelwert μ und die Varianz σ^2 aufweisen. Im Folgenden wird die Varianz der Stichprobe betrachtet.

Die Varianz einer Stichprobe x_1, x_2, \dots, x_N wird als Schätzwert der Varianz σ^2 der Grundgesamtheit angesehen. Die Schätzung ergibt

$$\sigma^2 \approx s^2 = \frac{1}{N-1} \cdot \sum_{n=1}^N (x_n - \bar{x})^2 = \frac{1}{N-1} \cdot \sum_{n=1}^N (x_n^2 - \bar{x}^2) \quad (5.11)$$

Wieder wird die Erwartungstreue der Schätzung untersucht. Der Erwartungswert der Stichprobenvarianz kann dargestellt werden als

$$E(s^2) = \frac{1}{N-1} \cdot \left(E\left(\sum_{n=1}^N x_n^2\right) - N \cdot E(\bar{x}^2) \right) = \frac{N}{N-1} \cdot (E(x^2) - E(\bar{x}^2)) \quad (5.12)$$

Auch die Varianz σ^2 der Grundgesamtheit kann über den Erwartungswert ausgedrückt werden

$$\sigma^2 = E((x - \nu)^2) = E(x^2 - \nu^2) = E(x^2) - \nu^2 \quad (5.13)$$

Auflösen nach $E(x^2)$ führt zu

$$E(x^2) = \sigma_x^2 + E^2(x) = \sigma^2 + \nu^2 \quad (5.14)$$

Analog ergibt sich für die Varianz des Stichprobenmittelwertes

$$E(\bar{x}^2) = \sigma_{\bar{x}}^2 + E^2(\bar{x}) \quad (5.15)$$

Die Varianz des Stichproben-Mittelwertes wird in Abschnitt 5.2.1 bestimmt zu σ^2/N , der Erwartungswert des Stichprobenwerten-Mittelwertes ist μ . Durch Einsetzen folgt

$$E(\bar{x}^2) = \sigma_{\bar{x}}^2 + E^2(\bar{x}) = \frac{\sigma^2}{N} + \mu^2 \quad (5.16)$$

Mit diesen Nebenrechnungen vereinfacht sich der Erwartungswert der Stichprobenvarianz aus Gleichung (5.12) zu

$$E(s^2) = \frac{N}{N-1} \cdot (E(x^2) - E(\bar{x}^2)) = \frac{N}{N-1} \cdot \left(\sigma^2 + \mu^2 - \frac{\sigma^2}{N} - \mu^2 \right) = \sigma^2 \quad (5.17)$$

Der Erwartungswert der Stichprobenvarianz stimmt also mit der Varianz der Grundgesamtheit überein. Die Schätzung der Varianz σ^2 durch die Stichprobenvarianz s^2 ist damit erwartungstreu.

Zur grafischen Darstellung wird eine Stichprobe aus einem Datensatz analysiert, der eine normalverteilten Grundgesamtheit mit einem Mittelwert von $\mu = 0$ und einer Standardabweichung von $\sigma = 0.5$ aufweist.

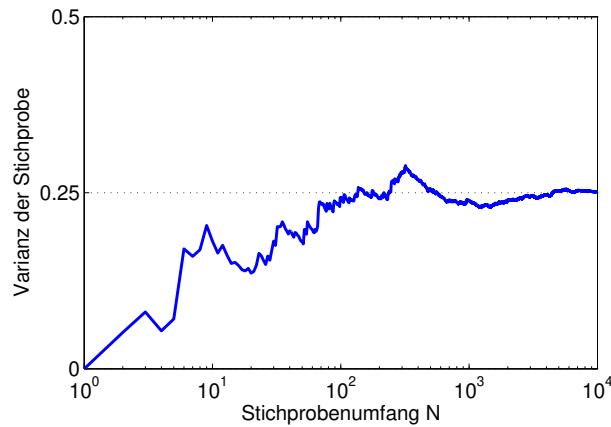


Bild 5.5: Varianz einer Stichprobe von einem Zufallsprozess mit $\mu = 0$ und $\sigma = 0.5$ als Funktion des Stichprobenumfangs N

Je größer der Stichprobenumfang ist, desto näher liegt die Varianz der Stichprobe an der Varianz der Grundgesamtheit $\sigma^2 = 0.25$.

5.3 Konfidenzbereiche für die Schätzung von Parametern

Mithilfe des arithmetischen Mittelwertes und der Varianz einer Stichprobe lassen sich die charakteristischen Parameter μ und σ einer Normalverteilung erwartungstreu schätzen. Sichere Schätzungen von einer Stichprobe auf eine Grundgesamtheit existieren jedoch nicht. Ziel der Wahrscheinlichkeitstheorie ist deshalb, eine Wahrscheinlichkeit dafür anzugeben, mit der zu schätzende Parameter in einem Intervall um ihren Schätzwert liegen. Zur Motivation wird dies am Beispiel eines Mittelwertes verdeutlicht.

Bei der Schätzung des Konfidenzbereiches für den Mittelwert μ der Grundgesamtheit muss ein Weg gefunden werden, die Intervallgrenzen μ_{C1} und μ_{C2} zu bestimmen, die mit einer vorgegebenen Wahrscheinlichkeit γ den wahren Mittelwert μ einschließen. Die Intervallgrenzen μ_{C1} und μ_{C2} müssen sich aus den zufälligen Stichprobenwerten x_1, x_2, \dots, x_N ergeben. Diese Intervallgrenzen μ_{C1} und μ_{C2} sind Funktionen von Zufallszahlen und damit selber Zufallszahlen. Wird die Wahrscheinlichkeit dafür, dass der Mittelwert μ in dem Intervall $\mu_{C1} < \mu \leq \mu_{C2}$ liegt, mit γ bezeichnet, gilt die Gleichung

$$P(\mu_{C1} < \mu \leq \mu_{C2}) = \gamma \quad (5.18)$$

Das Intervall mit den Werten μ_{C1} und μ_{C2} als Intervallgrenzen heißt Konfidenzintervall oder Vertrauensbereich für den unbekannten Parameterwert μ . Die Intervallgrenzen werden auch als Konfidenzgrenzen bezeichnet. Die Wahrscheinlichkeit γ ist die zugehörige Konfidenzzahl. Praktische Werte für γ sind 95 %, 99 % oder 99.9 %. Der Wert γ ist die Wahrscheinlichkeit dafür, dass ein mit einer Stichprobe bestimmtes Konfidenzintervall den wahren unbekannten Parameterwert μ enthält. Wird zum Beispiel eine Konfidenzzahl $\gamma = 95\%$ gewählt, wird davon ausgegangen, dass bei 95 % aller Stichproben die zugehörigen Konfidenzintervalle den Wert μ einschließen.

Für die Berechnung der Konfidenzbereiche wird von einer normalverteilten Grundgesamtheit ausgegangen. Auf Basis dieser Annahme werden die Verteilungen der beiden Stichprobenparameter für Mittelwert und Varianz einer normalverteilten Grundgesamtheit hergeleitet und zur Berechnung des Konfidenzintervalls verwendet.

5.3.1 Konfidenzbereich des Mittelwertes bei bekannter Varianz

Mit den Zufallsvariablen x_1, \dots, x_N werden N Stichprobenwerte bezeichnet, die Teil einer Normalverteilung mit dem unbekannten Mittelwert μ und der bekannten Varianz σ^2 sind. Die einzelnen Stichprobenwerte sind voneinander unabhängig. Der Stichprobenmittelwert ergibt sich aus

$$\bar{x} = \frac{1}{N} \cdot (x_1 + x_2 + \dots + x_N) = \frac{1}{N} \cdot \sum_{n=1}^N x_n \quad (5.19)$$

In Kapitel 8 wird gezeigt, dass die Summe von normalverteilten Werten selbst eine Normalverteilung besitzt. Weiterhin wird gezeigt, dass der Stichprobenmittelwert \bar{x} eine Normalverteilung mit dem Mittelwert

$$E(\bar{x}) = \frac{1}{n} \cdot (E(x_1) + E(x_2) + \dots + E(x_n)) = \frac{n \cdot \mu}{n} = \mu \quad (5.20)$$

und der Varianz

$$E((\bar{x} - \mu)^2) = \frac{\sigma^2}{N} \quad (5.21)$$

besitzt. Bei bekannter Varianz σ^2 der Grundgesamtheit ist auch die Varianz des Mittelwertes σ^2/N bekannt. Die Verteilung der Grundgesamtheit ist somit hinsichtlich aller benötigten Parameter spezifiziert. Auf Basis dieser Verteilung kann die Sicherheit angegeben werden, mit der der Stichprobenmittelwert \bar{x} in definierten Grenzen liegt. Mit der Standardisierung der Zufallsvariable

$$z = \frac{\bar{x} - \mu}{\sigma_{\bar{x}}} = \frac{\bar{x} - \mu}{\sigma/\sqrt{N}} \quad (5.22)$$

geht die Verteilung in eine Standardnormalverteilung über, sie weist also den Mittelwert $\mu_z = 0$ und die Standardabweichung $\sigma_z = 1$ auf. Mit dieser Verteilung wird nach Gleichung (5.18) die Wahrscheinlichkeit γ , mit der die Variable z in dem Intervall $c_1 \dots c_2$ liegt, definiert als

$$P(c_1 < z \leq c_2) = F(c_2) - F(c_1) = \gamma \quad (5.23)$$

Bei Annahme eines symmetrischen Konfidenzbereiches ergeben sich die Konstanten c_1 und c_2 aus den Bedingungen

$$F(c_1) = \frac{1 - \gamma}{2} \quad (5.24)$$

und

$$F(c_2) = 1 - \frac{1 - \gamma}{2} = \frac{1 + \gamma}{2} \quad (5.25)$$

Dabei ist $F(x)$ die Verteilungsfunktion der Standardnormalverteilung. Auflösen nach c_1 und c_2 führt zu

$$c_1 = F^{-1}\left(\frac{1 - \gamma}{2}\right) \quad (5.26)$$

und

$$c_2 = F^{-1}\left(\frac{1 + \gamma}{2}\right) \quad (5.27)$$

Durch Umformungen ergibt sich ein Ausdruck für den Konfidenzbereich des Mittelwertes der Grundgesamtheit.

$$\gamma = P\left(c_1 < \frac{\bar{x} - \mu}{\sigma/\sqrt{N}} \leq c_2\right) = P\left(\frac{c_1 \cdot \sigma}{\sqrt{N}} < \bar{x} - \mu \leq \frac{c_2 \cdot \sigma}{\sqrt{N}}\right) = P\left(\bar{x} - \frac{c_2 \cdot \sigma}{\sqrt{N}} < \mu \leq \bar{x} - \frac{c_1 \cdot \sigma}{\sqrt{N}}\right) \quad (5.28)$$

Mit der gewählten Wahrscheinlichkeit γ liegt der Mittelwert μ der Grundgesamtheit in dem angegebenen Konfidenzintervall. Das Vorgehen zur Bestimmung des Konfidenzintervalls für den Mittelwert einer Normalverteilung mit bekannter Varianz wird in Tabelle 5.2 zusammengefasst.

Tabelle 5.2: Vorgehen zur Bestimmung des Konfidenzintervalls für den Mittelwert einer Normalverteilung mit bekannter Varianz

Nr.	Prozessschritt
1	Wahl einer Konfidenzzahl γ
2	Bestimmung der zugehörigen Parameter c_1 und c_2 aus der inversen Standardnormalverteilung $c_1 = F^{-1}\left(\frac{1-\gamma}{2}\right) \quad \text{und} \quad c_2 = F^{-1}\left(\frac{1+\gamma}{2}\right)$
3	Berechnung des Mittelwertes aus der Stichprobe $\bar{x} = \frac{1}{N} \cdot (x_1 + x_2 + \dots + x_N) = \frac{1}{N} \cdot \sum_{n=1}^N x_n$
4	Bestimmung des Konfidenzintervalls $\bar{x} - \frac{c_2 \cdot \sigma}{\sqrt{N}} < \mu \leq \bar{x} + \frac{c_1 \cdot \sigma}{\sqrt{N}}$

Die Zusammenhänge zwischen Stichprobenumfang, Konfidenzintervall und Aussagesicherheit sollen in den folgenden Abbildungen interpretiert werden. Die Länge L eines Konfidenzintervalls ist nach Gleichung (5.28)

$$L = \frac{(c_2 - c_1) \cdot \sigma}{\sqrt{N}} \quad (5.29)$$

Zur Verkleinerung des Konfidenzintervalls muss der Stichprobenumfang vergrößert werden, oder es müssen Kompromisse hinsichtlich der Aussagesicherheit eingegangen werden. Bild 5.6 stellt für unterschiedliche Konfidenzzahlen γ die Länge des Konfidenzintervalls als Funktion des Stichprobenumfangs N dar.

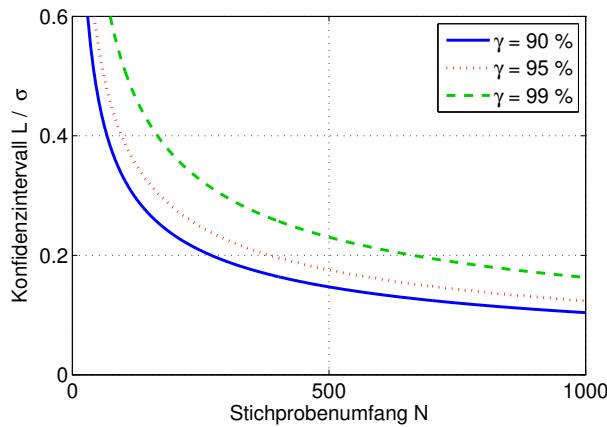


Bild 5.6: Länge des Konfidenzintervalls als Funktion des Stichprobenumfangs N

Durch Auflösen von Gleichung (5.29) nach N kann die Anzahl von Stichproben bestimmt werden, die notwendig ist, den Mittelwert mit einem Konfidenzintervall der Länge L und der Wahrscheinlichkeit γ zu bestimmen.

$$N = \frac{(c_2 - c_1)^2 \cdot \sigma^2}{L^2} \quad (5.30)$$

Bild 5.7 stellt für unterschiedliche Verhältnisse von Länge des Konfidenzbereiches L und Standardabweichung der Grundgesamtheit σ den Stichprobenumfang N als Funktion der Konfidenzzahl γ dar. Der Stichprobenumfang steigt mit steigendem Genauigkeitsanspruch an die Schätzung und mit steigender Sicherheit der Aussage.

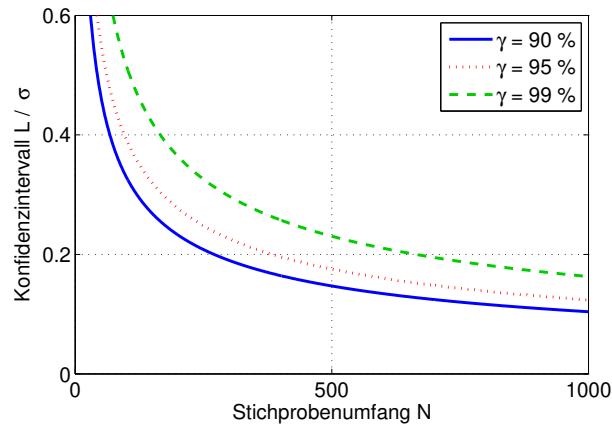


Bild 5.7: Stichprobenumfang n als Funktion der Konfidenzzahl γ

Beispiel: 5 V Linearregler

Bei der Vermessung von 5 V Linearreglern wird eine Stichprobe von $N = 100$ Teilen analysiert. Es soll das 95% Konfidenzintervall für den Mittelwert der entsprechenden Normalverteilung mit der bekannten Varianz $\sigma^2 = 0.09$ werden.

Der Mittelwert der Stichprobe beträgt 5.000 V. Es wird die Konfidenzzahl $\gamma = 0.95$ gefordert, zu der die kritischen Parameter $c_1 = -1.96$ und $c_2 = 1.96$ gehören. Mit dem berechneten Mittelwert ergeben sich die Grenzen des Konfidenzintervalls zu

$$\mu \geq \bar{x} - \frac{c_2 \cdot \sigma}{\sqrt{N}} = 5 - \frac{1.96 \cdot 0.3}{10} = 4.9412 \quad (5.31)$$

und

$$\mu \leq \bar{x} - \frac{c_1 \cdot \sigma}{\sqrt{N}} = 5 - \frac{-1.96 \cdot 0.3}{10} = 5.0588 \quad (5.32)$$

Für das Beispiel soll jetzt der notwendige Stichprobenumfang N für ein 95%-Konfidenzintervall mit der Länge L = 0.01 bestimmt werden. Nach Gleichung (5.30) ergibt sich für den vorliegenden Fall

$$N = \frac{(c_2 - c_1)^2 \cdot \sigma^2}{L^2} = \frac{(1.96 + 1.96)^2 \cdot 0.09}{0.0001} = 13830 \quad (5.33)$$

5.3.2 Konfidenzbereich der Varianz

Ist die Varianz der Grundgesamtheit nicht bekannt, muss die Varianz der Grundgesamtheit mithilfe der Stichprobe geschätzt werden. Allgemein gilt für die Varianz einer Stichprobe

$$s_x^2 = \frac{1}{N-1} \cdot \sum_{n=1}^N (x_n - \bar{x})^2 \quad (5.34)$$

Damit die in Kapitel 4 eingeführten Prüf- und Testverteilungen verwendet werden können, wird die Variable x standardisiert. Dazu wird die Variable z eingeführt.

$$z_n = \frac{x_n - \mu}{\sigma} \quad (5.35)$$

Sie besitzt den Mittelwert

$$\bar{z} = \frac{1}{N} \cdot \sum_{n=1}^N \frac{x_n - \mu}{\sigma} = \frac{\bar{x} - \mu}{\sigma} \quad (5.36)$$

Für die Standardabweichungen der Zufallsvariable z ergibt sich

$$s_z^2 = \frac{1}{N-1} \cdot \sum_{n=1}^N (z_n - \bar{z})^2 = \frac{1}{N-1} \cdot \sum_{n=1}^N \left(\frac{x_n - \mu}{\sigma} - \frac{\bar{x} - \mu}{\sigma} \right)^2 = \frac{1}{N-1} \cdot \sum_{n=1}^N \left(\frac{x_n - \bar{x}}{\sigma} \right)^2 \quad (5.37)$$

Bei dem Ausdruck für s_{Z2} handelt es sich um die Quadratsumme einer standardnormalverteilten Größe. Nach den Ausführungen zur Chi-Quadrat-Verteilung und Gleichung (5.37) ist damit die Größe

$$\chi = (N-1) \cdot s_z^2 = \sum_{n=1}^N \left(\frac{x_n - \bar{x}}{\sigma} \right)^2 \quad (5.38)$$

chi-quadrat-verteilt mit $N - 1$ Freiheitsgraden. Um die Varianz der Stichprobe s_{x2} und die Varianz der Grundgesamtheit σ^2 in Verbindung zu bringen, wird Gleichung (5.38) umgeformt zu

$$\chi = \sum_{n=1}^N \left(\frac{x_n - \bar{x}}{\sigma} \right)^2 = \frac{N-1}{\sigma^2} \cdot \frac{1}{N-1} \cdot \sum_{n=1}^N (x_n - \bar{x})^2 = \frac{N-1}{\sigma^2} \cdot s_x^2 = \frac{s_x^2}{\sigma^2} \cdot (N-1) \quad (5.39)$$

Das Verhältnis der beiden Varianzen kann demnach über die Chi-Quadrat-Verteilung beschrieben werden. Zur Herleitung des Konfidenzintervalls für σ^2 wird über die Wahrscheinlichkeit dafür berechnet, dass die Variable in einem Intervall $c_1 < \chi \leq c_2$ liegt.

$$P(c_1 \leq \chi \leq c_2) = \gamma \quad (5.40)$$

Bei Annahme eines symmetrischen Konfidenzbereiches ergeben sich die Konstanten c_1 und c_2 aus den Bedingungen

$$F(c_1) = \frac{1-\gamma}{2} \quad (5.41)$$

und

$$F(c_2) = 1 - \frac{1-\gamma}{2} = \frac{1+\gamma}{2} \quad (5.42)$$

Dabei ist $F(x)$ die Verteilungsfunktion der Chi-Quadrat-Verteilung mit $N - 1$ Freiheitsgraden. Auflösen nach c_1 und c_2 führt zu

$$c_1 = F^{-1} \left(\frac{1-\gamma}{2} \right) \quad (5.43)$$

und

$$c_2 = F^{-1}\left(\frac{1+\gamma}{2}\right) \quad (5.44)$$

Durch Umformungen ergibt sich der Konfidenzbereich für die Varianz der Grundgesamtheit.

$$\gamma = P\left(c_1 < \frac{s^2}{\sigma^2} \cdot (N-1) \leq c_2\right) = P\left(\frac{1}{c_1} \geq \frac{\sigma^2}{s^2 \cdot (N-1)} > \frac{1}{c_2}\right) = P\left(\frac{s^2 \cdot (N-1)}{c_2} < \sigma^2 \leq \frac{s^2 \cdot (N-1)}{c_1}\right) \quad (5.45)$$

Mit der gewählten Wahrscheinlichkeit γ liegt die Varianz σ^2 der Grundgesamtheit in dem angegebenen Konfidenzintervall. Das Vorgehen zur Bestimmung des Konfidenzintervalls für die Varianz einer Normalverteilung wird in Tabelle 5.3 zusammengefasst.

Tabelle 5.3: Vorgehen zur Bestimmung des Konfidenzintervalls für die Varianz einer Normalverteilung

Nr.	Prozessschritt
1	Wahl einer Konfidenzzahl γ
2	Bestimmung der zugehörigen Parameter c_1 und c_2 aus der inversen Chi-Quadrat-Verteilung mit $N - 1$ Freiheitsgraden $c_1 = F^{-1}\left(\frac{1-\gamma}{2}\right)$ und $c_2 = F^{-1}\left(\frac{1+\gamma}{2}\right)$
3	Berechnung der Varianz aus der Stichprobe $s^2 = \frac{1}{N-1} \cdot \sum_{n=1}^N (x_n - \bar{x})^2$
4	Bestimmung des Konfidenzintervalls $\frac{s^2 \cdot (N-1)}{c_2} < \sigma^2 \leq \frac{s^2 \cdot (N-1)}{c_1}$

Beispiel: Gewicht von Klebermengen

In einem Beispiel werden die Varianz für das Gewicht von Klebermengen eines Fertigungsprozesses und ihr Konfidenzbereich bestimmt. Es soll das 95%-Konfidenzintervall für die Varianz σ^2 der Grundgesamtheit für die folgende Stichprobe berechnet werden. Dabei wird vorausgesetzt, dass die Grundgesamtheit normalverteilt ist.

Tabelle 5.4: Stichprobe für das Gewicht von Klebermengen in einem Fertigungsprozess

Stichprobe	1	2	3	4	5	6	7	8	9	10
Gewicht x / g	4.3	4.5	4.2	4.3	4.3	4.7	4.4	4.2	4.3	4.5

Es ist eine Konfidenzzahl von $\gamma = 0.95$ gefordert. Wegen $N = 10$ ergeben sich aus der inversen Chi-Quadrat-Verteilung mit $10 - 1 = 9$ Freiheitsgraden $c_1 = 2.7$ und $c_2 = 19.02$. Die Auswertung der Stichprobe ergibt eine Varianz von

$$s^2 = \frac{1}{N-1} \cdot \sum_{n=1}^N (x_n - \bar{x})^2 = 0.0246g^2 \quad (5.46)$$

Damit lautet das Konfidenzintervall

$$0.0116g^2 < \sigma^2 \leq 0.0818g^2 \quad (5.47)$$

5.3.3 Konfidenzbereich des Mittelwertes bei unbekannter Varianz

Ist die Varianz der Grundgesamtheit nicht bekannt, kann die in Abschnitt 5.3.1 hergeleitete Standardnormalverteilung nicht verwendet werden. Zur Herleitung der zugrunde liegenden Verteilung wird zunächst wieder die Größe

$$z = \frac{\bar{x} - \mu}{\sigma / \sqrt{N}} \quad (5.48)$$

betrachtet. Sie weist nach den Ausführungen in Abschnitt 5.3.1 eine Standardnormalverteilung auf. Die Herleitung in Gleichung (5.49) zeigt, dass die Größe

$$\frac{\chi}{N-1} = \frac{s^2}{\sigma^2} \quad (5.49)$$

eine Chi-Quadrat-Verteilung mit $N - 1$ Freiheitsgraden besitzt. Nach den Ausführungen zu Testverteilungen besitzt die Zufallsvariable

$$t = \frac{\frac{\bar{x} - \mu}{\sigma / \sqrt{N}}}{\sqrt{\frac{s^2}{\sigma^2}}} = \frac{\bar{x} - \mu}{s / \sqrt{N}} \quad (5.50)$$

damit eine t-Verteilung mit $N - 1$ Freiheitsgraden. Damit ist auch bei unbekannter Varianz der Grundgesamtheit die Verteilung für den Stichprobenmittelwert \bar{x} bestimmt, und es kann die Sicherheit angegeben werden, mit der der Stichprobenmittelwert \bar{x} in definierten Grenzen liegt. Dazu wird eine Konfidenzzahl γ festgelegt, für die gilt

$$P(c_1 < t \leq c_2) = F(c_2) - F(c_1) = \gamma \quad (5.51)$$

Bei Annahme eines symmetrischen Konfidenzbereiches ergeben sich die Konstanten c_1 und c_2 aus den Bedingungen

$$F(c_1) = \frac{1 - \gamma}{2} \quad (5.52)$$

und

$$F(c_2) = 1 - \frac{1 - \gamma}{2} = \frac{1 + \gamma}{2} \quad (5.53)$$

Dabei ist $F(x)$ die Verteilungsfunktion der t-Verteilung mit $N - 1$ Freiheitsgraden. Auflösen nach c_1 und c_2 führt zu

$$c_1 = F^{-1}\left(\frac{1 - \gamma}{2}\right) \quad (5.54)$$

und

$$c_2 = F^{-1}\left(\frac{1 + \gamma}{2}\right) \quad (5.55)$$

Durch Umformungen ergibt sich ein Ausdruck für den Konfidenzbereich des Mittelwertes der Grundgesamtheit bei unbekannter Varianz.

$$\gamma = P\left(c_1 < \sqrt{N} \cdot \frac{\bar{x} - \mu}{s} \leq c_2\right) = P\left(\frac{c_1 \cdot s}{\sqrt{N}} < \bar{x} - \mu \leq \frac{c_2 \cdot s}{\sqrt{N}}\right) = P\left(\bar{x} - \frac{c_2 \cdot s}{\sqrt{N}} < \mu \leq \bar{x} - \frac{c_1 \cdot s}{\sqrt{N}}\right) \quad (5.56)$$

Mit der gewählten Wahrscheinlichkeit γ liegt der Mittelwert μ der Grundgesamtheit in dem angegebenen Konfidenzintervall. Das Vorgehen zur Bestimmung des Konfidenzintervalls für den Mittelwert einer Normalverteilung mit unbekannter Varianz wird in Tabelle 5.5 zusammengefasst.

Tabelle 5.5: Vorgehen zur Bestimmung des Konfidenzintervalls für den Mittelwert einer Normalverteilung mit unbekannter Varianz

Nr.	Prozessschritt
1	Wahl einer Konfidenzzahl γ
2	Bestimmung der zugehörigen Parameter c_1 und c_2 aus der inversen Chi-Quadrat-Verteilung mit $N - 1$ Freiheitsgraden $c_1 = F^{-1}\left(\frac{1-\gamma}{2}\right) \quad \text{und} \quad c_2 = F^{-1}\left(\frac{1+\gamma}{2}\right)$
3	Berechnung der Varianz aus der Stichprobe $\bar{x} = \frac{1}{N} \cdot \sum_{n=1}^N x_n \quad \text{und} \quad s = \sqrt{\frac{1}{N-1} \cdot \sum_{n=1}^N (x_n - \bar{x})^2}$
4	Bestimmung des Konfidenzintervalls $\bar{x} - \frac{c_2 \cdot s}{\sqrt{N}} < \mu \leq \bar{x} + \frac{c_1 \cdot s}{\sqrt{N}}$

Um den Unterschied zwischen den beiden Fällen einer bekannten Varianz und einer unbekannten Varianz zu diskutieren, zeigt Bild 5.8 das Verhältnis der Länge L_T des Konfidenzintervalls bei unbekannter Varianz und der Länge L_Z des Konfidenzintervalls bei bekannter Varianz als Funktion des Stichprobenumfangs N .

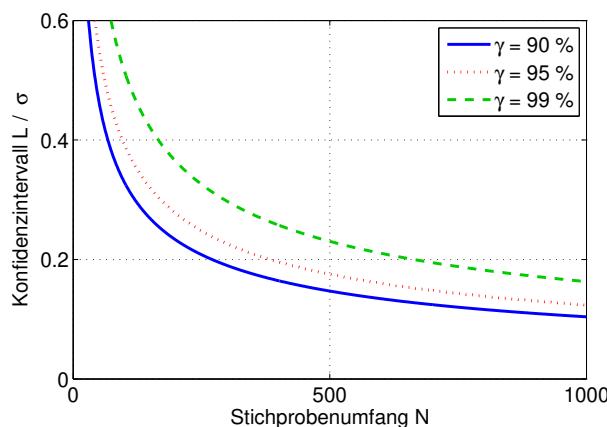


Bild 5.8: Verhältnis der Länge L_T des Konfidenzintervalls bei unbekannter Varianz und der Länge L_Z des Konfidenzintervalls bei bekannter Varianz als Funktion des Stichprobenumfangs N

Bei einer hinreichend großen Stichprobe unterscheiden sich die Längen der beiden Konfidenzintervalle nur wenig. Je kleiner der Stichprobenumfang N wird, desto größer ist der Unterschied. Zusätzlich steigt der Unterschied mit steigender Konfidenzzahl γ an.

Beispiel: Gewicht von Klebermengen

In einem Beispiel wird der Mittelwert und Konvergenzbereich für das Gewicht von Klebermengen eines Fertigungsprozesses bestimmt. Die Daten sind in Tabelle 5.5 aufgeführt. Es soll das 95%-Konfidenzintervall für den Mittelwert μ der Grundgesamtheit berechnet werden. Dabei wird vorausgesetzt, dass die Grundgesamtheit normalverteilt ist.

Es ist eine Konfidenzzahl von $\gamma = 0.95$ gefordert. Mit $N = 10$ Stichproben ergeben sich die Parameter c_1

= - 2.2622 und $c_2 = 2.2622$. Die Auswertung der Stichprobe ergibt für den arithmetischen Mittelwert

$$\bar{x} = \frac{1}{N} \cdot \sum_{n=1}^N x_n = 4.37g \quad (5.57)$$

und für die Standardabweichung

$$s = \sqrt{\frac{1}{N-1} \cdot \sum_{n=1}^N (x_n - \bar{x})^2} = 0.157g \quad (5.58)$$

Mit diesen Werten ergibt sich das Konfidenzintervall zu

$$KONF(4.258g \leq \mu \leq 4.482g) \quad (5.59)$$

5.3.4 Zusammenfassung der Konfidenzbereiche für die Schätzung von Parametern

Die Schätzung des Konfidenzbereiches beruht in allen Fällen auf einer Zufallsvariable mit einer bekannten Verteilung, in deren Beschreibung der gesuchte Parameter der Grundgesamtheit und der bekannte Parameter der Stichprobe vorkommen. Für die bekannte Zufallsvariable wird der Konfidenzbereich für ein Signifikanzniveau γ bestimmt. Durch Umformen dieser Gleichung ergibt sich der Konfidenzbereich der gesuchten Größe.

Zur Abschätzung der Aussagesicherheit von Mittelwert μ und Standardabweichung σ einer normalverteilten Grundgesamtheit ergeben sich die in Tabelle 5.6 dargestellten Verteilungen und Konfidenzbereiche.

Tabelle 5.6: Verteilungen und Konfidenzbereiche von Stichproben normalverteilter Grundgesamtheiten

	Verteilung der Schätzfunktion	Konfidenzbereich
Mittelwert μ bei bekannter Varianz	Standardnormalverteilung $z = \frac{\bar{x} - \mu}{\sigma / \sqrt{N}}$	$\bar{x} - \frac{c_2 \cdot \sigma}{\sqrt{N}} < \mu \leq \bar{x} + \frac{c_1 \cdot \sigma}{\sqrt{N}}$
Varianz σ^2	Chi-Quadrat-Verteilung mit $N - 1$ Freiheitsgraden $z = \frac{\bar{x} - \mu}{\sigma / \sqrt{N}}$	$\bar{x} - \frac{c_2 \cdot \sigma}{\sqrt{N}} < \mu \leq \bar{x} + \frac{c_1 \cdot \sigma}{\sqrt{N}}$
Mittelwert μ bei unbekannter Varianz	t-Verteilung mit $N - 1$ Freiheitsgraden $\bar{x} - \frac{c_2 \cdot s}{\sqrt{N}} < \mu \leq \bar{x} + \frac{c_1 \cdot s}{\sqrt{N}}$	$t = \frac{\bar{x} - \mu}{s / \sqrt{N}}$

5.4 Konfidenzbereiche für den Vergleich von Stichproben

In Abschnitt 5.3 werden Konfidenzbereiche für Parameter von Zufallsvariablen berechnet. Bei der Auswertung von Labor-Versuchen tritt außerdem der Fall auf, dass zwei Stichproben $x_{11}, x_{12}, \dots, x_{1N}$ und $x_{21}, x_{22}, \dots, x_{2M}$ miteinander verglichen werden sollen. In diesem Abschnitt wird gezeigt, welche Schlüsse dabei auf die Grundgesamtheiten möglich sind. Dabei wird wieder von einer normalverteilten Grundgesamtheit ausgegangen.

5.4.1 Konfidenzbereich der Differenz der Mittelwerte bei bekannter Varianz

Die Messwerte der beiden Stichproben entsprechen unabhängigen normalverteilten Zufallsvariablen mit dem arithmetischen Mittelwert

$$\bar{x}_1 = \frac{1}{N} \cdot \sum_{n=1}^N x_{1n} \quad (5.60)$$

beziehungsweise

$$\bar{x}_2 = \frac{1}{M} \cdot \sum_{m=1}^M x_{2m} \quad (5.61)$$

und der bekannten Varianz σ^2 . Nach den Rechenregeln für mehrere Zufallsvariablen besitzt die Differenz der Stichprobenmittelwerte

$$\bar{x} = \bar{x}_1 - \bar{x}_2 \quad (5.62)$$

den Erwartungswert

$$\mu = \mu_1 - \mu_2 \quad (5.63)$$

und die Varianz

$$\sigma_x^2 = \sigma_{\bar{x}_1}^2 + \sigma_{\bar{x}_2}^2 = \frac{\sigma^2}{N} + \frac{\sigma^2}{M} = \sigma^2 \cdot \left(\frac{1}{N} + \frac{1}{M} \right) \quad (5.64)$$

Mit der Standardisierung der Zufallsvariablen

$$z = \frac{\bar{x} - \mu}{\sqrt{\sigma_x^2}} = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\sigma^2 \cdot \left(\frac{1}{N} + \frac{1}{M} \right)}} \quad (5.65)$$

geht die Verteilung in eine Standardnormalverteilung über. Sie weist den Mittelwert $\mu_z = 0$ und die Standardabweichung $\sigma_z = 1$ auf. Die Wahrscheinlichkeit γ , mit der die Variable z in dem Intervall $c_1 \dots c_2$ liegt, ergibt sich zu

$$P(c_1 < z \leq c_2) = F(c_2) - F(c_1) = \gamma \quad (5.66)$$

Bei Annahme eines symmetrischen Konfidenzbereiches ergeben sich die Konstanten c_1 und c_2 aus den Bedingungen

$$F(c_1) = \frac{1 - \gamma}{2} \quad (5.67)$$

und

$$F(c_2) = 1 - \frac{1 - \gamma}{2} = \frac{1 + \gamma}{2} \quad (5.68)$$

Dabei ist $F(x)$ die Verteilungsfunktion der Standardnormalverteilung. Auflösen nach c_1 und c_2 führt zu

$$c_1 = F^{-1}\left(\frac{1-\gamma}{2}\right) \quad (5.69)$$

und

$$c_2 = F^{-1}\left(\frac{1+\gamma}{2}\right) \quad (5.70)$$

Durch Umformungen ergibt sich ein Ausdruck für den Konfidenzbereich der Differenz zweier Mittelwerte.

$$\begin{aligned} \gamma &= P\left(c_1 < \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\sigma^2 \cdot \left(\frac{1}{N} + \frac{1}{M}\right)}} \leq c_2\right) \\ &= P\left((\bar{x}_1 - \bar{x}_2) - c_2 \cdot \sqrt{\sigma^2 \cdot \left(\frac{1}{N} + \frac{1}{M}\right)} < (\mu_1 - \mu_2) \leq (\bar{x}_1 - \bar{x}_2) - c_1 \cdot \sqrt{\sigma^2 \cdot \left(\frac{1}{N} + \frac{1}{M}\right)}\right) \end{aligned} \quad (5.71)$$

Mit der gewählten Wahrscheinlichkeit γ liegt die Differenz zweier Mittelwerte $\mu_1 - \mu_2$ in dem angegebenen Konfidenzintervall. Das Vorgehen zur Bestimmung des Konfidenzintervalls für die Differenz zweier Mittelwerte bei bekannter Varianz der Grundgesamtheit wird in Tabelle 5.7 zusammengefasst.

Tabelle 5.7: Vorgehen zur Bestimmung des Konfidenzintervalls für die Differenz zweier Mittelwerte einer Normalverteilung mit bekannter Varianz

Nr.	Prozessschritt
1	Wahl einer Konfidenzzahl γ
2	Bestimmung der zugehörigen Parameter c_1 und c_2 aus der inversen Chi-Quadrat-Verteilung mit $N - 1$ Freiheitsgraden $c_1 = F^{-1}\left(\frac{1-\gamma}{2}\right)$ und $c_2 = F^{-1}\left(\frac{1+\gamma}{2}\right)$
3	Berechnung der Mittelwerte aus den Stichproben $\bar{x}_1 = \frac{1}{N} \cdot \sum_{n=1}^N x_{1n}$ und $\bar{x}_2 = \frac{1}{M} \cdot \sum_{m=1}^M x_{2m}$
4	Bestimmung des Konfidenzintervalls $(\bar{x}_1 - \bar{x}_2) - c_2 \cdot \sqrt{\sigma^2 \cdot \left(\frac{1}{N} + \frac{1}{M}\right)} < (\mu_1 - \mu_2) \leq (\bar{x}_1 - \bar{x}_2) - c_1 \cdot \sqrt{\sigma^2 \cdot \left(\frac{1}{N} + \frac{1}{M}\right)}$

Beispiel: Fertigung von Passstiften

Die Fertigung von Passstiften soll mithilfe des 95%-Konfidenzbereichs der Differenz zweier Mittelwerte untersucht werden. Von der Fertigung ist bekannt, dass sie eine Varianz von $\sigma^2 = 0.1 \text{ mm}$ besitzt. Zu diesem Zweck werden der Fertigung mit einer Woche Abstand je eine Stichprobe mit einem Umfang von $N = M = 15$ Passstiften entnommen.

Die Auswertung der ersten Stichprobe ergibt einen arithmetischen Mittelwert von

$$\bar{x}_1 = 50.1 \text{ mm} \quad (5.72)$$

die zweite Stichprobe besitzt einen Mittelwert von

$$\bar{x}_2 = 49.5 \text{ mm} \quad (5.73)$$

Es ist eine Konfidenzzahl von $\gamma = 0.95$ gefordert. Daraus ergeben sich die Parameter $c_1 = -1.96$ und $c_2 = 1.96$ aus der inversen Standardnormalverteilung. Damit ergibt sich der 95%-Konfidenzbereich zu

$$0.3737 < (\mu_1 - \mu_2) \leq 0.8263 \quad (5.74)$$

5.4.2 Konfidenzbereich der Differenz der Mittelwerte bei unbekannter Varianz

Sind die Varianzen der beiden Versuchsergebnisse nicht bekannt, muss die Berechnung des Mittelwertes auf die t-Verteilung zurückgeführt werden. Dies ist allerdings nur dann möglich, wenn die beiden Grundgesamtheiten dieselbe unbekannte Varianz σ^2 aufweisen. Es wird daher davon ausgegangen, dass die Stichprobe 1 einer normalverteilten Grundgesamtheit mit Mittelwert μ_1 und einer unbekannten Varianz σ^2 und die Stichprobe 2 einer normalverteilten Grundgesamtheit mit Mittelwert μ_2 und derselben Varianz σ^2 entstammt. Die Stichprobenmittelwerte ergeben sich zu

$$\bar{x}_1 = \frac{1}{N} \cdot \sum_{n=1}^N x_{1n} \quad (5.75)$$

beziehungsweise

$$\bar{x}_2 = \frac{1}{M} \cdot \sum_{m=1}^M x_{2m} \quad (5.76)$$

und die Varianz der Stichprobe folgt zu

$$s_1^2 = \frac{1}{N-1} \cdot \sum_{n=1}^N (x_{1n} - \bar{x}_1)^2 \quad (5.77)$$

beziehungsweise

$$s_2^2 = \frac{1}{M-1} \cdot \sum_{m=1}^M (x_{2m} - \bar{x}_2)^2 \quad (5.78)$$

Um den Mittelwert der Differenz der beiden Mittelwerte zu berechnen, wird die Zufallsvariable

$$\bar{x} = \bar{x}_1 - \bar{x}_2 \quad (5.79)$$

eingeführt. Die Differenz der Stichproben-Mittelwerte besitzt den Erwartungswert

$$\mu = \mu_1 - \mu_2 \quad (5.80)$$

und die Varianz

$$\sigma_{\bar{X}}^2 = \sigma_{\bar{X}_1}^2 + \sigma_{\bar{X}_2}^2 = \frac{\sigma^2}{N} + \frac{\sigma^2}{M} = \sigma^2 \cdot \left(\frac{1}{N} + \frac{1}{M} \right) \quad (5.81)$$

Damit besitzt die Zufallsvariable

$$z = \frac{\bar{x} - \mu}{\sqrt{\sigma_{\bar{X}}^2}} = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\sigma^2 \cdot \left(\frac{1}{N} + \frac{1}{M} \right)}} \quad (5.82)$$

eine Standardnormalverteilung. Da die Varianz σ^2 unbekannt ist, wird sie über die Stichproben geschätzt. Für die erste Stichprobe ergibt sich die chi-quadrat-verteilte Größe

$$\chi_1 = \frac{s_1^2}{\sigma^2} \cdot (N - 1) \quad (5.83)$$

und für die zweite Stichprobe

$$\chi_2 = \frac{s_2^2}{\sigma^2} \cdot (M - 1) \quad (5.84)$$

In [Ross06] wird gezeigt, dass die Summe zweier chi-quadrat-verteilter Stichproben mit $N - 1$ und $M - 1$ Freiheitsgraden ebenfalls eine chi-quadrat-verteilte Größe ist und $N + M - 2$ Freiheitsgrade besitzt.

$$\chi = \chi_1 + \chi_2 = \frac{s_1^2}{\sigma^2} \cdot (N - 1) + \frac{s_2^2}{\sigma^2} \cdot (M - 1) \quad (5.85)$$

Die Varianzen können zusammengefasst werden zu

$$s^2 = \frac{s_1^2 \cdot (N - 1) + s_2^2 \cdot (M - 1)}{N + M - 2} \quad (5.86)$$

Mit der Zufallsvariablen z aus Gleichung (5.82) und der Zufallsvariablen χ aus Gleichung (5.85) kann die Zufallsvariable t einer t-Verteilung gebildet werden.

$$t = \frac{z}{\sqrt{\frac{\chi}{N + M - 2}}} = \frac{\frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\sigma^2 \cdot \left(\frac{1}{N} + \frac{1}{M} \right)}}}{\sqrt{\frac{(N + M - 2) \cdot s^2}{N + M - 2}}} = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{1}{N} + \frac{1}{M}} \cdot s} \quad (5.87)$$

Die Zufallsvariable t aus Gleichung (5.87) besitzt eine t-Verteilung mit $N + M - 2$ Freiheitsgraden. Damit wird die Wahrscheinlichkeit γ , mit der die Variable t in dem Intervall $c_1 \dots c_2$ liegt, definiert werden als

$$P(c_1 < t \leq c_2) = F(c_2) - F(c_1) = \gamma \quad (5.88)$$

Bei Annahme eines symmetrischen Konfidenzbereiches ergeben sich die Konstanten c_1 und c_2 aus den Bedingungen

$$F(c_1) = \frac{1 - \gamma}{2} \quad (5.89)$$

und

$$F(c_2) = 1 - \frac{1 - \gamma}{2} = \frac{1 + \gamma}{2} \quad (5.90)$$

Dabei ist $F(x)$ die Verteilungsfunktion der t-Verteilung mit $N + M - 2$ Freiheitsgraden. Auflösen nach c_1 und c_2 führt zu

$$c_1 = F^{-1}\left(\frac{1-\gamma}{2}\right) \quad (5.91)$$

und

$$c_2 = F^{-1}\left(\frac{1+\gamma}{2}\right) \quad (5.92)$$

Durch Umformungen ergibt sich ein Ausdruck für den Konfidenzbereich der Differenz zweier Mittelwerte bei unbekannter Varianz.

$$\begin{aligned} \gamma &= P\left(c_1 < \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{1}{N} + \frac{1}{M}} \cdot s} \leq c_2\right) \\ &= P\left((\bar{x}_1 - \bar{x}_2) - c_2 \cdot \sqrt{\frac{1}{N} + \frac{1}{M}} \cdot s < (\mu_1 - \mu_2) \leq (\bar{x}_1 - \bar{x}_2) - c_1 \cdot \sqrt{\frac{1}{N} + \frac{1}{M}} \cdot s\right) \end{aligned} \quad (5.93)$$

Mit der gewählten Wahrscheinlichkeit γ liegt die Differenz zweier Mittelwerte $\mu_1 - \mu_2$ in dem angegebenen Konfidenzintervall. Das Vorgehen zur Bestimmung des Konfidenzintervalls für die Differenz zweier Mittelwerte bei unbekannter Varianz wird in Tabelle 5.8 zusammengefasst.

Tabelle 5.8: Vorgehen zur Bestimmung des Konfidenzintervalls für die Differenz zweier Mittelwerte einer normalverteilten Größe bei unbekannter Varianz

Nr.	Prozessschritt
1	Wahl einer Konfidenzzahl γ
2	Bestimmung der zugehörigen Parameter c_1 und c_2 aus der inversen t-Verteilung mit $N + M - 2$ Freiheitsgraden $c_1 = F^{-1}\left(\frac{1 - \gamma}{2}\right) \quad \text{und} \quad c_2 = F^{-1}\left(\frac{1 + \gamma}{2}\right)$
	Berechnung der Mittelwerte aus den Stichproben $\bar{x}_1 = \frac{1}{N} \cdot \sum_{n=1}^N x_{1n} \quad \text{und} \quad \bar{x}_2 = \frac{1}{M} \cdot \sum_{m=1}^M x_{2m}$
3	Berechnung der Varianzen aus den Stichproben $s_1^2 = \frac{1}{N-1} \cdot \sum_{n=1}^N (x_{1n} - \bar{x}_1)^2 \quad \text{und} \quad s_2^2 = \frac{1}{M-1} \cdot \sum_{m=1}^M (x_{2m} - \bar{x}_2)^2$
	Berechnung der Varianz s $s^2 = \frac{(N-1) \cdot s_1^2 + (M-1) \cdot s_2^2}{N+M-2}$
4	Bestimmung des Konfidenzintervalls $(\bar{x}_1 - \bar{x}_2) - c_2 \cdot \sqrt{\frac{1}{N} + \frac{1}{M}} \cdot s < (\mu_1 - \mu_2) \leq (\bar{x}_1 - \bar{x}_2) - c_1 \cdot \sqrt{\frac{1}{N} + \frac{1}{M}} \cdot s$

Beispiel: Speicherkapazität von Batterien

Als Beispiel wird die Speicherkapazität von Batterien untersucht, die nach zwei unterschiedlichen Fertigungsverfahren produziert wurden. Zur Untersuchung stehen zwei Stichproben zur Verfügung, die in Tabelle 5.9 aufgelistet sind. Es soll der 95%-Konfidenzbereich für die Differenz der beiden Mittelwerte bestimmt werden.

Tabelle 5.9: Messung der Speicherkapazität für Batterien unterschiedlicher Fertigungsverfahren

Index	Verfahren 1: Kapazität / mAh				
1 - 5	140	132	136	142	138
6 - 10	150	150	154	152	136
11- 12	144	142			
Index	Verfahren 2: Kapazität / mAh				
1 - 5	144	134	132	130	136
6 - 10	146	140	128	128	131
11- 14	150	137	130	135	

Es ist eine Konfidenzzahl von $\gamma = 0.95$ gefordert. Daraus ergeben sich die Parameter $c_1 = -2.0639$ und $c_2 = 2.0639$ aus der inversen t-Verteilung mit $N + M - 2 = 24$ Freiheitsgraden. Die Auswertung der

Stichproben führt zu den Stichprobenmittelwerten

$$\bar{x}_1 = 143 \text{ mAh} \quad (5.94)$$

$$\bar{x}_2 = 135.8 \text{ mAh} \quad (5.95)$$

und den Stichprobenvarianzen

$$s_1^2 = 50.55 \text{ mAh}^2 \quad (5.96)$$

$$s_2^2 = 47.8736 \text{ mAh}^2 \quad (5.97)$$

Daraus errechnet sich ein Konfidenzbereich

$$1.5106 < (\mu_1 - \mu_2) \leq 12.8894 \quad (5.98)$$

5.4.3 Konfidenzbereich des Verhältnisses der Varianzen

Ziel von Design For Six Sigma sind Prozesse mit geringer Varianz. Eine kleine Varianz der wesentlichen Spezifikationsmerkmale weist auf beherrschte Fertigungsprozesse hin. Ein Vergleich von aktuellen Fertigungsstücken mit älteren Fertigungsstücken wird daher in der Qualitätssicherung im Rahmen der statistischen Prozesskontrolle dazu verwendet, um Veränderungen im Fertigungsprozess zu erkennen. Die Messwerte der Stichproben können wiederum als unabhängige normalverteilte Zufallsvariablen mit den arithmetischen Mittelwerten

$$\bar{x}_1 = \frac{1}{N} \cdot \sum_{n=1}^N x_{1n} \quad (5.99)$$

$$\bar{x}_2 = \frac{1}{M} \cdot \sum_{m=1}^M x_{2m} \quad (5.100)$$

und den Varianzen

$$s_1^2 = \frac{1}{N-1} \cdot \sum_{n=1}^N (x_{1n} - \bar{x}_1)^2 \quad (5.101)$$

$$s_2^2 = \frac{1}{M-1} \cdot \sum_{m=1}^M (x_{2m} - \bar{x}_2)^2 \quad (5.102)$$

betrachtet werden. Das Verhältnis der beiden Stichprobenvarianzen

$$f = \frac{\frac{(N-1) \cdot s_1^2}{\sigma_1^2} \cdot \frac{1}{(N-1)}}{\frac{(M-1) \cdot s_2^2}{\sigma_2^2} \cdot \frac{1}{(M-1)}} = \frac{s_1^2}{s_2^2} \cdot \frac{\sigma_2^2}{\sigma_1^2} \quad (5.103)$$

besitzt nach Gleichung (4.237) eine f-Verteilung mit (N - 1, M - 1) Freiheitsgraden. Mit der Zufallsvariablen f aus Gleichung (5.103) wird die Wahrscheinlichkeit γ , mit der die Variable f in dem Intervall $c_1 \dots c_2$ liegt, definiert werden als

$$P(c_1 < f \leq c_2) = F(c_2) - F(c_1) = \gamma \quad (5.104)$$

Bei Annahme eines symmetrischen Konfidenzbereiches ergeben sich die Konstanten c_1 und c_2 aus den Bedingungen

$$F(c_1) = \frac{1 - \gamma}{2} \quad (5.105)$$

und

$$F(c_2) = 1 - \frac{1 - \gamma}{2} = \frac{1 + \gamma}{2} \quad (5.106)$$

Dabei ist $F(x)$ die Verteilungsfunktion einer f-Verteilung mit $(N - 1, M - 1)$ Freiheitsgraden. Auflösen nach c_1 und c_2 führt zu

$$c_1 = F^{-1}\left(\frac{1 - \gamma}{2}\right) \quad (5.107)$$

und

$$c_2 = F^{-1}\left(\frac{1 + \gamma}{2}\right) \quad (5.108)$$

Durch Umformungen ergibt sich ein Ausdruck für den Konfidenzbereich der Differenz zweier Mittelwerte bei unbekannter Varianz.

$$\gamma = P\left(c_1 < \frac{s_1^2}{s_2^2} \cdot \frac{\sigma_2^2}{\sigma_1^2} \leq c_2\right) = P\left(\frac{s_2^2}{s_1^2} \cdot c_1 < \frac{\sigma_2^2}{\sigma_1^2} \leq \frac{s_2^2}{s_1^2} \cdot c_2\right) \quad (5.109)$$

Mit der gewählten Wahrscheinlichkeit γ liegt das Verhältnis zweier Stichprobenvarianzen in dem angegebenen Konfidenzintervall. Das Vorgehen zur Bestimmung des Konfidenzintervalls für das Verhältnis zweier Stichprobenvarianzen wird in Tabelle 5.10 zusammengefasst.

Tabelle 5.10: Vorgehen zur Bestimmung des Konfidenzintervalls für das Verhältnis zweier Varianzen

Nr.	Prozessschritt
1	Wahl einer Konfidenzzahl γ
2	Bestimmung der zugehörigen Parameter c_1 und c_2 aus der inversen f-Verteilung mit ($N - 1$, $M - 1$) Freiheitsgraden $c_1 = F^{-1}\left(\frac{1 - \gamma}{2}\right) \quad \text{und} \quad c_2 = F^{-1}\left(\frac{1 + \gamma}{2}\right)$
3	Berechnung der Mittelwerte aus den Stichproben $s_1^2 = \frac{1}{N - 1} \cdot \sum_{n=1}^N (x_{1n} - \bar{x}_1)^2 \quad \text{und} \quad s_2^2 = \frac{1}{M - 1} \cdot \sum_{m=1}^M (x_{2m} - \bar{x}_2)^2$
4	Bestimmung des Konfidenzintervalls $\frac{s_2^2}{s_1^2} \cdot c_1 \leq \frac{\gamma_2^2}{\gamma_1^2} \leq \frac{s_2^2}{s_1^2} \cdot c_2$

Beispiel: Statistische Prozesskontrolle einer Gehäusefertigung

Im Rahmen der statistischen Prozesskontrolle einer Gehäusefertigung werden zwei Stichproben mit einem Umfang von $N = M = 5$ Gehäusen aufgenommen. Hierzu werde bei fünf gefertigten Gehäuseteilen die Stirnseite vermessen. Die ermittelten Messwerte sind in Tabelle 5.11 aufgelistet. Es soll das 95%-Konfidenzintervall für das Verhältnis der Varianzen berechnet werden.

Tabelle 5.11: Messung der Stirnseite von Kunststoffgehäusen

Index	Stichprobe 1: Stirnbreite / mm				
1 - 5	0.13	0.1	0.11	0.12	0.11
Index	Stichprobe 2: Stirnbreite / mm				
1 - 5	0.11	0.14	0.10	0.12	0.13

Es ist eine Konfidenzzahl von $\gamma = 0.95$ gefordert. Daraus ergeben sich die Parameter $c_1 = 0.1041$ und $c_2 = 9.6045$ aus der inversen f-Verteilung mit ($N - 1$, $M - 1$) = (4|4) Freiheitsgraden.

Die Auswertung der Stichproben führt zu den Stichprobenvarianzen

$$s_1^2 = 0.00013 \text{ mm}^2 \quad (5.110)$$

und

$$s_2^2 = 0.00025 \text{ mm}^2 \quad (5.111)$$

Daraus errechnet sich ein Konfidenzbereich

$$0.2002 < \frac{\sigma_2^2}{\sigma_1^2} \leq 18.4702 \quad (5.112)$$

5.4.4 Zusammenfassung der Konfidenzbereiche für den Vergleich von Stichproben

Die Schätzung des Konfidenzbereiches beruht wieder darauf, eine Zufallsvariable mit einer bekannten Verteilung zu finden, in deren Beschreibung der gesuchte Parameter der Grundgesamtheit und der bekannte Parameter der Stichprobe vorkommen. Das grundsätzliche Vorgehen ist damit identisch zu der Berechnung der Konfidenzbereiche in Abschnitt 5.3.

Zum Vergleich zweier Stichproben aus normalverteilten Grundgesamtheiten ergeben sich die in Tabelle 5.12 dargestellten Verteilungen und Konfidenzbereiche.

Tabelle 5.12: Verteilungen und Konfidenzbereiche von Stichproben normalverteilter Grundgesamtheiten

	Verteilung der Schätzfunktion	Konfidenzbereich
Differenz zweier Mittelwerte $\mu_1 - \mu_2$ bei bekannter Varianz	Standardnormalverteilung $z = \frac{\bar{x} - \mu}{\sqrt{\sigma_x^2}} = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\sigma^2 \cdot \left(\frac{1}{N} + \frac{1}{M}\right)}}$	$\mu_{C1} < (\mu_1 - \mu_2) \leq \mu_{C2}$ mit $\mu_{C1} = (\bar{x}_1 - \bar{x}_2) - c_2 \cdot \sqrt{\sigma^2 \cdot \left(\frac{1}{N} + \frac{1}{M}\right)}$ $\mu_{C2} = (\bar{x}_1 - \bar{x}_2) - c_1 \cdot \sqrt{\sigma^2 \cdot \left(\frac{1}{N} + \frac{1}{M}\right)}$
Differenz zweier Mittelwerte $\mu_1 - \mu_2$ bei unbekannter Varianz	t-Verteilung mit $N + M - 2$ Freiheitsgraden $t = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{1}{N} + \frac{1}{M}} \cdot s}$	$\mu_{C1} < (\mu_1 - \mu_2) \leq \mu_{C2}$ mit $\mu_{C1} = (\bar{x}_1 - \bar{x}_2) - c_2 \cdot \sqrt{\frac{1}{N} + \frac{1}{M}} \cdot s$ $\mu_{C2} = (\bar{x}_1 - \bar{x}_2) - c_1 \cdot \sqrt{\frac{1}{N} + \frac{1}{M}} \cdot s$
Verhältnis zweier Stichprobenvarianzen	f-Verteilung mit $(N - 1, M - 1)$ Freiheitsgraden $f = \frac{s_1^2}{s_2^2} \cdot \frac{\sigma_2^2}{\sigma_1^2}$	$\frac{s_2^2}{s_1^2} \cdot c_1 < \frac{\sigma_2^2}{\sigma_1^2} \leq \frac{s_2^2}{s_1^2} \cdot c_2$

5.5 Vorhersageintervalle für künftige Stichprobenwerte

Bei der Prognose von Toleranzen wird berechnet, in welchem Bereich eine zukünftige Beobachtung mit einer festgelegten Wahrscheinlichkeit liegen wird. Auf diese Weise ist es möglich, Toleranzgrenzen für ein Bauteil oder ein Produkt zu definieren, sodass eine festgelegte Ausschussquote nicht überschritten wird. Bei der Prognose muss wie bei den Konfidenzintervallen eine Annahme über die Verteilung der Grundgesamtheit getroffen werden. Da sich die meisten Verteilungen für eine große Anzahl Stichprobenwerte asymptotisch an eine Normalverteilung annähern, wird bei den folgenden Betrachtungen eine Normalverteilung mit dem Mittelwert μ und der Varianz σ^2 zu Grunde gelegt.

Bei der Berechnung der Vorhersageintervalle muss unterschieden werden, ob die Parameter der Normalverteilung μ und σ bekannt, teilweise bekannt oder vollkommen unbekannt sind. Entsprechend der bekannten und unbekannten Parameter wird die neue Verteilung unter Berücksichtigung der Ungenauigkeit der Parameterschätzung bestimmt und zur Berechnung der Intervallgrenzen herangezogen.

5.5.1 Vorhersageintervall bei bekanntem Mittelwert und bekannter Varianz

Zunächst wird die Berechnung eines Vorhersageintervalls für den Fall betrachtet werden, dass sowohl der Mittelwert μ als auch die Varianz σ^2 der Normalverteilung der Grundgesamtheit bekannt sind. Die Zufallsvariable

$$z = \frac{x - \mu}{\sigma} \quad (5.113)$$

ist standardnormalverteilt. Die Wahrscheinlichkeit γ , mit der ein zukünftiger Beobachtungswert in dem Intervall $c_1 \dots c_2$ liegt, ergibt sich aus

$$P(c_1 < z \leq c_2) = F(c_2) - F(c_1) = \gamma \quad (5.114)$$

Bei Annahme eines symmetrischen Vorhersagebereiches ergeben sich die Konstanten c_1 und c_2 aus den Bedingungen

$$F(c_1) = \frac{1 - \gamma}{2} \quad (5.115)$$

und

$$F(c_2) = 1 - \frac{1 - \gamma}{2} = \frac{1 + \gamma}{2} \quad (5.116)$$

Dabei ist $F(x)$ die Verteilungsfunktion einer Standardnormalverteilung. Auflösen nach c_1 und c_2 führt zu

$$c_1 = F^{-1}\left(\frac{1 - \gamma}{2}\right) \quad (5.117)$$

und

$$c_2 = F^{-1}\left(\frac{1 + \gamma}{2}\right) \quad (5.118)$$

Durch Umformungen ergibt sich ein Ausdruck für den Prognosebereich zukünftiger Werte bei bekanntem Mittelwert μ und bekannter Varianz σ^2 .

$$\gamma = P\left(c_1 < \frac{x - \mu}{\sigma} \leq c_2\right) = P(\mu + c_1 \cdot \sigma < x \leq \mu + c_2 \cdot \sigma) \quad (5.119)$$

In Tabelle 5.13 wird das Vorgehen zur Berechnung eines Vorhersageintervalls mithilfe der Standardnormalverteilung zusammengefasst.

Tabelle 5.13: Vorgehen zur Bestimmung des Vorhersageintervalls für künftige Stichprobenwerte einer Normalverteilung mit bekanntem Mittelwert und bekannter Varianz mithilfe der Standardnormalverteilung

Nr.	Prozessschritt
1	Wahl einer Konfidenzzahl γ
2	Bestimmung der zugehörigen Parameter c_1 und c_2 aus der inversen Standardnormalverteilung $c_1 = F^{-1}\left(\frac{1-\gamma}{2}\right) \quad \text{und} \quad c_2 = F^{-1}\left(\frac{1+\gamma}{2}\right)$
3	Bestimmung des Prognoseintervalls $\mu + c_1 \cdot \sigma < x \leq \mu + c_2 \cdot \sigma$

Beispiel: Widerstandsfertigung

Als Beispiel werden die Toleranzgrenzen für eine Widerstandsfertigung berechnet. Von dem Prozess ist der Mittelwert $\mu = 998 \Omega$ und die Standardabweichung von $\sigma = 5 \Omega$ bekannt. Die Toleranzgrenzen sollen so gewählt werden, dass maximal 5 % Ausschuss bei der Fertigung entstehen.

Die Grenzen $c_{1,2}$ berechnen sich aus der Standardnormalverteilung unter Berücksichtigung der geforderten Konfidenzzahl von $\gamma = 0.95$ zu $c_1 = -1.96$ und $c_2 = 1.96$. Damit folgt mit der Bedingung

$$\mu + c_1 \cdot \sigma < x \leq \mu + c_2 \cdot \sigma \quad (5.120)$$

für das Vorhersageintervall der Widerstandsfertigung

$$988.2\Omega < x \leq 1007.8\Omega \quad (5.121)$$

Zukünftige Widerstandswerte liegen mit einer Wahrscheinlichkeit von 95 % innerhalb der Grenzen 988.2Ω und 1007.8Ω . Die Toleranz des Bauteils kann damit auf $998 \pm 9.8\Omega$ definiert werden, damit ein Ausschussanteil von 5 % nicht überschritten wird.

5.5.2 Vorhersageintervall bei unbekanntem Mittelwert und bekannter Varianz

Im letzten Abschnitt wird davon ausgegangen, dass sowohl der Mittelwert μ als auch die Varianz σ^2 der Normalverteilung der Grundgesamtheit bekannt sind. In diesem Abschnitt wird angenommen, dass der Mittelwert μ mithilfe der Parameterschätzung geschätzt werden muss, während die Varianz der Grundgesamtheit bekannt ist.

Die Werte der Stichprobe können als unabhängige Zufallsvariable x_1, x_2, \dots, x_N mit demselben, unbekannten Mittelwert μ und derselben, bekannten Varianz σ^2 aufgefasst werden. Damit ist der arithmetische Mittelwert

$$\bar{x} = \frac{1}{N} \cdot \sum_{n=1}^N x_n \quad (5.122)$$

ebenfalls normalverteilt mit dem Mittelwert μ und der Varianz σ^2/N . Die Zufallsvariable

$$z = \frac{x - \bar{x}}{\sqrt{\sigma^2 + \frac{\sigma^2}{N}}} \quad (5.123)$$

ist unter diesen Voraussetzungen standardnormalverteilt. Die Wahrscheinlichkeit γ , mit der die Variable z in dem Intervall $c_1 \dots c_2$ liegt, ergibt sich aus

$$P(c_1 < z \leq c_2) = F(c_2) - F(c_1) = \gamma \quad (5.124)$$

Wieder wird ein symmetrischer Vorhersagebereich angenommen, sodass sich die Konstanten c_1 und c_2 aus den Bedingungen

$$F(c_1) = \frac{1 - \gamma}{2} \quad (5.125)$$

und

$$F(c_2) = 1 - \frac{1 - \gamma}{2} = \frac{1 + \gamma}{2} \quad (5.126)$$

ergeben. Dabei ist $F(x)$ die Verteilungsfunktion einer Standardnormalverteilung. Auflösen nach c_1 und c_2 führt zu

$$c_1 = F^{-1}\left(\frac{1 - \gamma}{2}\right) \quad (5.127)$$

und

$$c_2 = F^{-1}\left(\frac{1 + \gamma}{2}\right) \quad (5.128)$$

Durch Umformungen ergibt sich ein Ausdruck für den Prognosebereich zukünftiger Werte bei unbekanntem Mittelwert μ und bekannter Varianz σ^2 .

$$\gamma = P(c_1 < z \leq c_2) = P\left(c_1 < \frac{x - \bar{x}}{\sqrt{\sigma^2 + \frac{\sigma^2}{N}}} \leq c_2\right) = P\left(\bar{x} + c_1 \cdot \sqrt{\sigma^2 + \frac{\sigma^2}{N}} < x \leq \bar{x} + c_2 \cdot \sqrt{\sigma^2 + \frac{\sigma^2}{N}}\right) \quad (5.129)$$

In Tabelle 5.14 wird das Vorgehen zur Berechnung eines Vorhersageintervalls mithilfe der Standardnormalverteilung zusammengefasst.

Tabelle 5.14: Vorgehen zur Bestimmung des Vorhersageintervalls für künftige Stichprobenwerte einer Normalverteilung mit unbekanntem Mittelwert und bekannter Varianz mithilfe der Standardnormalverteilung

Nr.	Prozessschritt
1	Wahl einer Konfidenzzahl γ
2	Bestimmung der zugehörigen Parameter c_1 und c_2 aus der inversen Standardnormalverteilung $c_1 = F^{-1}\left(\frac{1 - \gamma}{2}\right)$ und $c_2 = F^{-1}\left(\frac{1 + \gamma}{2}\right)$
3	Berechnung des Mittelwertes der Stichprobe $\bar{x} = \frac{1}{n} \cdot \sum_{i=1}^n x_i$
4	Bestimmung des Prognoseintervalls $\bar{x} + c_1 \cdot \sqrt{\sigma^2 + \frac{\sigma^2}{n}} < x \leq \bar{x} + c_2 \cdot \sqrt{\sigma^2 + \frac{\sigma^2}{n}}$

Beispiel: Abfüllen von Granulat

Im folgenden Beispiel soll das Vorhersageintervall für das Gewicht von abgefülltem Granulat berechnet werden. Damit soll bestimmt werden, innerhalb welcher Grenzen $\gamma = 95\%$ der abgefüllten Mengen liegen, wenn der Prozess eine Standardabweichung von $\sigma = 10 \text{ g}$ besitzt. Für die Berechnung wurde die in Tabelle 5.15 aufgelistete Stichprobe gemessen.

Tabelle 5.15: Stichprobe für das Gewicht von Granulatmengen in einem Fertigungsprozess

Stichprobe	1	2	3	4	5	6	7	8	9	10
Gewicht x / kg	1.0729	1.0541	1.0566	1.0771	1.0556	1.0757	1.0862	1.0631	1.0786	1.0825

Daraus berechnet sich der arithmetische Mittelwert zu

$$\bar{x} = \frac{1}{N} \cdot \sum_{n=1}^N x_n = 1070.3 \text{ g} \quad (5.130)$$

Aus der Konfidenzzahl von $\gamma = 0.95$ folgen der inversen Standardnormalverteilung die beiden Konstanten $c_1 = -1.96$ und $c_2 = 1.96$. Aus der Bedingung

$$\bar{x} + c_1 \cdot \sqrt{\sigma^2 + \frac{\sigma^2}{N}} < x \leq \bar{x} + c_2 \cdot \sqrt{\sigma^2 + \frac{\sigma^2}{N}} \quad (5.131)$$

folgt das Vorhersageintervall für zukünftige Beobachtungen bei einer Konfidenzzahl von $\gamma = 0.95$ zu
 $1049.7 \text{ g} < x_{n+1} \leq 1090.9 \text{ g}$ (5.132)

95 % der abgefüllten Granulatmengen liegen innerhalb der Grenzen von 1049.7 g und 1090.9 g.

5.5.3 Vorhersageintervall bei bekanntem Mittelwert und unbekannter Varianz

Zur Bestimmung des Vorhersageintervalls für zukünftige Beobachtungen einer Normalverteilung mit bekanntem Mittelwert μ und unbekannter Varianz σ^2 muss zunächst die Standardabweichung der Stichprobe

$$s = \sqrt{\frac{1}{N-1} \cdot \sum_{n=1}^N (x_n - \bar{x})^2} \quad (5.133)$$

berechnet werden. In Kapitel 8 wird gezeigt, dass die Zufallsvariable

$$\chi = \frac{(N-1) \cdot s^2}{\sigma^2} \quad (5.134)$$

eine chi-quadrat-verteilte Größe mit $N - 1$ Freiheitsgraden ist. Ein zukünftiger Beobachtungswert der normalverteilten Grundgesamtheit ist wiederum normalverteilt mit dem Mittelwert μ und der Varianz σ^2 . Mithilfe dieser beiden Zufallsgrößen aus Gleichung (5.133) und Gleichung (5.134) kann die Zufallsvariable

$$t = \frac{x - \mu}{s} \quad (5.135)$$

mit t-Verteilung gebildet werden. Sie besitzt ebenfalls $N - 1$ Freiheitsgrade. Die Wahrscheinlichkeit γ , mit der die Variable t in dem Intervall $c_1 \dots c_2$ liegt, ergibt sich aus

$$P(c_1 < t \leq c_2) = F(c_2) - F(c_1) = \gamma \quad (5.136)$$

Bei Annahme eines symmetrischen Vorhersagebereiches ergeben sich die Konstanten c_1 und c_2 aus den Bedingungen

$$F(c_1) = \frac{1-\gamma}{2} \quad (5.137)$$

und

$$F(c_2) = 1 - \frac{1-\gamma}{2} = \frac{1+\gamma}{2} \quad (5.138)$$

Dabei ist $F(x)$ die Verteilungsfunktion einer t-Verteilung mit $N - 1$ Freiheitsgraden. Auflösen nach c_1 und c_2 führt zu

$$c_1 = F^{-1}\left(\frac{1-\gamma}{2}\right) \quad (5.139)$$

und

$$c_2 = F^{-1}\left(\frac{1+\gamma}{2}\right) \quad (5.140)$$

Durch Umformungen ergibt sich ein Ausdruck für den Prognosebereich zukünftiger Werte bei bekanntem Mittelwert μ und unbekannter Varianz σ^2 .

$$\gamma = P(c_1 < t \leq c_2) = P\left(c_1 < \frac{x-\mu}{s} \leq c_2\right) = P(\mu + c_1 \cdot s < x \leq \mu + c_2 \cdot s) \quad (5.141)$$

In Tabelle 5.16 wird das Vorgehen zur Berechnung eines Vorhersageintervalls mithilfe der t-Verteilung zusammengefasst.

Tabelle 5.16: Vorgehen zur Bestimmung des Vorhersageintervalls für künftige Stichprobenwerte einer Normalverteilung mit bekanntem Mittelwert und unbekannter Varianz mithilfe der t-Verteilung

Nr.	Prozessschritt
1	Wahl einer Konfidenzzahl γ
2	Bestimmung der zugehörigen Parameter c_1 und c_2 aus der inversen t-Verteilung mit $N - 1$ Freiheitsgraden $c_1 = F^{-1}\left(\frac{1-\gamma}{2}\right)$ und $c_2 = F^{-1}\left(\frac{1+\gamma}{2}\right)$
3	Berechnung der Standardabweichung aus der Stichprobe $s = \sqrt{\frac{1}{n-1} \cdot \sum_{i=1}^n (x_i - \bar{x})^2}$
4	Bestimmung des Prognoseintervalls $\mu + c_1 \cdot s < x \leq \mu + c_2 \cdot s$

Beispiel: Fertigung von Unterlegscheiben

Einem Hersteller von Unterlegscheiben ist bekannt, dass sein Fertigungsprozess zur Herstellung von Unterlegscheiben einen Mittelwert des Durchmessers von 49.5 mm besitzt. Die Streuung des Prozesses muss anhand einer Stichprobe abgeschätzt werden. Zu diesem Zweck wurden 15 Unterlegscheiben zufällig aus der Produktion vermessen. Die Ergebnisse sind in Tabelle 5.17 aufgelistet.

Tabelle 5.17: Stichprobe für den Durchmesser von Unterlegscheiben in einem Fertigungsprozess

Durchmesser der Unterlegscheiben d / mm							
49.99	46.45	47.50	50.00	49.48	51.91	50.33	50.42
49.59	51.49	48.94	49.63	48.59	50.59	48.24	

Mithilfe der Stichprobe sollen die Toleranzgrenzen bestimmt werden, wobei maximal 5 % der gefertigten Unterlegscheiben außerhalb dieser Spezifikation liegen dürfen. Hierzu wird zunächst die Standardabweichung der Stichprobe berechnet.

$$s = \sqrt{\frac{1}{N-1} \cdot \sum_{n=1}^N (x_n - \bar{x})^2} = 1.4395 \text{ mm} \quad (5.142)$$

Die Konstanten c_1 und c_2 berechnet sich aus der inversen t-Verteilung mit $N - 1 = 14$ Freiheitsgraden zu $c_1 = -2.1448$ und $c_2 = 2.1448$. Damit ergibt sich ein Vorhersageintervall von

$$46.4 \text{ mm} < x \leq 52.6 \text{ mm} \quad (5.143)$$

Der Fertigungsprozess hält somit bei 95 % der gefertigten Teile eine Toleranzgrenze von ± 3.1 mm um den Mittelwert von $\mu = 49.5$ mm ein.

5.5.4 Vorhersageintervall bei unbekanntem Mittelwert und unbekannter Varianz

Im allgemeisten Fall muss davon ausgegangen werden, dass weder der Mittelwert μ noch die Varianz σ^2 der normalverteilten Grundgesamtheit bekannt ist. Damit ist es erforderlich, beide Größen auf Basis einer Stichprobe abzuschätzen. Die Stichprobenfunktionen berechnen sich aus

$$\bar{x} = \frac{1}{N} \cdot \sum_{n=1}^N x_n \quad (5.144)$$

und

$$s = \sqrt{\frac{1}{N-1} \cdot \sum_{n=1}^N (x_n - \bar{x})^2} \quad (5.145)$$

Da die Werte des Mittelwertes und der Standardabweichung auf zufälligen Stichprobenwerten beruhen, sind auch die berechneten Schätzwerte zufällige Größen. Diese Unsicherheit muss bei der Berechnung des Vorhersageintervalls berücksichtigt werden. In Kapitel 8 wird gezeigt, dass die Zufallsvariable

$$z = \frac{x - \bar{x}}{\sqrt{\sigma^2 + \frac{\sigma^2}{N}}} \quad (5.146)$$

standardnormalverteilt ist und dass

$$\chi = \frac{(N-1) \cdot s^2}{\sigma^2} \quad (5.147)$$

eine chi-quadrat-verteilte Zufallsvariable mit $N - 1$ Freiheitsgraden ist. Aus Gleichung (5.146) und Gleichung (5.147) kann eine Zufallsvariable der t-Verteilung mit $N - 1$ Freiheitsgraden gebildet werden.

$$t = \frac{x - \bar{x}}{\sqrt{\sigma^2 + \frac{\sigma^2}{N}} \cdot \sqrt{\frac{(N-1) \cdot s^2}{\sigma^2 \cdot (N-1)}}} = \frac{x - \bar{x}}{s \cdot \sqrt{1 + \frac{1}{N}}} \quad (5.148)$$

Die Wahrscheinlichkeit γ , mit der die Variable t in dem Intervall $c_1 \dots c_2$ liegt, ergibt sich aus

$$P(c_1 < t \leq c_2) = F(c_2) - F(c_1) = \gamma \quad (5.149)$$

Bei Annahme eines symmetrischen Vorhersagebereiches ergeben sich die Konstanten c_1 und c_2 aus den Bedingungen

$$F(c_1) = \frac{1 - \gamma}{2} \quad (5.150)$$

und

$$F(c_2) = 1 - \frac{1 - \gamma}{2} = \frac{1 + \gamma}{2} \quad (5.151)$$

Dabei ist $F(x)$ die Verteilungsfunktion einer t-Verteilung mit $N - 1$ Freiheitsgraden. Auflösen nach c_1 und c_2 führt zu

$$c_1 = F^{-1}\left(\frac{1 - \gamma}{2}\right) \quad (5.152)$$

und

$$c_2 = F^{-1}\left(\frac{1 + \gamma}{2}\right) \quad (5.153)$$

Durch Umformungen ergibt sich ein Ausdruck für den Prognosebereich zukünftiger Werte bei unbekanntem Mittelwert μ und unbekannter Varianz σ^2 .

$$\gamma = P(c_1 < t \leq c_2) = P\left(c_1 < \frac{x - \bar{x}}{s \cdot \sqrt{1 + \frac{1}{N}}} \leq c_2\right) = P\left(\bar{x} + c_1 \cdot s \cdot \sqrt{1 + \frac{1}{N}} < x \leq \bar{x} + c_2 \cdot s \cdot \sqrt{1 + \frac{1}{N}}\right) \quad (5.154)$$

In Tabelle 5.18 wird das Vorgehen zur Berechnung eines Vorhersageintervalls mithilfe der t-Verteilung zusammengefasst.

Tabelle 5.18: Vorgehen zur Bestimmung des Vorhersageintervalls für künftige Stichprobenwerte einer Normalverteilung mit unbekanntem Mittelwert und unbekannter Varianz mithilfe der t-Verteilung

Nr.	Prozessschritt
1	Wahl einer Konfidenzzahl γ
2	Bestimmung der zugehörigen Parameter c_1 und c_2 aus der inversen Standardnormalverteilung $c_1 = F^{-1}\left(\frac{1 - \gamma}{2}\right)$ und $c_2 = F^{-1}\left(\frac{1 + \gamma}{2}\right)$
3	Berechnung des Mittelwertes aus der Stichprobe $\bar{x} = \frac{1}{N} \cdot \sum_{n=1}^N x_n$ und $s = \sqrt{\frac{1}{N-1} \cdot \sum_{n=1}^N (x_n - \bar{x})^2}$
4	Bestimmung des Prognoseintervalls $\bar{x} + c_1 \cdot s \cdot \sqrt{1 + \frac{1}{N}} < x \leq \bar{x} + c_2 \cdot s \cdot \sqrt{1 + \frac{1}{N}}$

Beispiel: Widerstandsfertigung

Die Berechnung des Vorhersageintervalls soll am Beispiel einer Fertigungseinrichtung für Widerstände verdeutlicht werden. Hierbei soll auf Grundlage einer Stichprobe von 40 Widerstandswerten ein Toleranzband definiert werden, das 95 % aller künftig gefertigten Widerstände enthält. Der Mittelwert und die Varianz der Grundgesamtheit sind unbekannt.

Tabelle 5.19: Stichprobe von 40 Widerständen zur Bestimmung eines Toleranzbandes

Widerstandsmessung R / k?				
985.35	987.95	987.84	984.71	986.76
985.90	985.69	984.48	986.35	985.79
987.84	986.46	986.63	986.23	987.30
985.07	986.00	986.28	986.49	986.32
985.72	988.96	987.74	986.48	987.06
987.39	988.32	985.57	984.43	984.16
989.40	983.61	987.08	988.70	989.64
985.28	986.38	986.13	989.06	987.48

Aus den Stichprobenwerten wird zunächst der arithmetische Mittelwert von

$$\bar{x} = \frac{1}{N} \cdot \sum_{n=1}^N x_n = 986.6\Omega \quad (5.155)$$

und die Standardabweichung von

$$s = \sqrt{\frac{1}{N-1} \cdot \sum_{n=1}^N (x_n - \bar{x})^2} = 1.4558\Omega \quad (5.156)$$

bestimmt. Die Konstanten c_1 und c_2 folgen aus der inversen t-Verteilung mit $N - 1 = 39$ Freiheitsgraden zu $c_1 = -2.0227$ und $c_2 = 2.0227$. Mit den berechneten Werten und der Bedingung

$$\bar{x} + c_1 \cdot s \cdot \sqrt{1 + \frac{1}{N}} < x \leq \bar{x} + c_2 \cdot s \cdot \sqrt{1 + \frac{1}{N}} \quad (5.157)$$

ergibt sich ein Vorhersageintervall für künftige Widerstandswerte zu

$$983.6\Omega < x \leq 989.6\Omega \quad (5.158)$$

Der Fertigungsprozess hält somit bei 95 % der gefertigten Teile eine Toleranzgrenze von $\pm 2.98\Omega$ um einen mittleren Widerstandswert von 986.6Ω ein.

5.5.5 Zusammenfassung der Vorhersageintervalle für künftige Stichprobenwerte

Die Schätzung der Vorhersageintervalle beruht ebenfalls auf einer Zufallsvariable mit einer bekannten Verteilung, in deren Beschreibung der gesuchte Parameter der Grundgesamtheit und der bekannte Parameter der Stichprobe vorkommen. Das grundsätzliche Vorgehen ist damit vergleichbar zu der Berechnung der Konfidenzbereiche in Abschnitt 5.3.

Zur Abschätzung der Vorhersageintervalle zukünftiger Stichprobenwerte bei einer normalverteilten Grundgesamtheit werden in Abhängigkeit der bekannten und unbekannten Parameter die in Tabelle 5.20 zusammengefassten Verteilungen und Prognoseintervalle verwendet.

Tabelle 5.20: Verteilungen und Vorhersagebereiche von Stichproben normalverteilter Grundgesamtheiten

	Verteilung der Schätzfunktion	Vorhersageintervall
bekannter Mittelwert ν bekannte Varianz σ^2	Standardnormalverteilung $z = \frac{x - \mu}{\sigma}$	$\mu + c_1 \cdot \sigma < x \leq \mu + c_2 \cdot \sigma$
unbekannter Mittelwert ν bekannte Varianz σ^2	Standardnormalverteilung $z = \frac{x - \bar{x}}{\sqrt{\sigma^2 + \frac{\sigma^2}{N}}}$	$\bar{x} + c_1 \cdot \sqrt{\sigma^2 + \frac{\sigma^2}{N}} < x \leq \bar{x} + c_2 \cdot \sqrt{\sigma^2 + \frac{\sigma^2}{N}}$
bekannter Mittelwert ν bekannte Varianz σ^2	t-Verteilung mit $N - 1$ Freiheitsgraden $t = \frac{x - \mu}{s}$	$\mu + c_1 \cdot s < x \leq \mu + c_2 \cdot s$
unbekannter Mittelwert ν bekannte Varianz σ^2	t-Verteilung mit $N - 1$ Freiheitsgraden $t = \frac{x - \bar{x}}{s \cdot \sqrt{1 + \frac{1}{N}}}$	$\bar{x} + c_1 \cdot s \cdot \sqrt{1 + \frac{1}{N}} < x \leq \bar{x} + c_2 \cdot s \cdot \sqrt{1 + \frac{1}{N}}$

5.6 Anwendungsbeispiel: Kontaktwiderstand eines Batteriesensors

Zur Überwachung des Batteriezustandes wird in Kraftfahrzeugen ein elektronischer Batteriesensor eingesetzt. Er erfasst die Batteriegrößen Strom, Spannung, Temperatur und die Zeit. Auf Basis der gemessenen Werte berechnet ein im Sensor implementierter Algorithmus den aktuellen Zustand der Batterie. Mit der Kenntnis dieses Batteriezustandes kann das Energiemanagement im Fahrzeug gezielt Energieverbraucher aus- und auch wieder einschalten, um sicherzustellen, dass die Ladung der Batterie für einen sicheren Fahrzeugstart ausreicht.



Bild 5.9: Elektronischer Batteriesensor (Robert Bosch GmbH, Geschäftsbereich AE Automotive Electronics)

Der Sensor wird mit einer Polklemme auf dem Batterie-Minus-Pol montiert. Die Verbindung zur Karosserie (Massekabel) erfolgt über einen Kabelschuh, der mit dem Batteriesensor verschraubt wird. Über diesen Schraubkontakt können Ströme im Bereich von mehreren 100 A fließen, sodass der Übergangswiderstand zwischen Sensor und Kabelschuh klein gehalten werden muss. Dieser Widerstand setzt sich aus dem Fremdschichtwiderstand, der durch immer vorhandene, dünne Oxid- oder Sulfidschichten verursacht wird, und dem sogenannten Engewiderstand zusammen. Die Summe beider Widerstände darf einen maximalen Wert von $R_{E,MAX}$ von $100\mu\Omega$ nicht überschreiten.

Der Engewiderstand ergibt sich aus der Berührfläche des Batteriesensorkontaktes und des Kabelschuhs. Beide Partner sind nicht ideal eben und weisen eine Oberflächenrauhigkeit auf, die effektive Kontaktfläche ist daher erheblich kleiner als die Grundfläche der Kontakte. Der fließende Strom muss durch diese Engstellen fließen. Daraus ergibt sich der Name Engewiderstand.

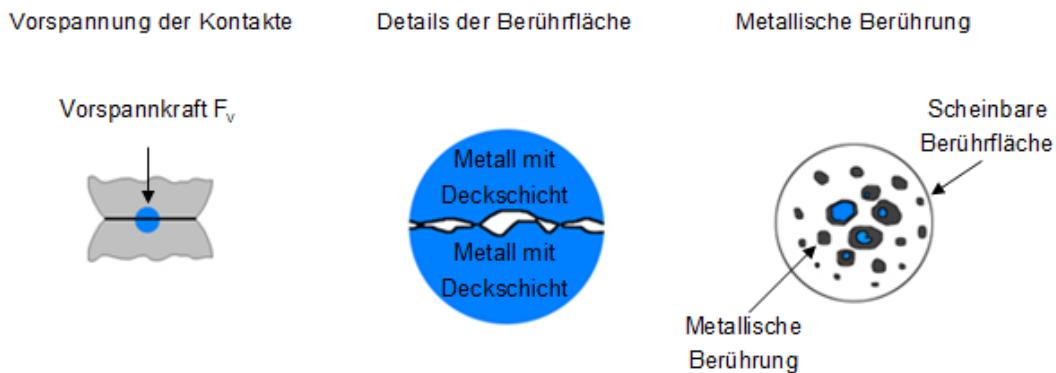


Bild 5.10: Beschreibung der effektiven Kontaktfläche eines Batteriesensors

Der elektrische Kontakt besteht nur an wenigen, sehr kleinen Berührflächen. Durch die Vorspannkraft F_V , die durch die Verschraubung aufgebracht wird, werden diese Kontaktstellen plastisch verformt, so dass sie an Größe und Anzahl wachsen.

Die mathematische Modellierung [Holm67, Höft80] zeigt, dass im Fall der plastischen Verformung der Engewiderstand R_E im Wesentlichen von der spezifischen Leitfähigkeit ρ , der Kontakthärte H und der

Vorspannkraft F_V abhängt und durch die Gleichung

$$R_E = \frac{\rho}{2} \cdot \sqrt{\frac{\pi \cdot H}{F_V}} \quad (5.159)$$

beschrieben werden kann. Um die theoretische Berechnung des Engewiderstandes zu hinterfragen, wurden Musterteile des Batteriesensors aufgebaut, an denen der Engewiderstand gemessen wurde. Bild 5.11 zeigt das Ergebnis der mathematischen Modellierung des Engewiderstandes und einer elektrischen Messung des Engewiderstandes an Neuteilen. Es zeigt sich, dass die theoretische Berechnung des Engewiderstandes für Vorspannkräfte, die größer als 1500 N sind, als konservative Abschätzung für den Engewiderstand verwendet werden kann.

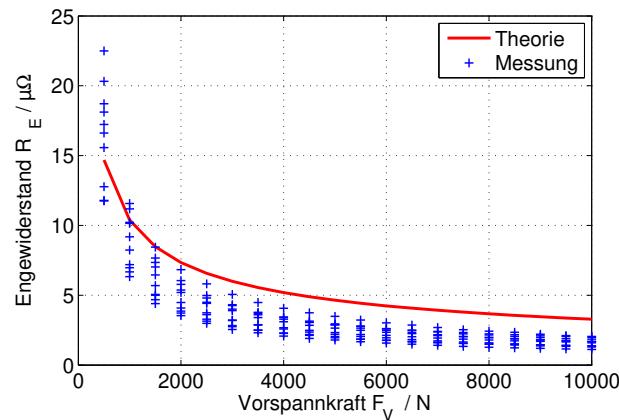


Bild 5.11: Ergebnis einer mathematischen Modellierung des Engewiderstandes R_E und einer elektrischen Messung des Engewiderstandes R_E an Neuteilen

Neben der Vorspannkraft hat die Oberflächenbeschaffenheit einen starken Einfluss auf den Kontaktwiderstand. Deshalb wurde in zwei Szenarien analysiert, wie sich unterschiedliche Vorbehandlungen auf den Widerstandswert auswirken. Die Sensoren wurden zum einen in Standardverpackungen als Seefracht transportiert, um den Einfluss der Seeatmosphäre zu bewerten. Zum anderen durchliefen die Teile mehrfach eine Löteinrichtung mit anschließender Abkühlung in einer Industrieatmosphäre, so dass die Oberfläche der Kontakte oxidierte. Bild 5.12 stellt für beide Vorbehandlungen die Ergebnisse der Nachmessung und die mit Gleichung (5.159) berechneten Widerstandswerte dar.

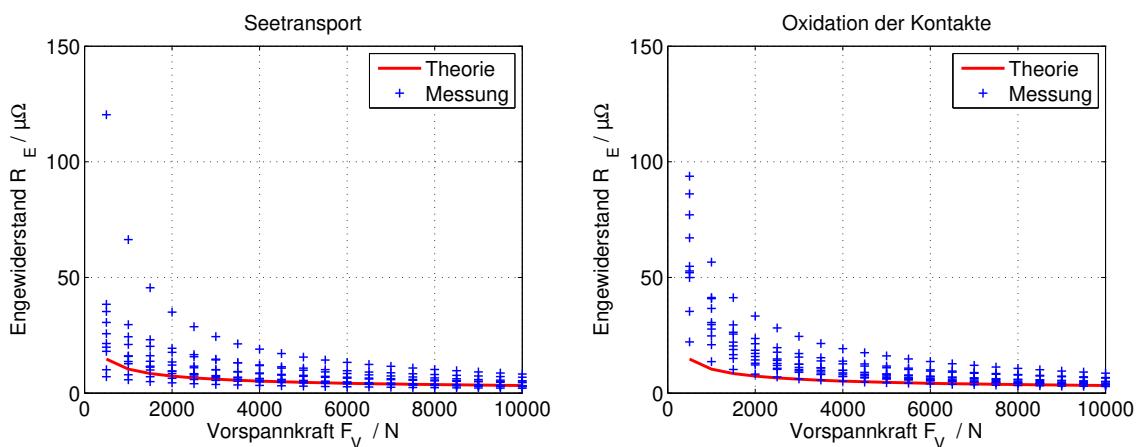


Bild 5.12: Elektrische Messung des Kontaktwiderstandes nach Vorbehandlung a) Simulierter Seetransport der Shunt-Widerstände b) Oxidierte Shunt-Widerstände

Durch die Vorbehandlung steigt der Kontaktwiderstand erwartungsgemäß an, der stärkste Anstieg wird bei Seetransport verzeichnet. Bei niedrigen Vorspannkräften steigt der Kontaktwiderstand gegenüber

der Messung bei Neuteilen um bis zu einem Faktor 10.

Für die Montage der Sensoren und die Auslegung der Schraubverbindung muss die Frage beantwortet werden, wie groß die von der Verschraubung aufzubringende Vorspannkraft F_V sein muss, damit der Widerstand unter der definierten Grenze von $R_{E,MAX} = 100 \mu\Omega$ bleibt. Hier wird vereinfachend davon ausgegangen, dass nur die beschriebenen Einflüsse auf den Kontaktwiderstand auftreten und dass sich die Effekte nicht überlagern. Die Sensoren dürfen in diesem Beispiel außerdem nur mit einer Wahrscheinlichkeit von 10 ppm einen Kontaktwiderstand oberhalb des definierten Grenzwertes $R_{E,MAX}$ aufweisen.

Für jede Kombination von Vorspannkraft und Vorbehandlung existieren 10 Stichprobenwerte. Sie sind in Tabelle 5.21 exemplarisch für die Vorbehandlung durch Oxidation und eine Vorspannkraft von 6000 N zusammengestellt.

Tabelle 5.21: Stichprobe von Messwerten des Kontaktwiderstandes bei Vorbehandlung durch Oxidation und einer Vorspannkraft von 6000 N

Nr.	1	2	3	4	5	6	7	8	9	10
$R_E / \mu\Omega$	3.984	7.349	6.451	5.952	7.076	13.633	10.065	5.478	9.529	5.263

Zur besseren Übersicht zeigt Bild 5.13 die Häufigkeitsverteilung und den Box-Plot der Stichprobe bei Vorbehandlung durch Seetransport und einer Vorspannkraft von 6000 N. Es wird von einer normalverteilten Grundgesamtheit ausgegangen.

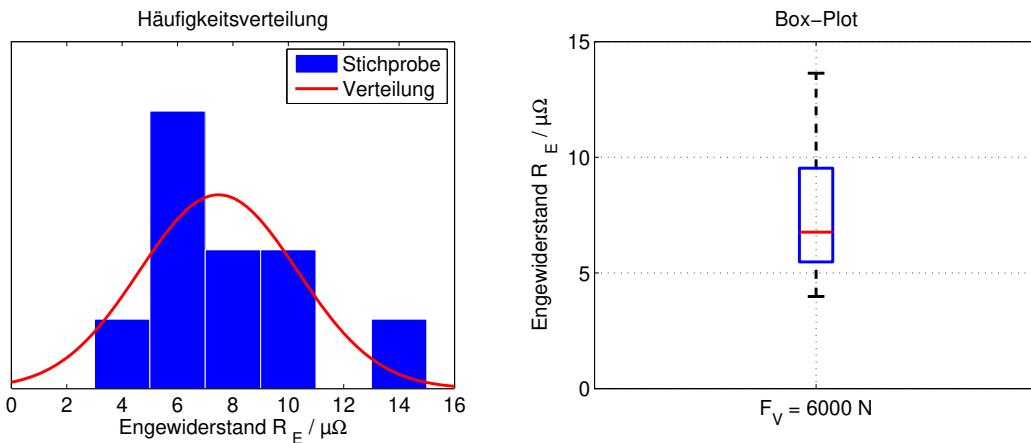


Bild 5.13: Verteilungsfunktion und Box-Plot der Stichprobe bei Vorbehandlung durch Seetransport und einer Vorspannkraft von 6000 N

Die Bestimmung der erforderlichen Vorspannkraft F_V ergibt sich aus der Prognose zukünftiger Kontaktwiderstände. Bei der Prognose sind weder Mittelwert, noch Standardabweichung bekannt. Damit ergibt sich die obere Grenze des Engewiderstandes zu

$$R_E \leq \bar{R}_E + c_2 \cdot s \cdot \sqrt{1 + \frac{1}{N}} \quad (5.160)$$

Dabei berechnet sich die Konstante c_2 über die inverse t-Verteilung mit $N - 1$ Freiheitsgraden und $\gamma = 10 \text{ ppm}$ zu

$$c_2 = F^{-1}(\gamma) = 8.1021 \quad (5.161)$$

Für den vorliegenden Fall wurden für jede Vorbehandlungsart $N = 10$ Teile untersucht. Mit den dabei aufgenommenen Messwerten ergeben sich für die Neuteile, die Teile nach Seetransport und die Teile mit oxidierten Kontakten die in Bild 5.14 dargestellten oberen Grenzen des Prognosebereiches des Kontaktwiderstandes als Funktion der Vorspannkraft.

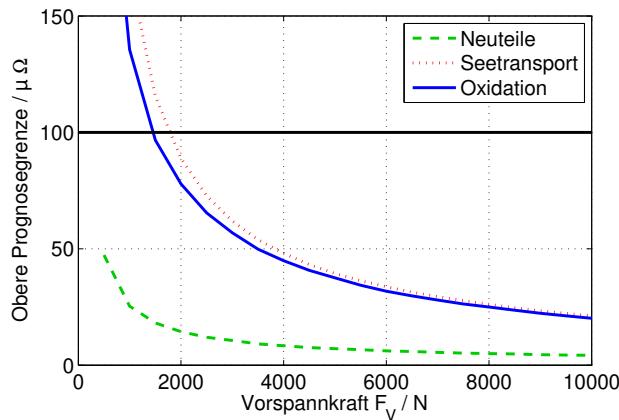


Bild 5.14: Obere Grenze des Prognosebereiches von Kontaktwiderständen auf Basis von Erprobungsergebnissen($N = 10$ Teile, $\gamma = 1 - 10$ ppm)

Es zeigt sich, dass die Kontaktwiderstände bei dem Seetransport zu den größten Prognosewerten führen. Nach diesen Erprobungsergebnissen und den beschriebenen Annahmen ist eine Vorspannkraft von $F_V = 1800$ N erforderlich, um das spezifizierte Ziel von $R_E < 100 \mu\Omega$ mit einer Sicherheit von $\gamma = 1 - 10$ ppm einzuhalten.

Die Berechnung erfolgt in folgendem MATLAB-Programm. Es geht davon aus, dass die Daten als Matrix organisiert sind, bei denen jede Zeile eine Vorspannkraft und jede Spalte ein Stichprobenwert repräsentiert.

```

1 % Messwerte einlesen
2 load Batteriesensor.mat;
3
4 % Berechnung der Konstante c2 über inverse t-Verteilung
5 % mit 9 Freiheitsgraden
6 c2 = tinv(1-1e-5,9);
7
8 % Berechnung der Prognoseintervalle nach Gleichung \eqref{eq:fivehundredfiftyfour}
9 ProgObenNeueShunts = mean(NeueShunts)' + ...
10    std(NeueShunts)' *c2*sqrt(1+0.1);
11 ProgObenSeetransport = mean(Seetransport)' + ...
12    std(Seetransport)' *c2*sqrt(1+0.1);
13 ProgObenOxidation = mean(Oxidation)' + ...
14    std(Oxidation)' *c2*sqrt(1+0.1);

```

Damit liegen die oberen Prognosegrenzen als Funktion der Vorspannkraft vor und können als Plot dargestellt werden.

```

1 % Grafische Darstellung der unterschiedlichen oberen Prognosegrenzen
2 f = figure(1)
3 plot(Kraft,ProgObenNeueShunts,'g--','Linewidth',2);
4 hold on;
5 plot(Kraft,ProgObenSeetransport,'r:','Linewidth',2);
6 plot(Kraft,ProgObenOxidation,'b','Linewidth',2);

```

7 | hold off;

Dasselbe Ergebnis ergibt sich bei der Umsetzung in Python:

```

1 % Berechnung der Prognosegrenzen
2 c2 = t.ppf(1-1e-5,9)
3 ProgObenNeueShunts = np.mean(NeueShunts, axis=1) + np.std(NeueShunts, axis
   =1)*c2*np.sqrt(1+0.1)
4 ProgObenOxidation = np.mean(Oxidation, axis=1) + np.std(Oxidation, axis=1)*
   c2*np.sqrt(1+0.1)
5 ProgObenSeetransport = np.mean(Seetransport, axis=1) + np.std(Seetransport,
   axis=1)*c2*np.sqrt(1+0.1)
6
7 % Darstellung der Verteilung und Erstellen eines Boxplot
8 fig = plt.figure(4, figsize=(6, 4))
9 ax = fig.subplots(1,1)
10 ax.plot(Kraft,ProgObenNeueShunts, 'g—', Linewidth = 2, label = 'Neue
   Shunts')
11 ax.plot(Kraft,ProgObenSeetransport, 'r:', Linewidth = 2, label = 'Seetransport')
12 ax.plot(Kraft,ProgObenOxidation, 'b', Linewidth = 2, label = 'Oxidation')\
   newline ax.axis([0, 10000, 0, 160]);
13 ax.set_xlabel('Vorspannkraft $F_V$ / N');
14 ax.set_ylabel('Prognose maximaler Engewiderstand $R_E$ / $\mu\$\Omega$');
15 ax.grid(True, which='both', axis='both', linestyle='—') ax.legend(loc='
   upper right')
```

5.7 Literatur

- [Krey91] Kreyszig, Erwin: Statistische Methoden und ihre Anwendungen
4., unveränderter Nachdruck der 7. Auflage
Vandenhoeck & Ruprecht, Göttingen, 1991
- [Fahr06] Fahrmeir, Ludwig; Künstler, Rita; Pigeot, Iris; Tutz, Gerhard: Der Weg zur Datenanalyse
6. Auflage
Springer Berlin Heidelberg New York, 2006
- [Ross06] Ross, M. Sheldon: Statistik für Ingenieure und Naturwissenschaftler
3. Auflage
Spektrum Akademischer Verlag, München, 2006
- [Papu01] Papula, Lothar: Mathematik für Ingenieure und Naturwissenschaftler Band 3
4., verbesserte Auflage
Vieweg Teubner, Braunschweig / Wiesbaden, 2008
- [Holm67] Holm, Ragnar: Electric Contacts - Theory and Application
Fourth completely rewritten edition
Springer-Verlag Berlin/Heidelberg/New York 1967
- [Höft80] Höft, Herbert: Elektrische Kontakte - Ausgewählte Beiträge
1. Auflage
Akademie-Verlag Berlin 1980

6 Motivation mit einem einführenden Beispiel

Unter einer Hypothese wird in der Statistik eine Annahme über einen Sachverhalt verstanden, der über die Verteilung einer Zufallsvariable beschrieben werden kann. Der Test einer Hypothese ist ein Prüfverfahren, das angewendet wird, um die Hypothese anzunehmen oder zu verwerfen. Hypothesentests sind damit Grundlage für Entscheidungen, die Theorie der Hypothesentests wird deshalb auch als Theorie der Entscheidungen bezeichnet.

Ähnlich wie bei der Bestimmung von Konfidenzbereichen wird auch beim Testen von Hypothesen von der Stichprobe auf die Grundgesamtheit geschlossen. Deshalb gibt es beim Hypothesentest keine vollkommen sicheren Schlüsse, eine Hypothese wird immer in Kombination mit einem definierten Fehlerrisiko getestet. Ein Hypothesentest sagt aus, ob auf Basis der vorliegenden Stichprobe eine Hypothese aufrechterhalten werden kann oder verworfen werden muss.

In der Praxis kann meist zumindest näherungsweise von einer normalverteilten Grundgesamtheit ausgegangen werden. Deshalb wird bei der folgenden Diskussion von Hypothesentests von normalverteilten Grundgesamtheiten ausgegangen.

6.1 Motivation mit einem einführenden Beispiel

Zur Kontrolle von Fertigungsprozessen werden an vielen Stellen Prozesskontrollen durchgeführt. Dadurch soll sichergestellt werden, dass die einzelnen Fertigungsschritte gemäß ihrer Spezifikation arbeiten. Signifikante Abweichungen von diesem Soll-Zustand werden durch die Prozesskontrolle erkannt und anschließend beseitigt. Als Beispiel wird das Gewicht einer Kleberaupe betrachtet. Zur Prozesskontrolle wird eine Stichprobe von $N = 5$ Teilen ausgewählt und das Gewicht der Kleberaupe vermessen. Das Gewicht aller gefertigten Teile soll als Mittelwert μ das spezifizierte Klebergewicht μ_0 aufweisen. Falls das Gewicht signifikant abweicht, muss bei den Maschineneinstellungen die Sollmenge korrigiert werden. Eine signifikante Abweichung soll auf Basis des Stichprobenmittelwertes \bar{x} erkannt werden. Die Standardabweichung σ für den Prozess ergibt sich aus der Fertigungseinrichtung, sie wird in diesem Kapitel als bekannt vorausgesetzt.

Die Aufgabe kann mit einem Hypothesentest gelöst werden, der mit den folgenden Hypothesen arbeitet:

- Nullhypothese H_0 : Mittelwert stimmt mit dem spezifizierten Wert überein, $\mu = \mu_0$
- Alternativhypothese H_1 : Mittelwert weicht signifikant von dem spezifizierten Wert ab, $\mu \neq \mu_0$

Eine starke Abweichung des Stichprobenmittelwertes \bar{x} von dem spezifizierten Sollwert μ_0 würde signalisieren, dass der Fertigungsprozess überprüft werden muss. Zur Bestimmung des Grenzwertes, bei dem die Hypothese gerade eben noch akzeptiert wird, wird davon ausgegangen, dass die Null-Hypothese H_0 gilt. In dem Fall weist der Stichprobenmittelwert eine Normalverteilung mit dem Mittelwert μ_0 und der Varianz σ^2/N auf. Die Verteilung ist in Bild 6.1 dargestellt.

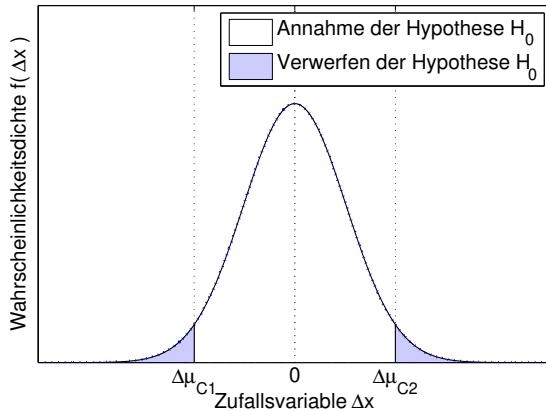


Bild 6.1: Darstellung des Hypothesentests in der Wahrscheinlichkeitsdichte der Normalverteilung, die blauen Flächen entsprechen der Wahrscheinlichkeit $P(\bar{x} < \bar{x}_1)$ und $P(\bar{x} > \bar{x}_2)$

Damit ein über die Stichprobe geschätzter Mittelwert \bar{x} mit einer spezifizierten Wahrscheinlichkeit zu der Normalverteilung mit dem Mittelwert μ_0 und der Varianz σ^2/N gehört, muss dieser in dem Intervall $\bar{x}_1 < \bar{x} \leq \bar{x}_2$ liegen. Wird die Wahrscheinlichkeit dafür mit γ bezeichnet, gilt die Gleichung

$$P(\bar{x}_1 < \bar{x} \leq \bar{x}_2) = \gamma \quad (6.1)$$

Das Intervall mit den Grenzen \bar{x}_1 und \bar{x}_2 wird als Annahmebereich für die Hypothese H_0 bezeichnet. Liegt ein geschätzter Mittelwert \bar{x} außerhalb des Intervalls $\bar{x}_1 < \bar{x} \leq \bar{x}_2$, wird die Hypothese H_0 verworfen, obwohl der geschätzte Mittelwert \bar{x} mit der Irrtumswahrscheinlichkeit

$$\alpha = 1 - \gamma \quad (6.2)$$

zu der in Bild 6.1 dargestellten Verteilung gehören kann. Die Irrtumswahrscheinlichkeit wird auch als Signifikanzniveau α des statistischen Tests bezeichnet und zur Berechnung der Grenzen \bar{x}_1 und \bar{x}_2 herangezogen.

Das Vorgehen zur Berechnung des Annahmebereiches entspricht weitgehend dem der Berechnung eines Konfidenzbereiches. Es beruht darauf, eine Zufallsvariable mit einer bekannten Verteilung zu finden, in deren Beschreibung die Hypothese H_0 und der bekannte Parameter der Stichprobe vorkommen. Für das Beispiel des Gewichts von Kleberaupen gilt dies für die standardnormalverteilte Zufallsvariable

$$z = \frac{\bar{x} - \mu_0}{\sigma / \sqrt{N}} \quad (6.3)$$

Mit dieser Verteilung wird nach Gleichung (6.1) die Wahrscheinlichkeit γ , mit der die Variable z innerhalb des Intervalls $c_1 \dots c_2$ liegt, definiert als

$$\gamma = P(c_1 < z \leq c_2) = F(c_2) - F(c_1) \quad (6.4)$$

Bei Annahme eines symmetrischen Tests ergeben sich die Konstanten c_1 und c_2 aus den Bedingungen

$$F(c_1) = \frac{1 - \gamma}{2} = \frac{\alpha}{2} \quad (6.5)$$

und

$$F(c_2) = 1 - \frac{1 - \gamma}{2} = 1 - \frac{\alpha}{2} \quad (6.6)$$

Auflösen nach c_1 und c_2 führt zu

$$c_1 = F^{-1}\left(\frac{\alpha}{2}\right) \quad (6.7)$$

und

$$c_2 = F^{-1} \left(1 - \frac{\alpha}{2} \right) \quad (6.8)$$

Durch Umformungen von Gleichung (6.4) ergibt sich ein Ausdruck für den Annahmebereich der Nullhypothese, nämlich dass der geschätzte Mittelwert \bar{x} , mit einer spezifizierten Wahrscheinlichkeit γ , zu der Normalverteilung mit dem Mittelwert μ_0 und der Varianz σ^2/N gehört.

$$\gamma = P \left(c_1 < \frac{\bar{x} - \mu_0}{\sigma/\sqrt{N}} \leq c_2 \right) = P \left(\mu_0 + \frac{c_1 \cdot \sigma}{\sqrt{N}} < \bar{x} \leq \mu_0 + \frac{c_2 \cdot \sigma}{\sqrt{N}} \right) \quad (6.9)$$

Für das Beispiel Prozesskontrolle von Kleberaupen soll der Annahmebereich dafür berechnet werden, dass der Mittelwert aus einer Stichprobe mit $N = 5$ Teilen dem spezifizierten Mittelwert von $\mu_0 = 5.3$ g entspricht. Die Standardabweichung des Prozesses beträgt $\sigma = 0.23$ g. Für den Hypothesentest wird ein Signifikanzniveau von $\alpha = 5\%$ festgelegt, zu dem die kritischen Parameter $c_1 = -1.96$ und $c_2 = 1.96$ gehören. Damit ergibt sich der Annahmebereich der Hypothese H_0 zu

$$\mu_0 + \frac{c_1 \cdot \sigma}{\sqrt{N}} = 5.3 - \frac{1.96 \cdot 0.23}{\sqrt{5}} = 5.0984 < \bar{x} \leq 5.5016 = 5.3 + \frac{1.96 \cdot 0.23}{\sqrt{5}} = \mu_0 + \frac{c_2 \cdot \sigma}{\sqrt{N}} \quad (6.10)$$

Liegt der aus der Stichprobe berechnete Mittelwert innerhalb dieser Grenzen, wird die Nullhypothese angenommen. Bei einem Wert außerhalb des berechneten Intervalls, muss die Nullhypothese auf Basis der vorliegenden Stichprobenwerte verworfen und die Alternativhypothese angenommen werden.

6.2 Praktisches Durchführen von Hypothesentests

Aufbauend auf dem einführenden Beispiel wird im Folgenden das Vorgehen beim Hypothesentest präzisiert. Nach der Darstellung des allgemeinen Vorgehens werden Hypothesentests mit unterschiedlichen Verwerfungsbereichen diskutiert. Dabei wird zunächst immer der Mittelwert bei bekannter Varianz bewertet. Eine Verallgemeinerung findet in den Abschnitten 6.5 und 6.6 statt.

6.2.1 Allgemeines Vorgehen bei der Durchführung eines Hypothesentests

Die Durchführung eines Hypothesentest teilt sich in folgende Schritte auf:

Schritt 1: Aufgabenstellung

Ein Hypothesentest baut auf einer inhaltlich klar formulierten Aufgabenstellung mit einem quantifizierten Ziel auf.

In dem einführenden Beispiel war anhand einer Stichprobe zu prüfen, ob die normalverteilte Grundgesamtheit das spezifizierte Klebergewicht mit dem Mittelwert $\mu_0 = 5.3$ g einhält.

Schritt 2: Modellannahmen

Im zweiten Schritt werden die Modellannahmen formuliert. Dazu gehören die Unabhängigkeit der Stichprobe, die Rückführung der Aufgabenstellung auf eine bekannte Verteilung und die Frage nach bekannten Parametern.

In dem Beispiel zur Prozesskontrolle wurde eine normalverteilte Grundgesamtheit mit bekannter Varianz σ^2 angenommen. Es handelte sich damit um die Abschätzung des Mittelwertes einer Grundgesamtheit mit bekannter Varianz, weshalb diese Aufgabe auf eine Standardnormalverteilung zurückgeführt werden konnte. Je nach Aufgabenstellung ergeben sich wie bei der Bestimmung des Konfidenzbereichs andere Zufallsvariablen und Verteilungen. In den Abschnitten 6.5 und 6.6 werden einige Zufallsvariablen vorgestellt.

Schritt 3: Festlegen des Signifikanzniveaus

Im nächsten Schritt wird das Signifikanzniveau α festgelegt. Es legt die Wahrscheinlichkeit fest, mit der die Hypothese H_0 verworfen wird, obwohl sie richtig gewesen wäre. Das Signifikanzniveau ergibt sich aus der Aufgabenstellung und beträgt typischerweise 1 oder 5 %. Bei der Definition des Signifikanzniveaus ist die Kenntnis der Fehler zweiter Art notwendig, auf sie wird in Abschnitt 6.4 eingegangen.

In dem Beispiel ergab sich das Signifikanzniveau α aus der Definition der Aussagesicherheit von $\gamma = 95\%$ zu $\alpha = 1 - \gamma = 5\%$.

Schritt 4: Bestimmung des Verwerfungsbereiches

Ist die Verteilung der Prüfgröße \bar{x} bekannt, kann die Hypothese $\mu = \mu_0$ getestet werden. Als Alternative sind generell drei Varianten denkbar:

$$\mu > \mu_0 \quad (6.11)$$

Eine Alternativhypothese $\mu > \mu_0$ ergibt sich zum Beispiel bei der Überwachung von Schadstoffbelastungen. Wird ein Grenzwert deutlich unterschritten, ist das unkritisch, vielleicht sogar gewünscht. Erst bei einer Überschreitung von Grenzwerten tritt eine Schädigung von Mensch und Natur ein.

$$\mu < \mu_0 \quad (6.12)$$

Die Variante $\mu < \mu_0$ tritt zum Beispiel bei Festigkeitsuntersuchungen auf, bei denen eine zu große Festigkeit unproblematisch ist, eine Festigkeit unterhalb eines Grenzwertes jedoch direkt zum Versagen des Werkstoffes führen kann.

$$\mu \neq \mu_0 \quad (6.13)$$

Die zweiseitige Variante $\mu \neq \mu_0$ ist die am häufigsten verwendete Alternativhypothese. Sie tritt zum Beispiel bei Maßen auf, die weder zu groß noch zu klein sein dürfen, wie etwa bei dem Durchmesser einer Welle. Bei dem zweiseitigen Test werden zwei Grenzwerte μ_{C1} und μ_{C2} benötigt und der Verwerfungsbereich besteht aus zwei Teilbereichen. Die Summe der Wahrscheinlichkeiten für beide Verwerfungsbereiche entspricht dem Signifikanzniveau α .

Bild 6.2 verdeutlicht die unterschiedlichen Annahme- und Verwerfungsbereiche beim Hypothesentest.

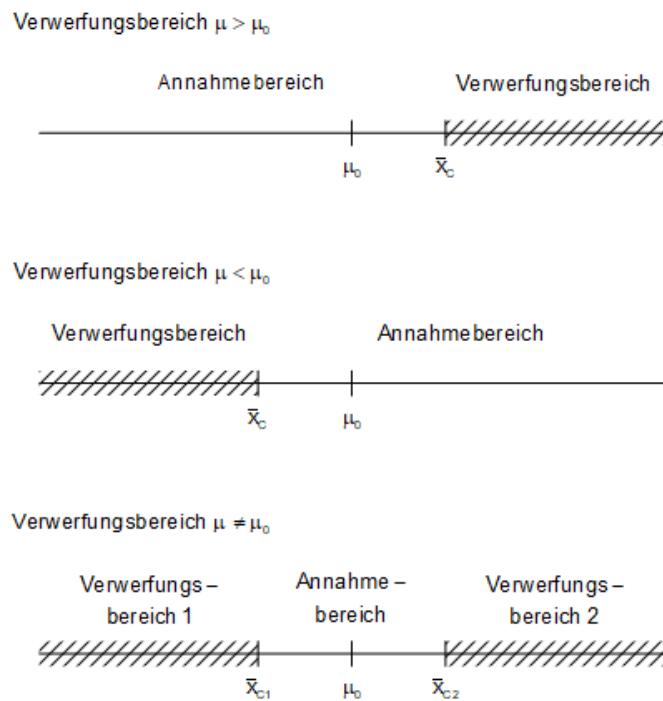


Bild 6.2: Grafische Darstellung von Verwerfungsbereichen und Annahmebereichen für einen Test mit den Alternativhypotesen $\mu > \mu_0$, $\mu < \mu_0$ und $\mu \neq \mu_0$

Nach der Festlegung des Verwerfungsbereichs und dem Signifikanzniveau können auf Basis der zu-grunde liegenden Verteilung die Grenzen μ_C beziehungsweise μ_{C1} und μ_{C2} bestimmt werden.

Im einführenden Beispiel zur Prozesskontrolle wurde ein zweiseitiger Test gewählt, da sowohl eine zu geringe, als auch eine zu große Klebermenge kritisch ist. Das Signifikanzniveau von $\alpha = 5\%$ wurde in Bild 6.1 über die blauen Flächen in der Wahrscheinlichkeitsdichte visualisiert.

Schritt 5: Vergleich der Prüfgröße mit den Grenzwerten

Für die konkret vorliegende Stichprobe wird abschließend die Prüfgröße \bar{x} berechnet und mit den Grenzwerten μ_C beziehungsweise μ_{C1} und μ_{C2} verglichen. Liegt die Prüfgröße im Annahmebereich, wird die Null-Hypothese angenommen, andernfalls verworfen.

In dem einleitenden Beispiel war die Prüfverteilung eine Standardnormalverteilung. Das hat die Berechnung vergleichsweise einfach und überschaubar gemacht. In vielen Fällen mit ausreichend großer Stichprobengröße ist es aufgrund des zentralen Grenzwertsatzes möglich, die Normalverteilung zu grunde zu legen.

6.2.2 Hypothesentest mit einseitigem Verwerfungsbereich $\mu > \mu_0$

Als Beispiel für einen Hypothesentest mit einseitigem Verwerfungsbereich $\mu > \mu_0$ wird der Kraftstoffverbrauch eines Fahrzeugtyps analysiert. In dem Beispiel liegen 40 Stichprobenwerte aus Fahrzeuguntersuchungen vor.

Tabelle 6.1: Beispiel zur Untersuchung des Kraftstoffverbrauchs eines Fahrzeugtyps

Benzinverbrauch V / l/100 km									
10.1	10.4	10.4	10.2	10.2	9.6	10.8	11.2	10.4	10.1
10.6	10.5	10.1	10.3	10.5	10.2	9.9	9.8	9	11.4
10.9	9.7	10.8	10.5	9.4	9.7	10.5	10.7	10	10.4
10	10.5	9.2	9.2	10.2	10.2	10.6	10.8	10.5	10.4

Es soll die Hypothese geprüft werden, dass die Grundgesamtheit, aus der die vorliegende Stichprobe entstammt, den Mittelwert $\mu_0 = 10 \text{ l}/100 \text{ km}$ hat. Die Alternativhypothese besagt, dass er größer ist. Als Signifikanzniveau wird der Wert $\alpha = 5\%$ gewählt. Es wird davon ausgegangen, dass die Grundgesamtheit normalverteilt ist und eine Standardabweichung von $\sigma = 0.5 \text{ l}/100 \text{ km}$ aufweist.

Es liegt eine Stichprobe mit bekannter Varianz vor, von der der Mittelwert getestet werden soll. Als geeignete Testgröße wird die Variable z gewählt.

$$z = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{N}} \quad (6.14)$$

Nach der Aufgabenstellung soll geprüft werden, ob der Mittelwert μ maximal $\mu_0 = 10 \text{ l}/100 \text{ km}$ beträgt. Wird der Mittelwert unterschritten, entspricht der Verbrauch trotzdem der Spezifikation. Der Verwerfungsbereich für die Hypothese ist deshalb $\mu > \mu_0$, er ist also einseitig. Damit kann die Aufgabe mithilfe der Wahrscheinlichkeitsdichte veranschaulicht werden, die ist in Bild 6.3 dargestellt.

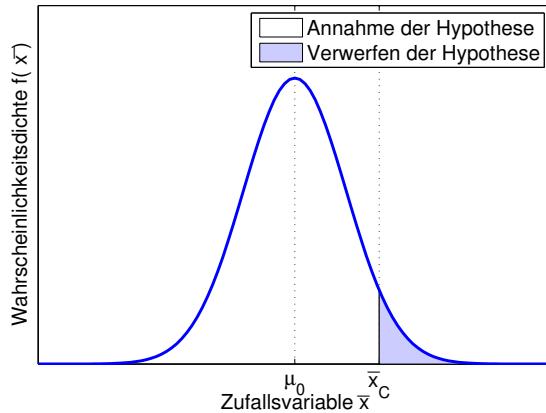


Bild 6.3: Darstellung des Hypothesentests mit einseitigem Verwerfungsbereich $\mu > \mu_0$

Damit ein Stichprobenwert \bar{x} mit einer Sicherheit von 95 % nicht zu der Normalverteilung gehört, muss er für diese Aufgabenstellung in dem Bereich der blauen Fläche unter der Kurve liegen. Die Fläche repräsentiert die Wahrscheinlichkeit von 5 %.

Mit dieser Verteilung wird die Wahrscheinlichkeit γ , mit der die Variable z im Annahmebereich liegt, definiert als

$$\gamma = 1 - \alpha = P(z < c) = F(c) \quad (6.15)$$

Bei Annahme eines einseitigen Tests mit dem Verwerfungsbereich $\mu > \mu_0$ ergibt sich die Konstante c aus der Bedingung

$$F(c) = 1 - \alpha = \gamma \quad (6.16)$$

Auflösen nach c führt zu

$$c = F^{-1}(1 - \alpha) \quad (6.17)$$

Durch Umformungen von Gleichung (6.15) ergibt sich ein Ausdruck für den Annahmebereich der Nullhypothese, nämlich dass der geschätzte Mittelwert \bar{x} , mit einer spezifizierten Wahrscheinlichkeit γ , zu der Normalverteilung mit dem Mittelwert μ_0 und der Varianz σ^2/N gehört.

$$\gamma = P\left(\frac{\bar{x} - \mu_0}{\sigma/\sqrt{N}} < c\right) = P\left(\bar{x} < \mu_0 + \frac{c \cdot \sigma}{\sqrt{N}}\right) \quad (6.18)$$

Die Konstante c ergibt sich zu

$$c = 1.6449 \quad (6.19)$$

Damit ergibt sich der Annahmebereich zu

$$\bar{x} < \mu_0 + \frac{c \cdot \sigma}{\sqrt{N}} = 10 + \frac{1.6449 \cdot 0.5}{\sqrt{40}} = 10.13 \quad (6.20)$$

Der Mittelwert der vorliegenden Stichprobe liegt bei $\bar{x}_0 = 10.25$ l/100 km und somit im Verwerfungsbereich des Hypothesentests.

Bei dem vorgestellten Test wird aus dem Signifikanzniveau α die Grenze μ_C für die Prüfgröße \bar{x} berechnet. Die Berechnung erfolgt über die zugrundeliegende Wahrscheinlichkeitsverteilung. Bei dieser Form des Hypothesentest kann nur ausgesagt werden, ob die Hypothese angenommen oder verworfen wird. Es fehlt eine quantitative Bewertung. Alternativ kann eine Überschreitungswahrscheinlichkeit p der Prüfgröße \bar{x}_0 bestimmt werden und mit dem Signifikanzniveau α verglichen werden. Diese Art des Tests ist standardmäßig in statistischen Software-Paketen implementiert, da sie eine quantitative

Bewertung ermöglicht. Bei Hypothesentests mit einseitigem Verwerfungsbereich $\mu > \mu_0$ muss für die Annahme der Nullhypothese die Bedingung

$$p = 1 - F(\bar{x}_0) > \alpha \quad (6.21)$$

erfüllt werden. Je größer der Wert p ist, desto sicherer wird die Hypothese H_0 bestätigt. Bild 6.4 stellt die Überschreitungswahrscheinlichkeit p grafisch dar.

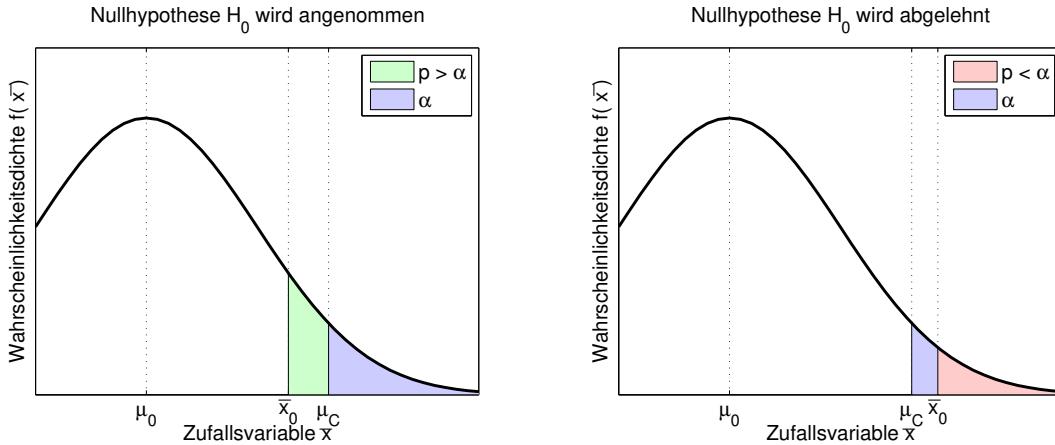


Bild 6.4: Überschreitungswahrscheinlichkeit p beim Hypothesentest mit einseitigem Verwerfungsbereich $\mu > \mu_0$

Für das Beispiel ergibt sich ein p -Wert aus der Standardnormalverteilung von

$$p = 1 - F\left(\frac{\bar{x} - \mu_0}{\sigma / \sqrt{N}}\right) = 1 - F\left(\frac{10.25 - 10}{0.5 / \sqrt{40}}\right) = 7.827 \cdot 10^{-4} < 0.05 = \alpha \quad (6.22)$$

Der Wert liegt deutlich unter der Grenze von 5 %, sodass die Nullhypothese verworfen werden muss. Der Kraftstoffverbrauch des Fahrzeugtyps liegt demnach statistisch gesehen über dem Spezifikationsbereich von 10 l/100 km.

6.2.3 Hypothesentest mit einseitigem Verwerfungsbereich $\mu < \mu_0$

Als Beispiel für einen Hypothesentest mit einseitigem Verwerfungsbereich $\mu < \mu_0$ wird die Zugfestigkeit von Folien untersucht. Dazu sind 30 Messwerte aus Zugversuchen gegeben, die in Tabelle 6.2 dargestellt sind.

Tabelle 6.2: Beispiel zur Untersuchung der Zugfestigkeit von Folien

Zugfestigkeit β_Z / N/cm ²					
44.00	44.50	44.50	44.40	42.50	40.80
43.30	43.50	46.70	44.90	43.90	43.70
42.90	44.10	42.00	44.00	44.80	43.50
43.80	43.80	42.80	44.30	41.10	45.80
43.20	44.20	42.50	46.10	44.50	43.20

Es soll die Hypothese geprüft werden, dass die Grundgesamtheit, aus der die Stichprobe von Messwerten genommen wurde, einen Mittelwert $\mu_0 \geq 44$ N/cm² hat. Es wird davon ausgegangen, dass die Grundgesamtheit normalverteilt ist und eine Standardabweichung von $\sigma = 1.29$ N/cm² aufweist. Als Signifikanzniveau wird der Wert $\alpha = 5$ % gefordert.

Es liegt eine Stichprobe mit bekannter Varianz vor, von der der Mittelwert getestet werden soll. Als geeignete Testgröße wird wieder die Variable z gewählt.

$$z = \frac{\bar{x} - \mu_0}{\sigma / \sqrt{N}} \quad (6.23)$$

Nach der Aufgabenstellung soll geprüft werden, ob der Mittelwert μ mindestens $\mu_0 = 44 \text{ N/cm}^2$ beträgt. Wird der Mittelwert überschritten, entspricht die Folie trotzdem der Spezifikation. Der Verwerfungsbereich für die Hypothese ist deshalb $\mu < \mu_0$, er ist also einseitig.

Wieder wird die Aufgabe mithilfe der Wahrscheinlichkeitsdichte veranschaulicht, sie ist in Bild 6.5 dargestellt.

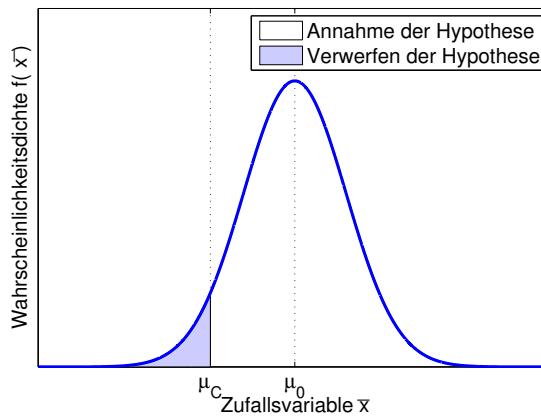


Bild 6.5: Darstellung des Hypothesentests mit einseitigem Verwerfungsbereich $\mu < \mu_0$

Damit ein Stichprobenwert \bar{x} mit einer Sicherheit von 95 % nicht zu der Normalverteilung gehört, muss er für diese Aufgabenstellung in dem Bereich der blauen Fläche unter der Kurve liegen. Die Fläche repräsentiert die Wahrscheinlichkeit von 5 %.

Mit dieser Verteilung wird wieder die Wahrscheinlichkeit γ , mit der die Variable z im Annahmebereich liegt, definiert als

$$\gamma = 1 - \alpha = P(c < z) = 1 - F(c) \quad (6.24)$$

Bei Annahme eines einseitigen Tests mit dem Verwerfungsbereich $\mu < \mu_0$ ergibt sich die Konstante c aus der Bedingung

$$F(c) = 1 - \gamma = \alpha \quad (6.25)$$

Auflösen nach c führt zu

$$c = F^{-1}(\alpha) \quad (6.26)$$

Durch Umformungen von Gleichung (6.24) ergibt sich ein Ausdruck für den Annahmebereich der Nullhypothese, dass der geschätzter Mittelwert \bar{x} mit einer spezifizierten Wahrscheinlichkeit γ zu der Normalverteilung mit dem Mittelwert μ_0 und der Varianz σ^2/N gehört.

$$\gamma = P\left(c < \frac{\bar{x} - \mu_0}{\sigma / \sqrt{N}}\right) = P\left(\mu_0 + \frac{c \cdot \sigma}{\sqrt{N}} < \bar{x}\right) \quad (6.27)$$

Die Konstante c ergibt sich zu

$$c = -1.6449 \quad (6.28)$$

Damit lautet der Annahmebereich

$$\bar{x} > \mu_0 + \frac{c \cdot \sigma}{\sqrt{N}} = 44 - \frac{1.6449 \cdot 1.29}{\sqrt{30}} = 43.6126 \quad (6.29)$$

Der Mittelwert der vorliegenden Stichprobe liegt bei $\bar{x}_0 = 43.78 \text{ N/cm}^2$ und somit im Annahmebereich des Hypothesentests.

Alternativ kann wie im Abschnitt zuvor eine Unterschreitungswahrscheinlichkeit p der Prüfgröße \bar{x} bestimmt werden und mit dem Signifikanzniveau α verglichen werden. Diese Art des Tests ist wie bereits erwähnt standardmäßig in statistischen Software-Paketen implementiert, da sie eine quantitative Bewertung ermöglicht. Bei Hypothesentests mit einseitigem Verwerfungsbereich $\mu < \mu_0$ muss für die Annahme der Nullhypothese die Bedingung

$$p = F(\bar{x}_0) > \alpha \quad (6.30)$$

erfüllt werden. Je größer der Wert p ist, desto sicherer wird die Hypothese H_0 bestätigt. Bild 6.6 stellt die Überschreitungswahrscheinlichkeit p grafisch dar.

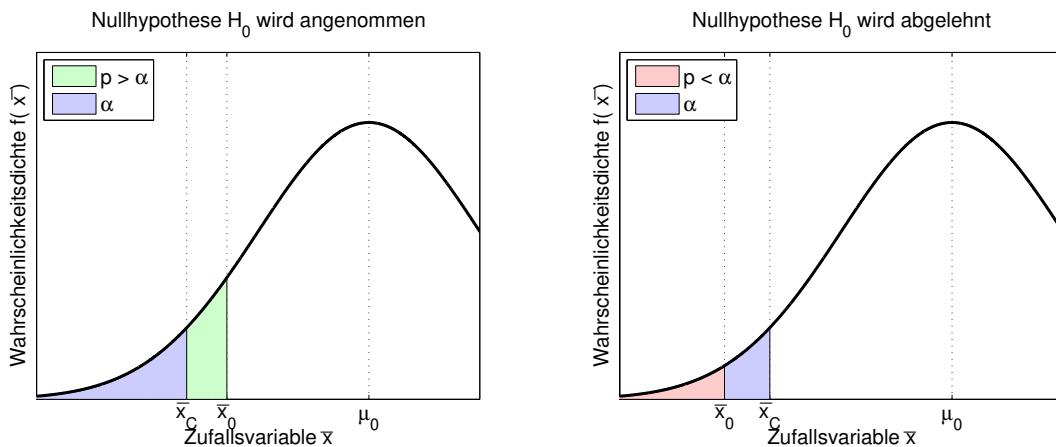


Bild 6.6: Überschreitungswahrscheinlichkeit p beim Hypothesentest mit einseitigem Verwerfungsbereich $\mu < \mu_0$

Für das Beispiel ergibt sich ein p -Wert von

$$p = F\left(\frac{\bar{x}_0 - \mu_0}{\sigma / \sqrt{N}}\right) = F\left(\frac{43.78 - 44}{1.29 / \sqrt{30}}\right) = 0.1751 > 0.05 = \alpha \quad (6.31)$$

Die Bedingung für eine Annahme der Nullhypothese ist damit erfüllt.

6.2.4 Hypothesentest mit zweiseitigem Verwerfungsbereich $\mu \neq \mu_0$

Der Fall eines Hypothesentests mit zweiseitigem Verwerfungsbereich $\mu \neq \mu_0$ wird an dem Beispiel von Drehteilen diskutiert. Dazu sind 20 Drehteile vermessen worden, deren Durchmesser in Tabelle 6.3 dargestellt sind.

Tabelle 6.3: Beispiel zur Untersuchung der Maßhaltigkeit von Drehteilen

Durchmesser Drehteile d / mm				
50.89	51.01	50.94	50.84	50.27
50.36	50.09	50.80	50.47	50.94
50.69	50.70	50.57	50.06	50.51
50.66	51.07	50.50	50.48	50.64

Es soll die Hypothese geprüft werden, dass die Grundgesamtheit, aus der die Stichprobe von Fertigungsteilen genommen wurde, einen Mittelwert $\mu_0 = 50.55$ mm hat. Die Gegenhypothese ist, dass der Mittelwert signifikant davon abweicht. Es wird davon ausgegangen, dass die Grundgesamtheit normalverteilt ist und eine Standardabweichung von $\sigma = 0.29$ mm aufweist. Als Signifikanzniveau wird der Wert $\alpha = 5\%$ gefordert.

Es liegt wieder eine Stichprobe mit bekannter Varianz vor, von der der Mittelwert getestet werden soll. Als geeignete Testgröße wird die Variable z gewählt.

$$z = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{N}} \quad (6.32)$$

Nach der Aufgabenstellung soll geprüft werden, ob der Mittelwert μ genau $\mu_0 = 50.55$ mm beträgt. Der Verwerfungsbereich für die Hypothese ist $\mu \neq \mu_0$, er ist also zweiseitig. Auch in diesem Fall kann die Aufgabe mithilfe der Wahrscheinlichkeitsdichte veranschaulicht werden, die ist in Bild 6.7 dargestellt.

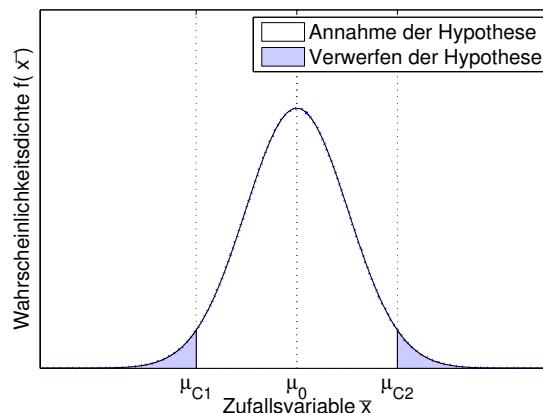


Bild 6.7: Darstellung des Hypothesentests mit zweiseitigem Verwerfungsbereich $\mu \neq \mu_0$

Damit ein Stichprobenwert \bar{x} mit einer Sicherheit von 95 % nicht zu der Normalverteilung mit dem Mittelwert μ_0 und der Varianz σ^2/N gehört, muss er für diese Aufgabenstellung in dem Bereich der blauen Fläche unter der Kurve liegen. Die Fläche repräsentiert insgesamt die Wahrscheinlichkeit von 5 %, ist in diesem Fall aber in zwei Bereiche von jeweils 2.5 % aufgeteilt.

Mit dieser Verteilung wird nach Gleichung (6.1) die Wahrscheinlichkeit γ , mit der die Variable z im Annahmebereich liegt, definiert als

$$\gamma = 1 - \alpha = P(c_1 < z \leq c_2) = F(c_2) - F(c_1) \quad (6.33)$$

Bei Annahme eines zweiseitigen Tests mit dem Verwerfungsbereich $\mu \neq \mu_0$ ergeben sich die Konstanten c_1 und c_2 aus den Bedingungen

$$F(c_1) = \frac{1 - \gamma}{2} = \frac{\alpha}{2} \quad (6.34)$$

$$F(c_2) = 1 - \frac{1 - \gamma}{2} = 1 - \frac{\alpha}{2} \quad (6.35)$$

Auflösen nach den Konstanten c_1 und c_2 führt zu

$$c_1 = F^{-1}\left(\frac{\alpha}{2}\right) \quad (6.36)$$

$$c_2 = F^{-1}\left(1 - \frac{\alpha}{2}\right) \quad (6.37)$$

Durch Umformungen von Gleichung (6.33) ergibt sich ein Ausdruck für den Annahmebereich der Nullhypothese, dass der geschätzter Mittelwert \bar{x} mit einer spezifizierten Wahrscheinlichkeit γ zu der Normalverteilung mit dem Mittelwert μ_0 und der Varianz σ^2/N gehört.

$$\gamma = P\left(c_1 < \frac{\bar{x} - \mu_0}{\sigma/\sqrt{N}} \leq c_2\right) = P\left(\mu_0 + \frac{c_1 \cdot \sigma}{\sqrt{N}} < \bar{x} \leq \mu_0 + \frac{c_2 \cdot \sigma}{\sqrt{N}}\right) \quad (6.38)$$

Die Konstante c_1 und c_2 ergeben sich zu

$$c_1 = -1.96 \quad (6.39)$$

$$c_2 = 1.96 \quad (6.40)$$

Damit ergibt sich der Annahmebereich in diesem Beispiel zu

$$\mu_0 + \frac{c_1 \cdot \sigma}{\sqrt{N}} = 50.55 - \frac{1.96 \cdot 0.29}{\sqrt{20}} = 50.42 < \bar{x} \leq 50.68 = 50.55 - \frac{1.96 \cdot 0.29}{\sqrt{20}} = \mu_0 + \frac{c_2 \cdot \sigma}{\sqrt{N}} \quad (6.41)$$

Der Mittelwert der vorliegenden Stichprobe liegt bei $\bar{x}_0 = 50.62$ mm und somit im Annahmebereich des Hypothesentests.

Alternativ kann wie im Abschnitt zuvor eine Unterschreitungswahrscheinlichkeit p der Prüfgröße \bar{x}_0 bestimmt werden und mit dem Signifikanzniveau α verglichen werden. Bei Hypothesentests mit beidseitigem Verwerfungsbereich $\mu \neq \mu_0$ müssen die Bedingungen

$$p = F(\bar{x}_0) > \frac{\alpha}{2} \quad (6.42)$$

und

$$p = F(\bar{x}_0) < 1 - \frac{\alpha}{2} \quad (6.43)$$

erfüllt werden. Je zentraler der Wert p zwischen den Grenzen $\alpha/2$ und $1 - \alpha/2$ liegt, desto sicherer wird die Hypothese H_0 bestätigt. Bild 6.8 stellt die Überschreitungswahrscheinlichkeit p mit den unterschiedlichen Annahme- und Verwerfungsszenarien grafisch dar.

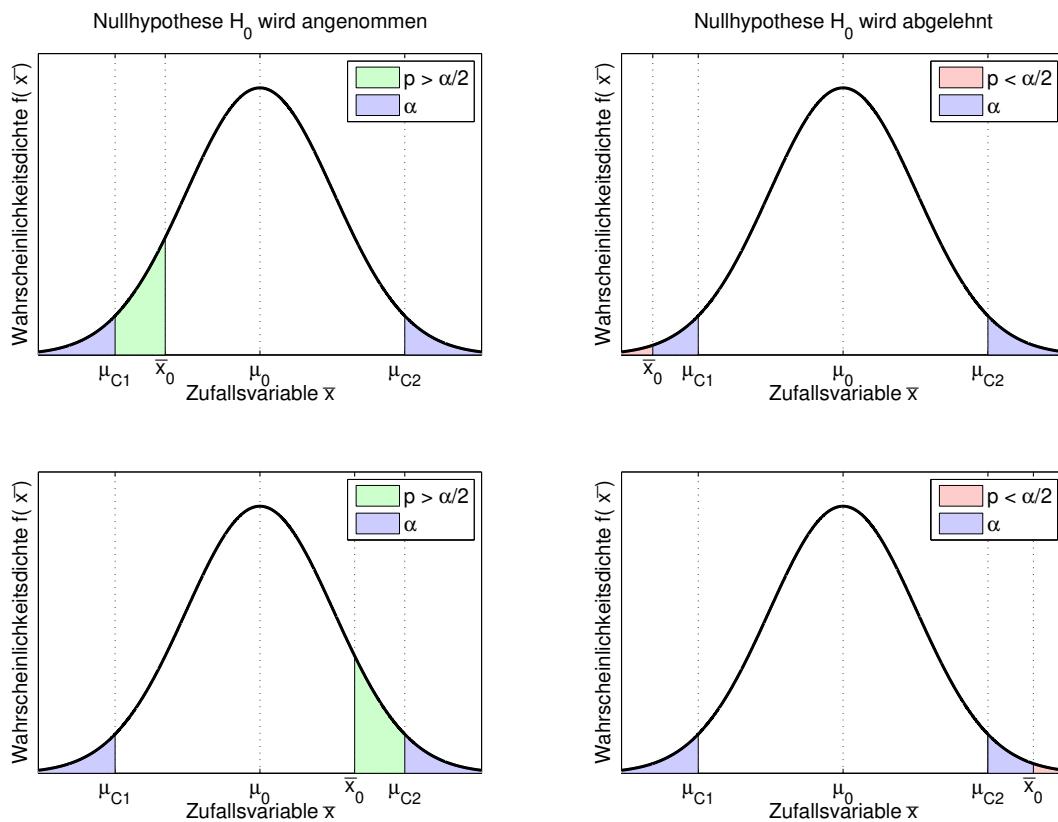


Bild 6.8: Überschreitungswahrscheinlichkeit p beim Hypothesentest mit zweiseitigem Verwerfungsbereich $\mu \neq \mu_0$

Mit dem Mittelwert der vorliegenden Stichprobe von $\bar{x}_0 = 50.62$ mm ergibt sich ein p -Wert von

$$p = F\left(\frac{\bar{x}_0 - \mu_0}{\sigma/\sqrt{N}}\right) = F\left(\frac{50.62 - 50.55}{0.29/\sqrt{20}}\right) = 0.8598 \quad (6.44)$$

Damit wird die Hypothese, dass die Grundgesamtheit einen Durchmesser von 50.55 mm besitzt, angenommen.

6.3 Hypothesentest und Konfidenzbereich

Bei einigen statistischen Verfahren wird statt des Hypothesentests eine Analyse des Konfidenzbereichs durchgeführt. Das Vorgehen zur Berechnung der Konfidenzbereiche und das Vorgehen zum Hypothesentest entsprechen einander. Beide Verfahren werden im Folgenden am Beispiel des Mittelwertes einer Grundgesamtheit mit bekannter Varianz miteinander verglichen.

6.3.1 Hypothesentest

Beim Hypothesentest wird geprüft, ob der Mittelwert einer vorliegenden Stichprobe mit der Wahrscheinlichkeit γ in dem Annahmebereich mit den Grenzen μ_{C1} und μ_{C2} liegt.

$$\gamma = 1 - \alpha = P(\mu_{C1} < \bar{x} \leq \mu_{C2}) \quad (6.45)$$

Die Berechnung der Grenzen μ_{C1} und μ_{C2} beruht darauf, eine Zufallsvariable mit einer bekannten Verteilung zu finden, in deren Beschreibung die Hypothese H_0 und der bekannte Parameter der Stichprobe vorkommen. Für das Beispiel des Mittelwertes gilt dies für die standardnormalverteilte Zufallsvariable

$$z = \frac{\bar{x} - \mu_0}{\sigma / \sqrt{N}} \quad (6.46)$$

Mit dieser Verteilung wird nach Gleichung (6.45) die Wahrscheinlichkeit γ , mit der die Variable z innerhalb des Intervalls $c_1 \dots c_2$ liegt, definiert als

$$\gamma = P(c_1 < z \leq c_2) = F(c_2) - F(c_1) \quad (6.47)$$

Durch Umformungen von Gleichung (6.47) ergibt sich ein Ausdruck für den Annahmebereich der Nullhypothese, dass der vorliegende Mittelwert \bar{x}_0 mit einer spezifizierten Wahrscheinlichkeit γ zu der Normalverteilung mit dem Mittelwert μ_0 und der Varianz σ^2/N gehört.

$$\gamma = P\left(c_1 < \frac{\bar{x}_0 - \mu_0}{\sigma / \sqrt{N}} \leq c_2\right) = P\left(\mu_0 + \frac{c_1 \cdot \sigma}{\sqrt{N}} < \bar{x}_0 \leq \mu_0 + \frac{c_2 \cdot \sigma}{\sqrt{N}}\right) = P(\mu_{C1} < \bar{x}_0 \leq \mu_{C2}) \quad (6.48)$$

Liegt er innerhalb des Annahmebereiches, wird die Hypothese angenommen, andernfalls abgelehnt.

6.3.2 Konfidenzbereich

Beim Konfidenzbereich wird auf Grundlage einer vorliegenden Stichprobe berechnet, in welchem Intervall der Mittelwert μ der Grundgesamtheit liegt.

$$\gamma = 1 - \alpha = P(\mu_{C1} < \mu \leq \mu_{C2}) \quad (6.49)$$

Die Berechnung der Grenzen μ_{C1} und μ_{C2} beruht darauf, eine Zufallsvariable mit einer bekannten Verteilung zu finden, in deren Beschreibung der Parameter der Grundgesamtheit und der bekannte Parameter der Stichprobe vorkommen. Für das Beispiel des Mittelwertes gilt dies für die standardnormalverteilte Zufallsvariable

$$z = \frac{\bar{x} - \mu}{\sigma / \sqrt{N}} \quad (6.50)$$

Mit dieser Verteilung wird nach Gleichung (6.49) die Wahrscheinlichkeit γ , mit der die Variable z innerhalb des Intervalls $c_1 \dots c_2$ liegt, definiert als

$$\gamma = P(c_1 < z \leq c_2) = F(c_2) - F(c_1) \quad (6.51)$$

Durch Umformungen von Gleichung (6.51) ergibt sich ein Ausdruck für Konfidenzbereich des Mittelwertes der Grundgesamtheit.

$$\gamma = P\left(c_1 < \frac{\bar{x} - \mu}{\sigma / \sqrt{N}} \leq c_2\right) = P\left(\bar{x} - \frac{c_2 \cdot \sigma}{\sqrt{N}} < \mu \leq \bar{x} - \frac{c_1 \cdot \sigma}{\sqrt{N}}\right) = P(\mu_{C1} < \mu \leq \mu_{C2}) \quad (6.52)$$

In dem in Gleichung (6.52) definierten Intervall liegt der Mittelwert der Grundgesamtheit mit einer spezifizierten Wahrscheinlichkeit γ .

6.3.3 Vergleich der beiden Verfahren

Beide statistischen Verfahren erfolgen auf derselben Basis, werden aber für unterschiedliche Größen durchgeführt. Während bei der Bestimmung des Konfidenzbereiches der Mittelwert der Grundgesamtheit betrachtet wird, wird beim Hypothesentest der Mittelwert der Stichprobe analysiert. Ein Hypothesentest kann in die Betrachtung eines Konfidenzbereiches überführt werden. Um dies zu verdeutlichen, wird in Gleichung (6.53) ausgehend vom Konfidenzbereich des Mittelwertes der Grundgesamtheit der Annahmebereich des Stichprobenmittelwerts bestimmt.

$$\gamma = P\left(\bar{x} - \frac{c_2 \cdot \sigma}{\sqrt{N}} < \mu \leq \bar{x} + \frac{c_1 \cdot \sigma}{\sqrt{N}}\right) = P\left(-\frac{c_2 \cdot \sigma}{\sqrt{N}} < \mu - \bar{x} \leq \frac{c_1 \cdot \sigma}{\sqrt{N}}\right) = P\left(\mu + \frac{c_1 \cdot \sigma}{\sqrt{N}} < \bar{x} \leq \mu + \frac{c_2 \cdot \sigma}{\sqrt{N}}\right) \quad (6.53)$$

Beide Verfahren führen aber nur zu derselben mathematischen Gleichung, wenn der Hypothesentest ein zweiseitiger Test mit der Alternativhypothese $\mu_1 \neq \mu_0$ ist. Der Konfidenzbereich deckt damit nur einen Teil der Möglichkeiten eines Hypothesentests ab.

Andererseits erlaubt er unter diesen Bedingungen eine einfache Interpretation: Schließt das Konfidenzintervall des Mittelwerts der Grundgesamtheit den Stichprobenwert \bar{x}_0 ein, ist der Mittelwert μ der Grundgesamtheit nicht signifikant von \bar{x}_0 verschieden.

6.4 Sicherheit bei Hypothesentests

Der Hypothesentest basiert auf einer Stichprobe und ist deshalb nicht sicher. Es kann zu Fehlentscheidungen kommen. Zur Bewertung der Aussagesicherheit wird an einem übersichtlichen Paar von Nullhypothese und Alternativhypothese die Definition von Fehlern erster und zweiter Art eingeführt. Darauf aufbauend wird die Gütfunktion eines Hypothesentests bestimmt und der notwendige Stichprobenumfang errechnet, der für eine geforderte Aussagesicherheit notwendig ist.

6.4.1 Fehler erster und zweiter Art

Zur Einführung der Fehler erster und zweiter Art wird wieder der Mittelwert einer Grundgesamtheit herangezogen. Ausgehend von der Nullhypothese

$$\mu = \mu_0 \quad (6.54)$$

wird zunächst gegen die Alternativhypothese

$$\mu = \mu_1 \quad (6.55)$$

getestet, wobei der Wert μ_1 größer ist als der Wert μ_0 . Bild 6.9 stellt die Situation mit zwei Wahrscheinlichkeitsdichten gleicher Varianz aber unterschiedlichen Mittelwerten μ_0 und μ_1 dar. Es existiert eine kritische Grenze μ_C , die zwischen den Werten μ_0 und μ_1 liegt. Aus der vorliegenden Stichprobe x_1, x_2, \dots, x_N wird ein Schätzwert für den Mittelwert berechnet.

$$\bar{x} = \frac{1}{N} \cdot \sum_{n=1}^N x_n \quad (6.56)$$

Ist der berechnete Stichprobenmittelwert \bar{x} größer als die Grenze μ_C , wird die Nullhypothese verworfen. Liegt der berechnete Stichprobenmittelwert unterhalb der kritischen Grenze μ_C , wird die Nullhypothese angenommen.

Bei dem Hypothesentest können zwei Arten von Fehlern auftreten, die als Fehler erster und zweiter Art bezeichnet werden. Bild 6.9 verdeutlicht diese Zusammenhänge grafisch.

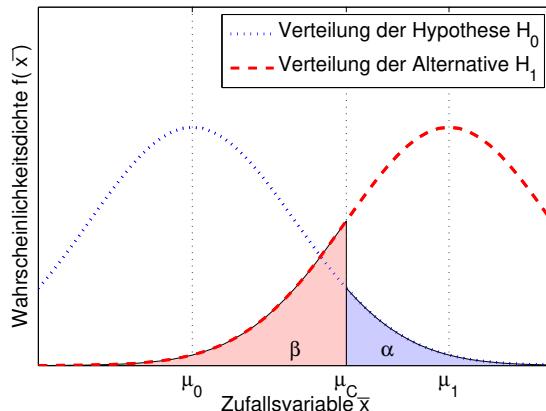


Bild 6.9: Darstellung des Fehlers erster und zweiter Art beim Hypothesentest

Fehler erster Art

Beim Fehler erster Art wird die Nullhypothese verworfen, obwohl sie richtig ist. Die Wahrscheinlichkeit, einen solchen Fehler zu begehen, entspricht der Irrtumswahrscheinlichkeit α , die auch als Signifikanzniveau des Tests bezeichnet wird. Ein solcher Fehler wird begangen, wenn die Hypothese richtig ist und der berechnete Stichprobenwert \bar{x} trotzdem einen Wert annimmt, der oberhalb der kritischen Grenze μ_C liegt. In diesem Fall gilt die bedingte Wahrscheinlichkeit

$$P(\bar{x} > \mu_C | \mu = \mu_0) = \alpha \quad (6.57)$$

In dem einführenden Beispiel zur Prozesskontrolle ist ein Fehler erster Art, dass der Stichprobenmittelwert \bar{x} den Grenzwert μ_{C2} überschreitet, obwohl die Grundgesamtheit den richtigen Mittelwert μ_0 besitzt.

Fehler zweiter Art

Beim Fehler zweiter Art wird die Nullhypothese angenommen, obwohl diese falsch ist. Die zugehörige Wahrscheinlichkeit wird mit β bezeichnet. Ein Fehler zweiter Art wird begangen, wenn die Nullhypothese falsch ist, der berechnete Stichprobenwert aber dennoch einen Wert annimmt, der unterhalb der kritischen Grenze μ_C liegt. In diesem Fall gilt

$$P(\bar{x} \leq \mu_C | \mu = \mu_1) = \beta \quad (6.58)$$

Der Wert $(1 - \beta)$ ist die Wahrscheinlichkeit, einen Fehler zweiter Art zu vermeiden. Der Wert wird als Güte des Hypothesentests bezeichnet.

Diskussion von Fehlern erster und zweiter Art

Tabelle 6.4 stellt die Situation von Annahme und Ablehnung einer Nullhypothese beim Hypothesentest und die damit verbundenen Fehler tabellarisch zusammen.

Tabelle 6.4: Übersicht über richtige und falsche Entscheidungen beim Hypothesentest mit der entsprechenden Wahrscheinlichkeitsangabe

		Unbekannte Wirklichkeit	
		$\mu = \mu_0$	$\mu = \mu_1$
Angenommen	$\mu = \mu_0$	richtige Entscheidung $p = 1 - \alpha$	Fehler 2. Art $p = \beta$
	$\mu = \mu_1$	Fehler 1. Art $p = \alpha$	richtige Entscheidung $p = 1 - \beta$

Die Wahl des Parameters μ_C bestimmt die Wahrscheinlichkeit der Fehlentscheidung. Der Parameter μ_C sollte daher so gewählt werden, dass die Fehlerwahrscheinlichkeiten α und β möglichst klein werden. Bild 6.9 zeigt, dass diese Forderungen sich gegenseitig widersprechen. Um α zu minimieren, muss die kritische Grenze μ_C nach rechts verschoben werden. Dann wird aber die Fehlerwahrscheinlichkeit β größer. Bei der praktischen Durchführung von Hypothesentests wird zunächst das Signifikanzniveau α festgelegt. Daraus ergibt sich die Grenze des Annahmebereiches μ_C und mit dem Parameter μ_C wird die Fehlerwahrscheinlichkeit β des Fehlers zweiter Art berechnet.

Die Annahme einer Hypothese ist stets von der vorliegenden Stichprobe und dem gewählten Signifikanzniveau α abhängig. Aus der Annahme einer Hypothese auf Basis eines Hypothesentests folgt daher nicht, dass die Hypothese die einzige mögliche ist. Ein Hypothesentest sagt lediglich aus, ob auf Basis

der vorliegenden Stichprobe und dem gewählten Signifikanzniveau α eine Hypothese aufrechterhalten werden kann oder verworfen werden muss. Oftmals ist auch die Fehlerwahrscheinlichkeit β des Hypothesentests nicht bekannt, sodass weitere Bewertungen der Entscheidung nicht möglich sind. Das Risiko erster Art kann dann zwar klein gewählt werden, dadurch steigt jedoch das Risiko zweiter Art.

6.4.2 Gütfunktion eines Hypothesentests

Die Gütfunktion eines Hypothesentests erlaubt Aussagen über die Qualität des statistischen Tests zu machen. Sie wird deshalb zum Vergleich unterschiedlicher Tests zu einem Testproblem herangezogen. Falls mehrere konkurrierende Tests existieren, wird der Test ausgewählt, der bei gleichem Stichprobenumfang die größte Güte beziehungsweise die größte Trennschärfe besitzt. Für die bekannten Alternativhypotesen

$$\mu_1 > \mu_0 \quad (6.59)$$

$$\mu_1 < \mu_0 \quad (6.60)$$

$$\mu_1 \neq \mu_0 \quad (6.61)$$

wird die Güte $(1 - \beta)$ als Funktion der alternativen Prüfgröße μ_1 beschrieben. Es wird deshalb nicht mehr von der Güte, sondern von einer Gütfunktion gesprochen.

Die Gütfunktion eines Hypothesentests wird an einem Beispiel mit einer normalverteilten Grundgesamtheit und bekannter Varianz $\sigma^2 = 9$ diskutiert. Unter Verwendung einer Stichprobe mit einem Umfang von $N = 10$ Messwerten und dem Stichprobenmittelwert \bar{x} soll die Nullhypothese $\mu = \mu_0 = 24$ gegen die drei Varianten der Alternativhypothese aus Gleichungen (6.59) - (6.61) getestet werden.

Für die weiteren Betrachtungen wird ein Signifikanzniveau von $\alpha = 5\%$ gewählt. Trifft die Nullhypothese zu, ist der Mittelwert der Grundgesamtheit normalverteilt mit dem Mittelwert $\mu_0 = 24$ und der Standardabweichung

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{N}} = \sqrt{0.9} \quad (6.62)$$

Für die Berechnung des kritischen Wertes μ_C wird die standardnormalverteilte Prüfgröße

$$z = \frac{\bar{x} - \mu_0}{\sigma_{\bar{x}}} = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{N}} = \frac{\bar{x} - 24}{\sqrt{0.9}} \quad (6.63)$$

herangezogen. Mit dieser Verteilung kann der Annahmebereich der Nullhypothese berechnet werden. Die unterschiedlichen Alternativhypotesen werden im Folgenden einzeln hinsichtlich ihrer Annahme- und Verwerfungsbereiche sowie der Güte des Tests diskutiert.

Alternativhypothese $\mu_1 > \mu_0$

Zunächst wird die Alternativhypothese $\mu_1 > \mu_0$ untersucht. Die Berechnung des Annahmebereiches erfolgt in diesem Fall durch die Wahrscheinlichkeit

$$P(\bar{x} < \mu_C | \mu = \mu_0) = P\left(\bar{x} < \mu_0 + \frac{c \cdot \sigma}{\sqrt{N}} | \mu = \mu_0\right) = \gamma = 1 - \alpha \quad (6.64)$$

Die Grenze c ergibt sich aus der inversen Standardnormalverteilung zu $c = 1.6449$. Damit folgt die Grenze des Annahmebereiches mit der Nullhypothese $\mu_0 = 24$ zu

$$\bar{x} < \mu_0 + \frac{c \cdot \sigma}{\sqrt{N}} = 25.56 \quad (6.65)$$

Liegt der Stichprobenmittelwert unterhalb der kritischen Grenze $\mu_C = 25.56$ wird die Nullhypothese angenommen, andernfalls wird die Nullhypothese abgelehnt. Nach Festlegung des Grenzwertes μ_C kann die Güte des Hypothesentests mit der Alternativhypothese $\mu_1 > \mu_0$ berechnet werden zu

$$1 - \beta(\mu_1) = P(\bar{x} > \mu_C | \mu = \mu_1) = 1 - P(\bar{x} < \mu_C | \mu = \mu_1) \quad (6.66)$$

Wegen der Standardisierung der Zufallsvariable \bar{x} kann die Gütfunktion mit der Standardnormalverteilung ausgedrückt werden.

$$1 - \beta(\mu_1) = 1 - P(\bar{x} < \mu_C | \mu = \mu_1) = 1 - F\left(\frac{\mu_C - \mu_1}{\sigma_{\bar{x}}}\right) = 1 - F\left(\frac{25.56 - \mu_1}{\sqrt{0.9}}\right) \quad (6.67)$$

Die Güte ist davon abhängig, wie groß der Alternativmittelwert μ_1 ist. Die entsprechende Gütfunktion ist in Bild 6.10 dargestellt.

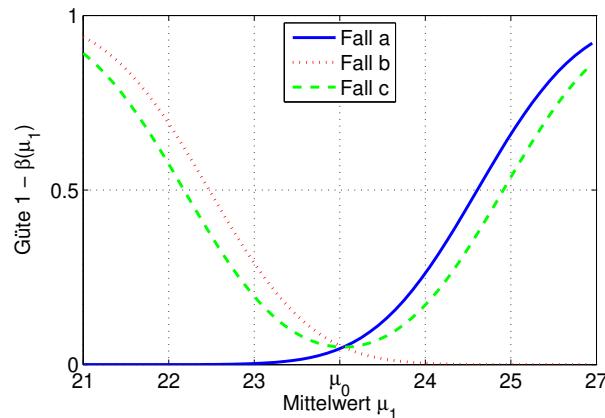


Bild 6.10: Darstellung der Gütfunktion für die Alternativhypothese $\mu_1 > \mu_0$

Je größer der Wert μ_1 ist, desto sicherer ist die Aussage, dass die Mittelwerte voneinander abweichen. Für $\mu_1 = \mu_0$ hat die Güte den Wert des Signifikanzniveaus α . Angenommen der wahre Mittelwert liegt bei $\mu_1 = 26$, dann ist die Güte dieses Tests ca. 70 %. Ein Mittelwert $\mu_1 = 26$ wird demnach nur mit einer Wahrscheinlichkeit von 70 % als Fehler erkannt.

Alternativhypothese $\mu_1 < \mu_0$

Für die Alternativhypothese $\mu_1 < \mu_0$ wird der Annahmebereich mit der Wahrscheinlichkeit

$$P(\bar{x} > \mu_C | \mu = \mu_0) = P\left(\bar{x} > \mu_0 + \frac{c \cdot \sigma}{\sqrt{N}} | \mu = \mu_0\right) = \gamma = 1 - \alpha \quad (6.68)$$

berechnet. Die Grenze c ergibt sich dabei aus der inversen Standardnormalverteilung zu $c = -1.6449$. Mit diesem Wert berechnet sich der Annahmebereich bei der Alternativhypothese $\mu_1 < \mu_0$ mit der Nullhypothese $\mu_0 = 24$ zu

$$\bar{x} > \mu_0 + \frac{c \cdot \sigma}{\sqrt{N}} = 22.44 \quad (6.69)$$

Liegt der Stichprobenmittelwert oberhalb der kritischen Grenze $\mu_C = 22.44$, wird die Nullhypothese angenommen, andernfalls wird die Nullhypothese abgelehnt.

Nach Festlegung des Grenzwertes μ_C kann die Güte des Hypothesentests mit der Alternativhypothese $\mu_1 < \mu_0$ berechnet werden zu

$$1 - \beta(\mu_1) = P(\bar{x} < \mu_C | \mu = \mu_1) = F\left(\frac{\mu_C - \mu_1}{\sigma_{\bar{x}}}\right) = F\left(\frac{22.44 - \mu_1}{\sqrt{0.9}}\right) \quad (6.70)$$

Wieder ist die Güte davon abhängig, wie groß der Alternativmittelwert μ_1 ist. Die Gütfunktion ist in Bild 6.11 dargestellt.

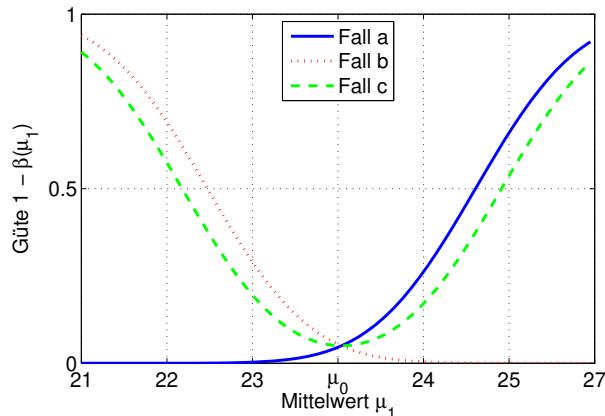


Bild 6.11: Darstellung der Gütfunktion für die Alternativhypothese $\mu_1 < \mu_0$

Je kleiner der Wert μ_1 ist, desto sicherer ist die Aussage, dass die Mittelwerte voneinander abweichen. Für $\mu_1 = \mu_0$ hat die Güte den Wert des Signifikanzniveaus α . Angenommen der wahre Mittelwert liegt bei $\mu_1 = 22$, dann ist die Güte dieses Tests ca. 70 %. Ein Mittelwert $\mu_1 = 22$ wird demnach nur mit einer Wahrscheinlichkeit von ca. 70 % als Fehler erkannt.

Alternativhypothese $\mu_1 \neq \mu_0$

Die Alternativhypothese $\mu_1 \neq \mu_0$ führt zu einem zweiseitigen Verwerfungsbereich der Nullhypothese. Der Annahmebereich ergibt sich aus der Definition der Annahmewahrscheinlichkeit

$$P(\mu_{C1} < \bar{x} \leq \mu_{C2} | \mu = \mu_0) = P\left(\mu_0 + \frac{c_1 \cdot \sigma}{\sqrt{N}} < \bar{x} \leq \mu_0 + \frac{c_2 \cdot \sigma}{\sqrt{N}} | \mu = \mu_0\right) = \gamma = 1 - \alpha \quad (6.71)$$

Damit ergibt sich der Annahmebereich in diesem Beispiel zu

$$22.14 = \mu_0 + \frac{c_1 \cdot \sigma}{\sqrt{N}} < \bar{x} \leq \mu_0 + \frac{c_2 \cdot \sigma}{\sqrt{N}} = 25.86 \quad (6.72)$$

Liegt der Stichprobenmittelwert innerhalb der Grenzen $\mu_{C1} = 22.14$ und $\mu_{C2} = 25.86$ wird die Nullhypothese angenommen, andernfalls wird die Nullhypothese abgelehnt.

Bei der Berechnung der Güte müssen bei der Alternativhypothese $\mu_1 \neq \mu_0$ zwei Bereiche berücksichtigt werden. Damit ergibt sich bei einem alternativen Mittelwert μ_1 die Gütfunktion

$$\begin{aligned} 1 - \beta(\mu_1) &= P(\bar{x} < \mu_{C1} | \mu = \mu_1) + P(\bar{x} > \mu_{C2} | \mu = \mu_1) \\ &= P(\bar{x} < \mu_{C1} | \mu = \mu_1) + 1 - P(\bar{x} < \mu_{C2} | \mu = \mu_1) \end{aligned} \quad (6.73)$$

Damit berechnet sich die Gütfunktion für das vorliegende Beispiel durch die Standardnormalverteilung aus

$$1 - \beta(\mu_1) = 1 + F\left(\frac{\mu_{C1} - \mu_1}{\sigma / \sqrt{N}}\right) - F\left(\frac{\mu_{C2} - \mu_1}{\sigma / \sqrt{N}}\right) = 1 + F\left(\frac{22.14 - \mu_1}{\sqrt{0.9}}\right) - F\left(\frac{25.86 - \mu_1}{\sqrt{0.9}}\right) \quad (6.74)$$

Bild 6.12 stellt die Güte des Hypothesentests mit der Alternativhypothese $\mu_1 \neq \mu_0$ als Funktion des alternativen Mittelwertes μ_1 dar.

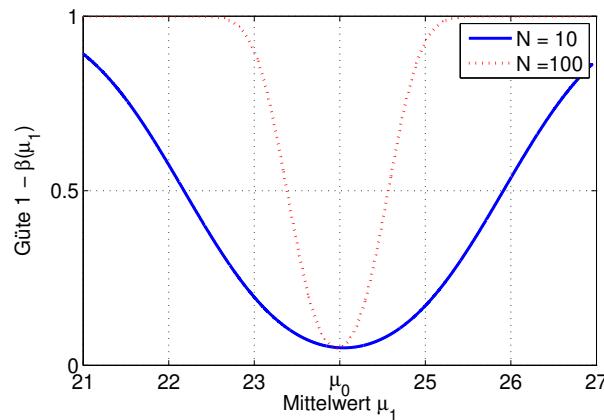


Bild 6.12: Darstellung der Gütfunktion für die Alternativhypothese $\mu_1 \neq \mu_0$

Je weiter der Wert μ_1 von dem Wert μ_0 abweicht, desto sicherer ist erwartungsgemäß die Aussage, dass die Mittelwerte voneinander abweichen. Für $\mu_1 = \mu_0$ besitzt die Güte den Wert des Signifikanzniveaus α . Angenommen der wahre Mittelwert liegt bei $\mu_1 = 26$, dann ist die Güte dieses Tests ca. 55 %. Ein Mittelwert $\mu_1 = 26$ wird nur mit einer Wahrscheinlichkeit von ca. 55 % als Fehler erkannt.

Würde der Test mit einem größeren Stichprobenumfang von $N = 100$ durchgeführt, würde sich die Varianz des Mittelwertes verkleinern und die kritischen Grenzen ergäben sich zu $\mu_{C1} = 23.41$ und $\mu_{C2} = 24.59$. Bild 6.13 verdeutlicht für die Alternativhypothese $\mu_1 \neq \mu_0$, dass mit höherem Stichprobenumfang die Gütfunktion des Hypothesentests einen steileren Verlauf bekommt, also eine größere Trennschärfe besitzt als bei kleinerem Stichprobenumfang.

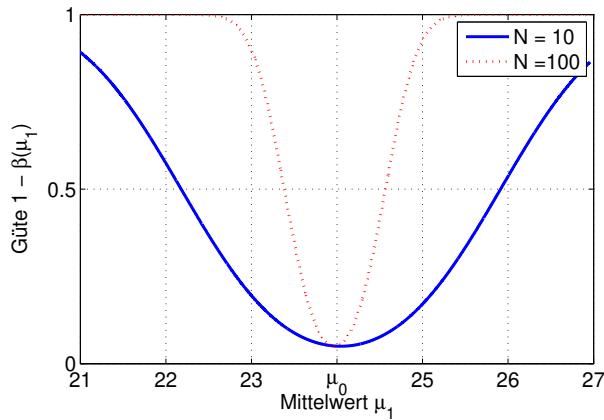


Bild 6.13: Darstellung der Gütfunktion für die Alternativhypothese $\mu_1 \neq \mu_0$ mit unterschiedlichen Stichprobenumfängen

Wird der Stichprobenumfang erhöht, wird ein wahrer Mittelwert von $\mu_1 = 26$ mit einer Güte von praktisch 100 % erkannt.

Allerdings muss der Stichprobenumfang in der Praxis aus Kostengründen so klein wie möglich gehalten werden. Interessiert zum Beispiel eine Abweichung von ± 2 Einheiten, so ist für dieses Beispiel eine Stichprobe vom Umfang von $N = 10$ Werten zu gering, denn für die Werte $\mu = 22$ und 26 beträgt das Risiko eines Fehlers zweiter Art noch fast 50 %. Mit einem Stichprobenumfang von $N = 100$ Werten reicht die Aussagesicherheit sicher aus, die Wahrscheinlichkeit für einen Fehler zweiter Art beträgt für die Werte $\mu = 22$ und 26 weniger als 1 ppm. Realistisch ist ein Stichprobenumfang zwischen diesen Werten. In Abschnitt 6.4.3 wird die Abschätzung des erforderlichen Stichprobenumfangs für die Erkennung einer Abweichung von Mittelwerten diskutiert.

Die Eigenschaften der Gütfunktion lassen sich wie folgt zusammenfassen:

- Mit sinkendem Signifikanzniveau α sinkt die Güte des Tests und die Wahrscheinlichkeit für den Fehler zweiter Art steigt an.
- Die Güte wird mit wachsendem Abstand der Parameter μ_0 und μ_1 größer.
- Mit wachsendem Stichprobenumfang N wird die Trennschärfe eines Hypothesentests größer.

Abschätzung des notwendigen Stichprobenumfang

Auf Basis des Signifikanzniveaus α und der Gütfunktion $(1 - \beta)$ kann der Stichprobenumfang abgeschätzt werden, der für eine Unterscheidung von zwei Mittelwerten mit der Differenz $\Delta\mu$ mit einer definierten Güte notwendig ist. Ausgangspunkt für die hier dargestellte Abschätzung ist aus Gründen der Übersichtlichkeit eine normalverteilte Grundgesamtheit mit der bekannten Standardabweichung σ . Die zu prüfende Nullhypothese ist, dass die Grundgesamtheit den Mittelwert μ_0 aufweist. Als Alternativhypothese wird ein Mittelwert $\mu_1 > \mu_0$ geprüft. Die Nullhypothese wird mit dem Signifikanzniveau α geprüft. Damit ergibt sich für die Annahme des Hypothesentests eine Grenze μ_C für die Annahme mit der Bedingung

$$P(\bar{x} < \mu_C | \mu = \mu_0) = P\left(\bar{x} < \mu_0 + \frac{c \cdot \sigma}{\sqrt{N}} | \mu = \mu_0\right) = F\left(\frac{\mu_C - \mu_0}{\sigma / \sqrt{N}}\right) = \gamma = 1 - \alpha \quad (6.75)$$

Durch Auflösen der Gleichung (6.75) ergibt sich mit der inversen Standardnormalverteilung die Grenze des Annahmebereich μ_C zu

$$\mu_C = F^{-1}(1 - \alpha) \cdot \frac{\sigma}{\sqrt{N}} + \mu_0 \quad (6.76)$$

Die Gütefunktion ergibt sich unter Annahme der Alternativhypothese $\mu_1 > \mu_0$ zu

$$1 - \beta(\mu_1) = P(\bar{x} > \mu_C | \mu = \mu_1) = 1 - P(\bar{x} < \mu_C | \mu = \mu_1) = 1 - F\left(\frac{\mu_C - \mu_1}{\sigma/\sqrt{N}}\right) \quad (6.77)$$

Durch Umformen von Gleichung (6.77) ergibt sich ein weiterer Ausdruck für die Grenze des Annahmebereichs μ_C zu

$$\mu_C = F^{-1}(\beta(\mu_1)) \cdot \frac{\sigma}{\sqrt{N}} + \mu_1 \quad (6.78)$$

Um eine Gleichung für die Abschätzung des Stichprobenumfangs zu erhalten, werden Gleichung (6.76) und Gleichung (6.78) zusammengeführt

$$\mu_C = F^{-1}(1 - \alpha) \cdot \frac{\sigma}{\sqrt{N}} + \mu_0 = F^{-1}(\beta(\mu_1)) \cdot \frac{\sigma}{\sqrt{N}} + \mu_1 \quad (6.79)$$

und nach dem Stichprobenumfang N aufgelöst

$$N = \left(\frac{\sigma}{\mu_1 - \mu_0} \right)^2 \cdot \left(F^{-1}(1 - \alpha) - F^{-1}(\beta(\mu_1)) \right)^2 = \left(\frac{\sigma}{\Delta\mu} \right)^2 \cdot \left(F^{-1}(1 - \alpha) - F^{-1}(\beta(\mu_1)) \right)^2 \quad (6.80)$$

Der Stichprobenumfang steigt erwartungsgemäß mit wachsendem Verhältnis von Standardabweichung σ zum Abstand $\Delta\mu$ der zu unterscheidenden Mittelwerte an. Für ein typisches Signifikanzniveau von $\alpha = 5\%$ und eine Güte von $1 - \beta = 90\%$ ergibt sich

$$N = \left(\frac{\sigma}{\Delta\mu} \right)^2 \cdot 8.56 \quad (6.81)$$

Die hier durchgeführte Herleitung gilt unter der Annahme einer einseitigen Alternative und bekannter Standardabweichung σ . Analog kann auch für eine zweiseitige Alternativhypothese $\mu_1 \neq \mu_0$ und eine unbekannte Standardabweichung σ vorgegangen werden. Beide Änderungen führen zu einer Vergrößerung des notwendigen Stichprobenumfangs. In der Literatur wird deshalb für den Stichprobenumfang die Abschätzung

$$N = \left(\frac{\sigma}{\Delta\mu} \right)^2 \cdot 30 \quad (6.82)$$

angegeben.

6.5 Hypothesentests für die Parameter einer Normalverteilung

In den folgenden Abschnitten werden verschiedenen Hypothesentests für die Parameter einer Normalverteilung vorgestellt. Dabei wird von einem zweiseitig begrenzten Annahmebereich ausgegangen. Bei Hypothesentests für einseitig begrenzte Annahmebereiche muss die Grenze des Annahmebereichs μ_C entsprechend der Alternativhypothese angepasst werden. Dieses Vorgehen wird anhand einiger Beispiele dargestellt.

6.5.1 Test auf Mittelwert μ_0 bei bekannter Varianz (Ein-Stichproben-z-Test)

Der Test auf einen bestimmten Mittelwert μ_0 bei bekannter Varianz wird bereits als einführendes Beispiel in Abschnitt 6.1 herangezogen. Der Hypothesentest arbeitet mit den Hypothesen

- Nullhypothese $H_0 : \mu = \mu_0$
- Alternativhypothese $H_1 : \mu \neq \mu_0$

Dieser Test wird auch als Gauß- oder z-Test bezeichnet. Er wird bei einer zumindest näherungsweise normalverteilten Stichprobe mit bekannter Varianz angewendet. Das Vorgehen kann in die folgenden Prozessschritte aufgeteilt werden:

Tabelle 6.5: Vorgehen zum Hypothesentest für den Mittelwert einer Normalverteilung mit bekannter Varianz

Nr.	Prozessschritt		
1	Wahl eines Signifikanzniveaus α		
2	Bestimmung der zugehörigen Parameter c_1 und c_2 aus der inversen Standardnormalverteilung $F(c_1) = \frac{\alpha}{2}$ und $F(c_2) = 1 - \frac{\alpha}{2}$		
3	Berechnung des Mittelwertes aus der Stichprobe $\bar{x}_0 = \frac{1}{N} \cdot (x_1 + x_2 + \dots + x_N) = \frac{1}{N} \cdot \sum_{n=1}^N x_n$		
4	Bestimmung des Konfidenzintervalls $\mu_{C1} = \mu_0 + \frac{c_1 \cdot \sigma}{\sqrt{N}} < \bar{x}_0 \leq \mu_0 + \frac{c_2 \cdot \sigma}{\sqrt{N}} = \mu_{C2}$	Berechnung des p-Wertes mit der Standardnormalverteilung $p = F\left(\frac{\bar{x}_0 - \mu_0}{\sigma / \sqrt{N}}\right)$	
5	Für $\nu_{C1}^2 \leq \bar{x}_0^2 < \nu_{C2}^2$ wird die Hypothese angenommen, für $\bar{x}_0^2 \leq \nu_{C1}^2$ oder $\bar{x}_0^2 > \nu_{C2}^2$ wird die Hypothese verworfen	Für $\alpha/2 \leq p < 1 - \alpha/2$ wird die Hypothese angenommen, für $p < \alpha/2$ und $p \geq 1 - \alpha/2$ wird die Hypothese verworfen	

Für einseitige Testbedingungen müssen die Grenzen μ_C entsprechend angepasst werden. Beispiele für den z-Test wurde bereits im einführenden Beispiel zur Prozesskontrolle und in der Diskussion zu den verschiedenen Alternativhypotesen in Abschnitt 6.1 vorgestellt.

6.5.2 Test auf eine Standardabweichung σ_0 (Ein-Stichproben-Chi-Quadrat-Test)

Bei der Herleitung des Konfidenzbereichs der Varianz wird in Abschnitt 4.7.1 gezeigt, dass die Zufallsvariable

$$\chi = \frac{s^2}{\sigma^2} \cdot (N - 1) \quad (6.83)$$

eine Chi-Quadrat-Verteilung mit $N - 1$ Freiheitsgraden besitzt. Die Wahrscheinlichkeit γ , mit der die Variable χ innerhalb des Intervalls $c_1 \dots c_2$ liegt, ist definiert als

$$\gamma = P(c_1 < \chi \leq c_2) = F(c_2) - F(c_1) \quad (6.84)$$

Durch Einsetzen von Gleichung (6.83) in die Definitionsgleichung der Wahrscheinlichkeit γ ergibt sich ein Ausdruck für den Annahmebereich der Nullhypothese.

$$\gamma = P\left(c_1 < \frac{s^2}{\sigma_0^2} \cdot (N - 1) \leq c_2\right) = P\left(\frac{c_1}{(N - 1)} \cdot \sigma_0^2 < s^2 \leq \frac{c_2}{(N - 1)} \cdot \sigma_0^2\right) = P\left(\sigma_{C1}^2 < s^2 \leq \sigma_{C2}^2\right) \quad (6.85)$$

Liegt die vorliegende Stichprobenvarianz s_0^2 in dem definierten Intervall, wird die Nullhypothese angenommen, andernfalls muss sie verworfen werden.

Alternativ kann auch der p-Wert aus der Chi-Quadrat-Verteilung mit $N - 1$ Freiheitsgraden aus der Gleichung

$$p = F\left(\frac{s_0^2}{\sigma_0^2} \cdot (N - 1)\right) \quad (6.86)$$

berechnet werden. Das Vorgehen kann in die folgenden Prozessschritte aufgeteilt werden:

Tabelle 6.6: Vorgehen zum Hypothesentest für die Varianz einer Normalverteilung

Nr.	Prozessschritt	
1	Wahl eines Signifikanzniveaus α	
2	Bestimmung der zugehörigen Parameter c_1 und c_2 aus der inversen Chi-Quadrat-Verteilung mit $N - 1$ Freiheitsgraden $F(c_1) = \frac{\alpha}{2}$ und $F(c_2) = 1 - \frac{\alpha}{2}$	
3	Berechnung des Mittelwertes aus der Stichprobe $s_0^2 = \frac{1}{N-1} \cdot \sum_{n=1}^N (x_n - \bar{x})^2$	
4	Bestimmung des Annahmebereichs $\sigma_{C1}^2 = \frac{c_1}{(N-1)} \cdot \sigma_0^2 < s_0^2 \leq \frac{c_2}{(N-1)} \cdot \sigma_0^2 = \sigma_{C2}^2$	Berechnung des p-Wertes mit der Chi-Quadrat-Verteilung $p = 1 - F\left(\frac{s_0^2}{\sigma_0^2} \cdot (N-1)\right)$
5	Für $\sigma_{C1}^2 \leq s_0^2 < \sigma_{C2}^2$ wird die Hypothese angenommen, für $s_0^2 \leq \sigma_{C1}^2$ oder $s_0^2 > \sigma_{C2}^2$ wird die Hypothese verworfen	Für $\alpha/2 \leq p < 1 - \alpha/2$ wird die Hypothese angenommen, für $p < \alpha/2$ und $p \geq 1 - \alpha/2$ wird die Hypothese verworfen

Beispiel: Analyse einer Messeinrichtung

Als Beispiel wird eine Messeinrichtung analysiert, die eine Restmenge im Kraftstofftank erfassen soll. Die Messeinrichtung gilt als wirksam, wenn die Standardabweichung σ_0 der Restmenge mit einer spezifizierten Wahrscheinlichkeit von $\alpha = 5\%$ unter 0.15 l liegt. Bei einer Stichprobe von 16 Restmengen ergibt sich eine Standardabweichung von $s_0 = 0.158$ l.

Mithilfe des Hypothesentests soll überprüft werden, ob die Genauigkeit der Messeinrichtung für den Anwendungsfall ausreichend ist oder nicht. Dazu wird die Nullhypothese getestet, ob die Standardabweichung der zugrundeliegenden Grundgesamtheit $\sigma_0 \leq 0.15$ l ist. Die Alternativhypothese folgt entsprechend zu $\sigma_0 > 0.15$ l.

Die Aufgabenstellung führt zu einem einseitigen Verwerfungsbereich nach Abschnitt 6.2.2. Mit der Verteilung der Zufallsvariablen aus Gleichung (6.89) ergibt sich ein Annahmebereich der Nullhypothese mit der spezifizierten Wahrscheinlichkeit α von

$$\gamma = 1 - \alpha = P(\chi^2 \leq c) = P\left(\frac{s^2}{\sigma_0^2} \cdot (N-1) \leq c\right) = P\left(s^2 \leq \frac{c}{(N-1)} \cdot \sigma_0^2\right) \quad (6.87)$$

Mit der inversen Chi-Quadrat-Verteilung mit 15 Freiheitsgraden berechnet sich die Grenze c zu 24.9958. Durch Einsetzen in Gleichung (6.87) berechnet sich der Annahmebereich des Hypothesentests zu

$$s^2 \leq \frac{c}{(N-1)} \cdot \sigma_0^2 = \frac{24.9958}{(16-1)} \cdot 0.15^2 = 0.0375 \quad (6.88)$$

Da die aus der Stichprobe berechnete Varianz s_0^2 mit einem Wert von 0.0250 im Annahmebereich des Hypothesentests liegt, wird die Nullhypothese angenommen. Das Messsystem ist somit für die Messung der Restmenge im Kraftstoff mit der spezifizierten Genauigkeit geeignet.

Um eine quantitative Bewertung der Aufgabenstellung durchführen zu können, wird der p-Wert mit den vorliegenden Stichprobenwerten berechnet. Dieser folgt zu

$$p = 1 - F\left(\frac{s_0^2}{\sigma_0^2} \cdot (N - 1)\right) = 1 - P\left(\frac{0.158^2}{0.15^2} \cdot (16 - 1)\right) = 0.3407 > 0.05 = \alpha \quad (6.89)$$

Der Wert liegt mit 34.07 % deutlich über dem Signifikanzniveau von 5 %, sodass die Nullhypothese nicht verworfen werden muss. Dies entspricht der Bewertung des Messsystems in Gleichung (6.88).

6.5.3 Test auf Mittelwert μ_0 bei unbekannter Varianz (Ein-Stichproben-t-Test)

Der Test auf einen bestimmten Mittelwert μ_0 bei unbekannter Varianz wird in der Literatur als Ein-Stichproben-t-Test bezeichnet. Er arbeitet mit den Hypothesen

- $H_0 : \mu = \mu_0$
- Alternativhypothese $H_1 : \mu \neq \mu_0$

Der Test auf Mittelwert μ_0 bei normalverteilter Grundgesamtheit aber unbekannter Varianz ist vergleichbar mit dem Ein-Stichproben-z-Test. Der wesentliche Unterschied besteht in der zugrunde liegenden Verteilung. Trifft die Nullhypothese zu, besitzt die Zufallsvariable

$$t = \frac{\bar{x} - \mu_0}{s/\sqrt{N}} \quad (6.90)$$

nach den Ausführungen in Kapitel 5 eine t-Verteilung mit $N - 1$ Freiheitsgraden. Damit berechnet sich der Annahmebereich für die Nullhypothese zu

$$\gamma = P\left(c_1 < \frac{\bar{x} - \mu_0}{s/\sqrt{N}} \leq c_2\right) = P\left(\mu_0 + \frac{c_1 \cdot s}{\sqrt{N}} < \bar{x} \leq \mu_0 + \frac{c_2 \cdot s}{\sqrt{N}}\right) \quad (6.91)$$

Die Konstanten c_1 und c_2 ergeben sich dabei aus der inversen t-Verteilung mit $N - 1$ Freiheitsgraden. Liegt der Stichprobenmittelwert in dem Annahmebereich

$$\mu_{C1} = \mu_0 + \frac{c_1 \cdot s}{\sqrt{N}} < \bar{x}_0 \leq \mu_0 + \frac{c_2 \cdot s}{\sqrt{N}} = \mu_{C2} \quad (6.92)$$

kann die Nullhypothese beibehalten werden, andernfalls gilt die Alternativhypothese. Alternativ kann auch der p-Value mit der Gleichung

$$p = F\left(\frac{\bar{x}_0 - \mu_0}{s/\sqrt{N}}\right) \quad (6.93)$$

berechnet werden. Das Vorgehen kann in die folgenden Prozessschritte aufgeteilt werden:

Tabelle 6.7: Vorgehen zur zum Hypothesentest für den Mittelwert einer Normalverteilung mit unbekannter Varianz

Nr.	Prozessschritt		
1	Wahl eines Signifikanzniveaus α		
2	Bestimmung der zugehörigen Parameter c_1 und c_2 aus der inversen t-Verteilung mit $N - 1$ Freiheitsgraden $F(c_1) = \frac{\alpha}{2}$ und $F(c_2) = 1 - \frac{\alpha}{2}$		
3	Berechnung des Mittelwertes aus der Stichprobe $\bar{x}_0 = \frac{1}{N} \cdot (x_1 + x_2 + \dots + x_N) = \frac{1}{N} \cdot \sum_{n=1}^N x_n$		
4	Bestimmung des Annahmebereichs $\mu_{C1} = \mu_0 + \frac{c_1 \cdot s}{\sqrt{N}} < \bar{x}_0 \leq \mu_0 + \frac{c_2 \cdot s}{\sqrt{N}} = \mu_{C2}$	Berechnung des p-Wertes mit der t-Verteilung $p = F\left(\frac{\bar{x}_0 - \mu_0}{s/\sqrt{N}}\right)$	
5	Für $\mu_{C1} \leq \bar{x}_0 < \mu_{C2}$ wird die Hypothese angenommen, für $\bar{x}_0 \leq \mu_{C1}$ oder $\bar{x}_0 > \mu_{C2}$ wird die Hypothese verworfen	Für $\alpha/2 \leq p < 1 - \alpha/2$ wird die Hypothese angenommen, für $p < \alpha/2$ und $p \geq 1 - \alpha/2$ wird die Hypothese verworfen	

Beispiel: Bruchkraft von Seilen

Anhand eines Beispiels mit unbekannter Varianz wird dieser Hypothesentest weiter vertieft. Es handelt sich um einen Zugversuch mit $N = 16$ Stichproben, bei dem die Bruchkraft von Seilen untersucht wird. Getestet wird die Hypothese $\mu_0 = 45000$ N gegen die Alternative $\mu < 45000$ N. Dabei wird vorausgesetzt, dass die Bruchkraft eine normalverteilte Zufallsvariable ist. Der Wert μ_0 ist zum Beispiel der vom Hersteller angegebene Sollwert, der bei Eingangstests getestet werden soll. Als Signifikanzniveau wird $\alpha = 5\%$ angenommen.

Die Alternativhypothese $\mu < \mu_0$ führt zu einem einseitigen Verwerfungsbereich nach Abschnitt 6.2.3. Trifft die Nullhypothese der Aufgabenstellung zu, besitzt die Zufallsvariable

$$t = \frac{\bar{x} - \mu_0}{s/\sqrt{N}} \quad (6.94)$$

eine t-Verteilung mit $N - 1 = 15$ Freiheitsgraden. Bei Annahme eines einseitigen Tests mit dem Verwerfungsbereich $\mu < \mu_0$ ergibt sich die Konstante c aus der Bedingung

$$F(c) = 1 - \gamma = \alpha \quad (6.95)$$

Das Auflösen von Gleichung (6.95) nach c führt mit der inversen t-Verteilung mit 15 Freiheitsgraden zu

$$c = F^{-1}(\alpha) = F^{-1}(0.05) = -1.7531 \quad (6.96)$$

Für die Modellannahme ergibt sich ein Annahmebereich für den Hypothesentest aus

$$\gamma = 1 - \alpha = P\left(c < \frac{\bar{x} - \mu_0}{s/\sqrt{N}}\right) = P\left(\mu_0 + \frac{c \cdot s}{\sqrt{N}} < \bar{x}\right) \quad (6.97)$$

Mit einem Stichprobenmittelwert von

$$\bar{x}_0 = 44820N \quad (6.98)$$

und einer Standardabweichung von

$$s = 1150N \quad (6.99)$$

aus der Versuchsdurchführung ergibt sich die Grenze des Annahmebereich zu

$$\mu_C = \mu_0 + \frac{c \cdot s}{\sqrt{N}} = 45000 + \frac{-1.7531 \cdot 1150}{\sqrt{16}} = 44495.98N \quad (6.100)$$

Da der Stichprobemittelwert größer als die kritische Grenze μ_C ist, wird der Hypothesentest angenommen.

Um die Fragestellung quantitativ bewerten zu können, wird alternativ zum Vergleich des Stichprobemittelwertes \bar{x}_0 mit der kritischen Grenzen μ_C die Überschreitungswahrscheinlichkeit p der Prüfgröße \bar{x}_0 bestimmt und mit dem Signifikanzniveau α verglichen. Für die obige Modellannahme mit der Alternativhypothese $\mu < \mu_0$ berechnet sich die Überschreitungswahrscheinlichkeit p zu

$$p = F\left(\frac{\bar{x}_0 - \mu_0}{s/\sqrt{N}}\right) = F\left(\frac{44820 - 45000}{1150/\sqrt{16}}\right) = 0.2703 > 0.05 = \alpha \quad (6.101)$$

Die Bewertung der Bruchkraft von Seilen mithilfe des p-Values führt zu einem vergleichbaren Ergebnis wie der Vergleich des Stichprobemittelwertes \bar{x}_0 mit der kritischen Grenzen μ_C . Die Nullhypothese muss auf Basis der vorliegenden Stichprobenwerte nicht verworfen werden.

6.5.4 Zusammenfassung der Hypothesentests für die Parameter einer Normalverteilung

Die in diesem Abschnitt beschriebenen Tests sind Testverfahren, die in den gebräuchlichen Statistik-Programmen implementiert sind. Zum Nachschlagen sind in Tabelle 6.8 die beschriebenen Testverfahren noch einmal kurz zusammengefasst.

Tabelle 6.8: Übersicht über die Hypothesentests für die Parameter einer Normalverteilung

Testverfahren für die Parameter einer Normalverteilung				
Parameter	Funktion	Hypothesentest	Verteilung	Bemerkungen
Mittelwert bei bekannter Varianz	Prüfung von einer Stichprobe mit Stichprobenumfang N einer normalverteilten Zufallsvariable mit bekannter Varianz auf einen Mittelwert μ_0	Ein-Stichproben-z-Test	Normalverteilung	Test kann einseitig oder zweiseitig definiert werden
Varianz	Prüfung von einer Stichprobe mit Stichprobenumfang N einer normalverteilten Zufallsvariable auf die Varianz σ_0^2	Ein-Stichproben-Chi ² -Test	Chi ² -Verteilung mit N - 1 Freiheitsgraden	Test kann einseitig oder zweiseitig definiert werden
Mittelwert bei unbekannter Varianz	Prüfung von einer Stichprobe mit Stichprobenumfang N einer normalverteilten Zufallsvariable mit unbekannter Varianz auf einen Mittelwert μ_0	Ein-Stichproben-t-Test	t-Verteilung mit N - 1 Freiheitsgraden	Test kann einseitig oder zweiseitig definiert werden

6.6 Hypothesentests für den Vergleich zweier Normalverteilungen

Bislang wurden Hypothesentests auf Basis einer Stichprobe durchgeführt. Bei der Auswertung von Labor-Versuchen tritt oft der Fall auf, dass zwei Stichproben $x_{11}, x_{12}, \dots, x_{1N}$ und $x_{21}, x_{22}, \dots, x_{2M}$ miteinander verglichen werden sollen. In der Praxis kann dabei meist von einer normalverteilten Grundgesamtheit ausgegangen werden.

6.6.1 Test auf gleichen Mittelwert bei gepaarten Stichproben (Ein-Stichproben-t-Test)

Der Test basiert auf zwei gleich großen Stichproben $x_{11}, x_{12}, \dots, x_{1N}$ und $x_{21}, x_{22}, \dots, x_{2N}$, bei denen je ein Wert der einen und ein Wert der anderen Stichprobe zusammengehören, weil sie von demselben Individuum kommen. Zum Beispiel kann das ein Maß eines Werkstücks sein, das mit zwei unterschiedlichen Messgeräten ermittelt wurde.

In diesem Fall wird die Differenz der paarweise zusammengehörigen Werte gebildet und die Hypothese getestet, dass die Grundgesamtheit, aus der die Differenzen stammen, den Mittelwert 0 aufweist. Der Hypothesentest arbeitet somit mit den Hypothesen

- Nullhypothese $H_0: \mu = 0$
- Alternativhypothese $H_1: \mu \neq 0$

Ist die Varianz bekannt, entspricht das Vorgehen dem Ein-Stichproben-z-Test. Bei unbekannter Varianz wird der ein-Stichproben-t-Test mit $N - 1$ Freiheitsgraden verwendet.

Falls ein Experiment noch in der Planungsphase ist, ist diese Variante des Vergleichs der Mittelwerte zweier Verteilungen zu bevorzugen, weil dabei die Variabilität zwischen den Versuchsobjekten eliminiert wird und sich so klarere Ergebnisse ergeben.

6.6.2 Test auf gleiche Mittelwerte bei bekannter Varianz (Zwei-Stichproben-z-Test)

Der Test auf einen bestimmten Mittelwert μ_0 bei bekannter Varianz wurde bereits als einführendes Beispiel in Abschnitt 6.1 beschrieben. Der Hypothesentest auf gleiche Mittelwerte zweier Stichproben geht analog vor. Dabei wird die Differenz der Mittelwerte $\mu_1 - \mu_2$ gebildet. Der Hypothesentest arbeitet mit den Hypothesen

- Nullhypothese $H_0: \mu = \mu_1 - \mu_2 = \mu_0$
- Alternativhypothese $H_1: \mu = \mu_1 - \mu_2 \neq \mu_0$

Die Messwerte der beiden Stichproben entsprechen unabhängigen normalverteilten Zufallsvariablen mit dem arithmetischen Mittelwert

$$\bar{x}_1 = \frac{1}{N} \cdot \sum_{n=1}^N x_{1n} \quad (6.102)$$

beziehungsweise

$$\bar{x}_2 = \frac{1}{M} \cdot \sum_{m=1}^M x_{2m} \quad (6.103)$$

und der Varianz σ^2 der Grundgesamtheit. Der Mittelwert der Differenz der beiden Stichproben berechnet sich aus

$$\bar{x} = \bar{x}_1 - \bar{x}_2 \quad (6.104)$$

Mit den Rechenregeln für mehrere Zufallsvariablen in Kapitel 4 ist bekannt, dass der Mittelwert der Differenz zweier Zufallsvariablen den Mittelwert

$$\mu = \mu_1 - \mu_2 \quad (6.105)$$

und eine Varianz von

$$\sigma_x^2 = \sigma_{\bar{x}_1}^2 + \sigma_{\bar{x}_2}^2 = \frac{\sigma^2}{N} + \frac{\sigma^2}{M} = \sigma^2 \cdot \left(\frac{1}{N} + \frac{1}{M} \right) \quad (6.106)$$

besitzt. Mit der Standardisierung der Zufallsvariablen

$$z = \frac{\bar{x} - \mu}{\sqrt{\sigma_x^2}} = \frac{\bar{x} - \mu}{\sqrt{\sigma^2 \cdot \left(\frac{1}{N} + \frac{1}{M} \right)}} \quad (6.107)$$

geht die Verteilung in eine Standardnormalverteilung über, sie weist also den Mittelwert $\mu_z = 0$ und die Standardabweichung $\sigma_z = 1$ auf. Mit dieser Verteilung wird die Wahrscheinlichkeit γ , mit der die Variable z in dem Intervall $c_1 \dots c_2$ liegt, definiert als

$$P(c_1 < z \leq c_2) = F(c_2) - F(c_1) = \gamma \quad (6.108)$$

Bei Annahme eines symmetrischen Hypothesentests ergeben sich die Konstanten c_1 und c_2 aus den Bedingungen

$$F(c_1) = \frac{1 - \gamma}{2} \quad (6.109)$$

und

$$F(c_2) = 1 - \frac{1 - \gamma}{2} = \frac{1 + \gamma}{2} \quad (6.110)$$

Auflösen nach c_1 und c_2 führt zu

$$c_1 = F^{-1} \left(\frac{1 - \gamma}{2} \right) \quad (6.111)$$

und

$$c_2 = F^{-1} \left(\frac{1 + \gamma}{2} \right) \quad (6.112)$$

Mit der gewählten Wahrscheinlichkeit γ liegt die Differenz zweier Mittelwerte $\mu_1 - \mu_2$ unter Annahme der Nullhypothese in dem Intervall

$$\gamma = P \left(c_1 < \frac{\bar{x} - \mu_0}{\sqrt{\sigma^2 \cdot \left(\frac{1}{N} + \frac{1}{M} \right)}} \leq c_2 \right) = P \left(\mu_0 + c_1 \cdot \sqrt{\sigma^2 \cdot \left(\frac{1}{N} + \frac{1}{M} \right)} < \bar{x} \leq \mu_0 + c_2 \cdot \sqrt{\sigma^2 \cdot \left(\frac{1}{N} + \frac{1}{M} \right)} \right) \quad (6.113)$$

Aus Gleichung (6.113) folgt der Annahmebereich der Nullhypothese zu

$$\mu_{C1} = \mu_0 + c_1 \cdot \sqrt{\sigma^2 \cdot \left(\frac{1}{N} + \frac{1}{M} \right)} < \bar{x} \leq \mu_0 + c_2 \cdot \sqrt{\sigma^2 \cdot \left(\frac{1}{N} + \frac{1}{M} \right)} = \mu_{C2} \quad (6.114)$$

Liegt die Differenz der Stichprobenmittelwerte in dem durch Gleichung (6.113) definierten Intervall, kann die Nullhypothese beibehalten werden.

Das Vorgehen zur Bestimmung des Annahmebereichs für die Differenz zweier Mittelwerte bei bekannter Varianz der Grundgesamtheit wird in Tabelle 6.9 zusammengefasst.

Tabelle 6.9: Durchführung eines Hypothesentests auf gleiche Mittelwerte bei bekannter Varianz

Nr.	Prozessschritt	
1	Wahl eines Signifikanzniveaus α	
2	Bestimmung der zugehörigen Parameter c_1 und c_2 aus der inversen Standardnormalverteilung $F(c_1) = \frac{\alpha}{2}$ und $F(c_2) = 1 - \frac{\alpha}{2}$	
3	Berechnung des Mittelwertes aus der Stichprobe $\bar{x}_0 = \bar{x}_1 - \bar{x}_2 = \frac{1}{N} \cdot \sum_{n=1}^N x_{1n} - \frac{1}{M} \cdot \sum_{m=1}^M x_{2m}$	
4	Bestimmung der unteren Annahmegrenze $\mu_{C1} = \mu_0 + c_1 \cdot \sqrt{\sigma^2 \cdot \left(\frac{1}{N} + \frac{1}{M} \right)}$	Bestimmung der oberen Annahmegrenze $\mu_{C2} = \mu_0 + c_2 \cdot \sqrt{\sigma^2 \cdot \left(\frac{1}{N} + \frac{1}{M} \right)}$
5	Bestimmung des Annahmebereichs $\mu_{C1} < \bar{x}_0 \leq \mu_{C2}$	Berechnung des p-Values mit der Standardnormalverteilung $p = F\left(\frac{\bar{x}_0 - \mu_0}{\sqrt{\sigma^2 \cdot \left(\frac{1}{N} + \frac{1}{M} \right)}} \right)$
6	Für $\mu_{C1} \leq \bar{x}_0 < \mu_{C2}$ wird die Hypothese angenommen, für $\bar{x}_0 \leq \mu_{C1}$ oder $\bar{x}_0 > \mu_{C2}$ wird die Hypothese verworfen	Für $\alpha/2 \leq p < 1 - \alpha/2$ wird die Hypothese angenommen, für $p < \alpha/2$ und $p \geq 1 - \alpha/2$ wird die Hypothese verworfen

Für einseitige Testbedingungen müssen die Grenzen μ_C entsprechend angepasst werden.

6.6.3 Test auf gleiche Mittelwerte bei unbekannter Varianz (Zwei-Stichproben-t-Test)

Sind die Varianzen der beiden Versuchsergebnisse nicht bekannt, muss die Berechnung des Mittelwertes auf die t-Verteilung zurückgeführt werden. Dies ist allerdings nur dann möglich, wenn die beiden Grundgesamtheiten dieselbe unbekannte Varianz σ^2 aufweisen. Es wird daher davon ausgegangen, dass die Stichprobe 1 einer normalverteilten Grundgesamtheit mit Mittelwert μ_1 und einer unbekannten Varianz σ^2 und die Stichprobe 2 einer normalverteilten Grundgesamtheit mit Mittelwert μ_2 und derselben Varianz σ^2 entstammt. Der Hypothesentest arbeitet wieder mit den Hypothesen

- Nullhypothese $H_0: \mu = \mu_1 - \mu_2 = \mu_0$
- Alternativhypothese $H_1: \mu = \mu_1 - \mu_2 \neq \mu_0$

Die Stichprobenmittelwerte ergeben sich zu

$$\bar{x}_1 = \frac{1}{N} \cdot \sum_{n=1}^N x_{1n} \quad (6.115)$$

beziehungsweise

$$\bar{x}_2 = \frac{1}{M} \cdot \sum_{m=1}^M x_{2m} \quad (6.116)$$

und die Varianz der Stichprobe folgt zu

$$s_1^2 = \frac{1}{N-1} \cdot \sum_{n=1}^N (x_{1n} - \bar{x}_1)^2 \quad (6.117)$$

beziehungsweise

$$s_2^2 = \frac{1}{M-1} \cdot \sum_{m=1}^M (x_{2m} - \bar{x}_2)^2 \quad (6.118)$$

Um den Mittelwert der Differenz der beiden Mittelwerte zu berechnen, wird eine neue Zufallsvariable eingeführt.

$$\bar{x} = \bar{x}_1 - \bar{x}_2 \quad (6.119)$$

Mit den Ausführungen bezüglich der Rechenregeln zum Umgang mit mehreren Zufallsvariablen besitzt diese eine Varianz von

$$s^2 = \frac{\sum_{n=1}^N (x_{1n} - \bar{x}_1)^2 + \sum_{m=1}^M (x_{2m} - \bar{x}_2)^2}{N + M - 2} = \frac{(N-1) \cdot s_1^2 + (M-1) \cdot s_2^2}{N + M - 2} \quad (6.120)$$

Weiterhin gilt, dass der Mittelwert der Differenz zweier Zufallsvariablen den Mittelwert

$$\mu = \mu_1 - \mu_2 \quad (6.121)$$

und eine Varianz von

$$\sigma_{\bar{X}}^2 = \sigma_{\bar{X}_1}^2 + \sigma_{\bar{X}_2}^2 = \frac{\sigma^2}{N} + \frac{\sigma^2}{M} = \sigma^2 \cdot \left(\frac{1}{N} + \frac{1}{M} \right) \quad (6.122)$$

besitzt. Damit besitzt die Zufallsvariable

$$z = \frac{\bar{x} - \mu}{\sqrt{\sigma_{\bar{X}}^2}} = \frac{\bar{x} - \mu}{\sqrt{\sigma^2 \cdot \left(\frac{1}{N} + \frac{1}{M} \right)}} \quad (6.123)$$

eine Standardnormalverteilung. Analog zu der chi-quadrat-verteilten Zufallsvariablen aus Gleichung (5.39) besitzt die Zufallsvariable

$$\chi = \frac{(N + M - 2) \cdot s^2}{\sigma^2} \quad (6.124)$$

eine Chi-Quadrat-Verteilung mit $N + M - 2$ Freiheitsgraden. Mit der Zufallsvariablen aus Gleichung (6.123) und der Zufallsvariablen aus Gleichung (6.124) kann nach Gleichung (4.240) die Zufallsvariable t der t-Verteilung gebildet werden.

$$t = \frac{z}{\sqrt{\frac{\chi}{\nu}}} = \frac{\bar{x} - \mu}{\sqrt{\sigma^2 \cdot \left(\frac{1}{N} + \frac{1}{M}\right) \cdot \frac{(N + M - 2) \cdot s^2}{\sigma^2 \cdot (N + M - 2)}}} = \frac{\bar{x} - \mu}{\sqrt{\frac{1}{N} + \frac{1}{M}} \cdot s} \quad (6.125)$$

Die Zufallsvariable t aus Gleichung (6.125) besitzt eine t-Verteilung mit $N + M - 2$ Freiheitsgraden. Damit wird die Wahrscheinlichkeit γ , mit der die Variable t in dem Intervall $c_1 \dots c_2$ liegt, definiert werden als

$$P(c_1 < t \leq c_2) = F(c_2) - F(c_1) = \gamma \quad (6.126)$$

Bei Annahme einer symmetrischen Alternativhypothese ergeben sich die Konstanten c_1 und c_2 aus den Bedingungen

$$F(c_1) = \frac{1 - \gamma}{2} \quad (6.127)$$

und

$$F(c_2) = 1 - \frac{1 - \gamma}{2} = \frac{1 + \gamma}{2} \quad (6.128)$$

Auflösen nach c_1 und c_2 führt zu

$$c_1 = F^{-1}\left(\frac{1 - \gamma}{2}\right) \quad (6.129)$$

und

$$c_2 = F^{-1}\left(\frac{1 + \gamma}{2}\right) \quad (6.130)$$

Mit der gewählten Wahrscheinlichkeit γ liegt die Differenz zweier Mittelwert $\mu_1 - \mu_2$ unter Annahme der Nullhypothese in dem Annahmeintervall

$$\gamma = P\left(c_1 < \frac{\bar{x} - \mu_0}{\sqrt{\frac{1}{N} + \frac{1}{M}} \cdot s} \leq c_2\right) = P\left(\mu_0 + c_1 \cdot \sqrt{\frac{1}{N} + \frac{1}{M}} \cdot s < \bar{x} \leq \mu_0 + c_2 \cdot \sqrt{\frac{1}{N} + \frac{1}{M}} \cdot s\right) \quad (6.131)$$

Durch Umformungen ergibt sich ein Ausdruck für den Annahmebereich der Differenz zweier Mittelwerte bei unbekannter Varianz.

$$\mu_{C1} = \mu_0 + c_1 \cdot \sqrt{\frac{1}{N} + \frac{1}{M}} \cdot s < \bar{x} \leq \mu_0 + c_2 \cdot \sqrt{\frac{1}{N} + \frac{1}{M}} \cdot s = \mu_{C2} \quad (6.132)$$

Das Vorgehen zur Bestimmung des Konfidenzintervalls für die Differenz zweier Mittelwerte bei unbekannter Varianz wird in Tabelle 6.10 zusammengefasst.

Tabelle 6.10: Vorgehen zur Durchführung eines Hypothesentests auf gleiche Mittelwerte bei unbekannter Varianz

Nr.	Prozessschritt	
1	Wahl eines Signifikanzniveaus α	
2	Bestimmung der zugehörigen Parameter c_1 und c_2 aus der inversen Standardnormalverteilung $F(c_1) = \frac{\alpha}{2}$ und $F(c_2) = 1 - \frac{\alpha}{2}$	
3	Berechnung des Mittelwertes aus der Stichprobe $\bar{x}_0 = \bar{x}_1 - \bar{x}_2 = \frac{1}{N} \cdot \sum_{n=1}^N x_{1n} - \frac{1}{M} \cdot \sum_{m=1}^M x_{2m}$	
4	Berechnung der Varianz der Stichproben $s^2 = \frac{\sum_{n=1}^N (x_{1n} - \bar{x}_1)^2 + \sum_{m=1}^M (x_{2m} - \bar{x}_2)^2}{N + M - 2} = \frac{(N - 1) \cdot s_1^2 + (M - 1) \cdot s_2^2}{N + M - 2}$	
5	Bestimmung der unteren Annahmegrenze $\mu_{C1} = \mu_0 + c_1 \cdot \sqrt{\frac{1}{N} + \frac{1}{M}} \cdot s$	Bestimmung der oberen Annahmegrenze $\mu_{C2} = \mu_0 + c_2 \cdot \sqrt{\frac{1}{N} + \frac{1}{M}} \cdot s$
6	Bestimmung des Annahmebereichs $\mu_{C1} < \bar{x}_0 \leq \mu_{C2}$	Berechnung des p-Values mit der t-Verteilung $p = F\left(\frac{\bar{x}_0 - \mu_0}{\sqrt{\frac{1}{N} + \frac{1}{M}} \cdot s}\right)$
7	Für $\mu_{C1} \leq \bar{x}_0 < \mu_{C2}$ wird die Hypothese angenommen, für $\bar{x}_0 \leq \mu_{C1}$ oder $\bar{x}_0 > \mu_{C2}$ wird die Hypothese verworfen	Für $\alpha/2 \leq p < 1 - \alpha/2$ wird die Hypothese angenommen, für $p < \alpha/2$ und $p \geq 1 - \alpha/2$ wird die Hypothese verworfen

Für einseitige Testbedingungen müssen die Grenzen μ_C entsprechend angepasst werden.

Beispiel: Reproduzierbarkeit zweier Messmethoden

Eine Messmethode muss reproduzierbar sein, das bedeutet, dass bei mehrmaligen Messen der gleichen Eigenschaft unter gleichen Bedingungen die Ergebnisse vergleichbar sein müssen. Jedes neue Messverfahren muss hinsichtlich seiner Reproduzierbarkeit verifiziert werden. Im Folgenden soll eine bei der Abkühlung von Gasen verwendete Messmethode mithilfe eines Hypothesentests auf seine Reproduzierbarkeit überprüft werden. Dabei soll die Frage beantwortet werden, ob der Unterschied der Mittelwerte beider Versuchsreihen signifikant ist.

Tabelle 6.11: Stichprobenwerte zur Überprüfung des Messverfahrens hinsichtlich seiner Reproduzierbarkeit

n	1	2	3	4	5
$T_1 / ^\circ\text{C}$	106.9	106.3	107.0	106.0	104.9
$T_2 / ^\circ\text{C}$	106.5	106.7	106.8	106.1	105.6

Die Auswertung der Stichproben liefert die arithmetischen Mittelwerte von

$$\bar{x}_1 = \frac{1}{N} \cdot \sum_{n=1}^N x_{1n} = 106.22 \quad (6.133)$$

beziehungsweise

$$\bar{x}_2 = \frac{1}{M} \cdot \sum_{m=1}^M x_{2m} = 106.34 \quad (6.134)$$

und die Varianzen der Stichprobe von

$$s_1^2 = \frac{1}{N-1} \cdot \sum_{n=1}^N (x_{1n} - \bar{x}_1)^2 = 0.717 \quad (6.135)$$

beziehungsweise

$$s_2^2 = \frac{1}{M-1} \cdot \sum_{m=1}^M (x_{2m} - \bar{x}_2)^2 = 0.243 \quad (6.136)$$

Die für die Bewertung der Problemstellung benötigte Differenz der beiden Mittelwerte folgt zu

$$\bar{x}_0 = \bar{x}_1 - \bar{x}_2 = -0.12 \quad (6.137)$$

und die Varianz berechnet sich aus Gleichung (6.120) zu

$$s^2 = \frac{(N-1) \cdot s_1^2 + (M-1) \cdot s_2^2}{N+M-2} = 0.48 \quad (6.138)$$

Es wird die Hypothese $\mu_1 = \mu_2$ gegen die Hypothese $\mu_1 \neq \mu_2$ für eine Signifikanz von $\alpha = 5\%$ getestet. Damit folgt, dass die Differenz der Mittelwerte μ_1 und μ_2 auf 0 getestet wird. Mit der inversen t-Verteilung lassen sich die Konstanten c_1 und c_2 berechnen. Diese folgen mit dem gewählten Signifikanzniveau zu $c_1 = -2.3060$ und $c_2 = 2.3060$.

Der Annahmebereich für die Nullhypothese berechnet sich mit den Kenngrößen der Stichproben und den obigen Annahmen zu

$$-0.7001 = \mu_0 + c_1 \cdot \sqrt{\frac{1}{N} + \frac{1}{M}} \cdot s < \bar{x}_0 = -0.12 \leq \mu_0 + c_2 \cdot \sqrt{\frac{1}{N} + \frac{1}{M}} \cdot s = 0.7001 \quad (6.139)$$

Da der berechnete Wert der Abweichung der Stichprobenmittelwerte in dem Annahmebereich liegt, wird die Nullhypothese $\mu_0 = 0$ beibehalten. Die Mittelwerte μ_1 und μ_2 unterscheiden sich demnach nicht signifikant. Die Reproduzierbarkeit der Messmethode ist somit gegeben.

Um eine quantitative Bewertung der Aufgabenstellung durchführen zu können, wird der p-Value mit den vorliegenden Stichprobenwerten berechnet. Dieser folgt zu

$$p = F\left(\frac{\bar{x}_0 - \mu_0}{\sqrt{\frac{1}{N} + \frac{1}{M}} \cdot s}\right) = P\left(\frac{-0.12}{\sqrt{\left(\frac{1}{5} + \frac{1}{5}\right) \cdot 0.48}}\right) = 0.3956 \quad (6.140)$$

Da der berechnete p-Value in dem Bereich

$$0.025 = \frac{\alpha}{2} \leq p = 0.3956 < 1 - \frac{\alpha}{2} = 0.975 \quad (6.141)$$

liegt, muss die Nullhypothese nicht verworfen werden. Dies entspricht der Bewertung des Messsystems in Gleichung (6.139).

6.6.4 Test auf gleiche Varianz zweier Normalverteilungen (F-Test)

In der Praxis müssen auch immer wieder die Varianzen überprüft werden, zum Beispiel um die Gleichmäßigkeit einer Produktion zu untersuchen. Ob zwei Normalverteilungen, deren Mittelwerte μ nicht bekannt zu sein brauchen, gleiche Varianzen haben, kann mit dem im Folgenden dargestellten Hypothesentest überprüft werden. Er beruht auf einer f-Verteilung. Der Stichprobenumfang muss nicht gleich sein und wird im Folgenden mit N und M bezeichnet. Der Hypothesentest arbeitet mit den Hypothesen

- Nullhypothese $H_0: \sigma_{12}^2 / \sigma_{22}^2 = v_0$
- Alternativhypothese $H_1: \sigma_{12}^2 / \sigma_{22}^2 \neq v_0$

Die Messwerte der Stichproben werden hierbei wiederum als unabhängige normalverteilte Zufallsvariablen mit den arithmetischen Mittelwerten

$$\bar{x}_1 = \frac{1}{N} \cdot \sum_{n=1}^N x_{1n} \quad (6.142)$$

beziehungsweise

$$\bar{x}_2 = \frac{1}{M} \cdot \sum_{m=1}^M x_{2m} \quad (6.143)$$

und den Stichprobenvarianzen

$$s_1^2 = \frac{1}{N-1} \cdot \sum_{n=1}^N (x_{1n} - \bar{x}_1)^2 \quad (6.144)$$

beziehungsweise

$$s_2^2 = \frac{1}{M-1} \cdot \sum_{m=1}^M (x_{2m} - \bar{x}_2)^2 \quad (6.145)$$

betrachtet. Das Verhältnis der beiden Stichprobenvarianzen

$$f = \frac{\frac{(N-1) \cdot s_1^2}{\sigma_1^2} \cdot \frac{1}{(N-1)}}{\frac{(M-1) \cdot s_2^2}{\sigma_2^2} \cdot \frac{1}{(M-1)}} = \frac{s_1^2}{s_2^2} \cdot \frac{\sigma_2^2}{\sigma_1^2} \quad (6.146)$$

besitzt nach Gleichung (4.236) eine f-Verteilung mit $(N-1|M-1)$ Freiheitsgraden. Mit der Zufallsvariablen f aus Gleichung (5.103) wird die Wahrscheinlichkeit γ , mit der die Variable f in dem Intervall $c_1 \dots c_2$ liegt, definiert als

$$P(c_1 < f \leq c_2) = F(c_2) - F(c_1) = \gamma \quad (6.147)$$

Bei Annahme einer symmetrischen Alternativhypothese ergeben sich die Konstanten c_1 und c_2 mit der inversen f-Verteilung mit $(N-1|M-1)$ Freiheitsgraden aus den Bedingungen

$$F(c_1) = \frac{1-\gamma}{2} \quad (6.148)$$

und

$$F(c_2) = 1 - \frac{1-\gamma}{2} = \frac{1+\gamma}{2} \quad (6.149)$$

Auflösen nach c_1 und c_2 führt zu

$$c_1 = F^{-1} \left(\frac{1 - \gamma}{2} \right) \quad (6.150)$$

und

$$c_2 = F^{-1} \left(\frac{1 + \gamma}{2} \right) \quad (6.151)$$

Mit der gewählten Wahrscheinlichkeit γ liegt das Verhältnis zweier Stichprobenvarianzen unter Annahme der Nullhypothese in dem Bereich

$$\gamma = P \left(c_1 < \frac{s_1^2}{s_2^2} \cdot \frac{\sigma_2^2}{\sigma_1^2} \leq c_2 \right) = P \left(c_1 \cdot \frac{\sigma_1^2}{\sigma_2^2} < \frac{s_1^2}{s_2^2} \leq c_2 \cdot \frac{\sigma_1^2}{\sigma_2^2} \right) = P \left(c_1 \cdot v_0 < \frac{s_1^2}{s_2^2} \leq c_2 \cdot v_0 \right) \quad (6.152)$$

Durch Umformungen ergibt sich ein Ausdruck für den Annahmebereich der Nullhypothese des Verhältnisses zweier Varianzen

$$c_1 \cdot v_0 = c_1 \cdot \frac{\sigma_1^2}{\sigma_2^2} < \frac{s_1^2}{s_2^2} \leq c_2 \cdot \frac{\sigma_1^2}{\sigma_2^2} = c_2 \cdot v_0 \quad (6.153)$$

Das Vorgehen zur Durchführung des Hypothesentests für das Verhältnis zweier Stichprobenvarianzen wird in Tabelle 6.12 zusammengefasst.

Tabelle 6.12: Vorgehen zur Durchführung eines Hypothesentests auf gleiche Varianz zweier Normalverteilungen

Nr.	Prozessschritt	
1	Wahl eines Signifikanzniveaus α	
2	Bestimmung der zugehörigen Parameter c_1 und c_2 aus der inversen Standardnormalverteilung $F(c_1) = \frac{\alpha}{2}$ und $F(c_2) = 1 - \frac{\alpha}{2}$	
3	Berechnung der Stichprobenvarianzen $s_1^2 = \frac{1}{N-1} \cdot \sum_{n=1}^N (x_{1n} - \bar{x}_1)^2$ und $s_2^2 = \frac{1}{M-1} \cdot \sum_{m=1}^M (x_{2m} - \bar{x}_2)^2$	
4	Bestimmung des Annahmebereichs $c_1 \cdot v_0 = c_1 \cdot \frac{\sigma_1^2}{\sigma_2^2} < \frac{s_1^2}{s_2^2} \leq c_2 \cdot \frac{\sigma_1^2}{\sigma_2^2} = c_2 \cdot v_0$	Berechnung des p-Values mit der f-Verteilung $p = 1 - F\left(\frac{s_1^2}{s_2^2} \cdot \frac{1}{v_0}\right)$
5	Für $c_1 \cdot v_0 < s_1^2/s_2^2 < c_2 \cdot v_0$ wird die Hypothese angenommen, für $c_1 \cdot v_0 > s_1^2/s_2^2$ oder $s_1^2/s_2^2 > c_2 \cdot v_0$ wird die Hypothese verworfen	Für $\alpha/2 \leq p < 1 - \alpha/2$ wird die Hypothese angenommen, für $p < \alpha/2$ und $p \geq 1 - \alpha/2$ wird die Hypothese verworfen

Für einseitige Testbedingungen müssen die Grenzen des Annahmebereichs entsprechend angepasst werden.

Beispiel: Untersuchung von Schrauben

Bei der Untersuchung von 16 Schrauben mit gewalztem und gefrästem Gewinde wurde der Flanken-durchmesser bestimmt. Für gewalzte Gewinde ergibt sich der Mittelwert

$$\bar{x}_1 = 23.189 \quad (6.154)$$

und für gefräste Gewinde der Mittelwert

$$\bar{x}_2 = 23.277 \quad (6.155)$$

Die entsprechenden Varianzen berechnen sich zu

$$s_1^2 = 0.001382 \quad (6.156)$$

und

$$s_2^2 = 0.000433 \quad (6.157)$$

Unter der Annahme, dass Normalverteilungen vorliegen, wird die Nullhypothese $\sigma_1^2 = \sigma_2^2$ gegen die Alternativhypothese $\sigma_1^2 > \sigma_2^2$ für ein Signifikanzniveau $\alpha = 0.05$ getestet. Die Stichprobenumfänge sind $N = M = 16$. Damit ergeben sich die Konstante c aus der inversen f-Verteilung mit (15|15)-Freiheitsgraden zu $c = 2.4034$. Mit diesen Werten berechnet sich der Annahmebereich der Nullhypothese zu

$$\frac{s_1^2}{s_2^2} \leq c_2 \cdot v_0 = 2.4034 \quad (6.158)$$

Das Verhältnis der Stichprobenvarianzen liegt mit 3.1916 nicht in dem durch Gleichung (6.158) berechneten Annahmebereich. Die Nullhypothese muss auf Basis der vorliegenden Stichprobenwerte verworfen werden.

Für die quantitative Bewertung der Fragestellung wird zusätzlich der p-Value berechnet. Dieser folgt zu

$$p = 1 - F\left(\frac{s_1^2}{s_2^2} \cdot \frac{\sigma_2^2}{\sigma_1^2}\right) = 1 - P\left(\frac{0.001382}{0.000433}\right) = 0.0156 \quad (6.159)$$

Da der berechnete p-Value kleiner ist als das gewählte Signifikanzniveau α des Hypothesentests, muss die Nullhypothese verworfen werden. Dies entspricht der Bewertung des Messsystems mithilfe des Annahmebereichs. Der Unterschied zwischen den beiden Varianzen ist also signifikant, sodass die gebräuchlichen Schrauben als qualitativ höherwertig anzusehen sind.

6.6.5 Zusammenfassung der Hypothesentests für den Vergleich zweier Normalverteilungen

Die in diesem Abschnitt beschriebenen Tests sind Testverfahren, die in den gebräuchlichen Statistik-Programmen implementiert sind. Zum Nachschlagen sind in Tabelle 6.13 die beschriebenen Testverfahren noch einmal kurz beschrieben. Alle Tests können ein- oder zweiseitig durchgeführt werden.

Tabelle 6.13: Übersicht über die Hypothesentests für den Vergleich zweier Normalverteilungen

Testverfahren für die Parameter einer Normalverteilung			
Parameter	Funktion	Hypothesentest	Verteilung
Differenz der Mittelwerte bei gepaarten Stichproben	Prüfung von zwei gepaarten Stichproben mit Stichprobenumfang N auf die Differenz der Mittelwerte μ_0	Ein-Stichproben-z-Test oder Ein-Stichproben-t-Test	Normalverteilung oder t-Verteilung
Differenz der Mittelwerte bei bekannter Varianz	Prüfung von zwei Stichproben mit unterschiedlichem Stichprobenumfang mit bekannter Varianz auf die Differenz der Mittelwerte μ_0	Zwei-Stichproben-z-Test	Normalverteilung
Differenz der Mittelwert bei unbekannter Varianz	Prüfung von zwei Stichproben mit unterschiedlichem Stichprobenumfang mit unbekannter Varianz auf die Differenz der Mittelwert μ_0	Zwei-Stichproben-z-Test	t-Verteilung mit $N + M - 2$ Freiheitsgraden
Verhältnis der Varianzen	Prüfung von zwei Stichproben mit unterschiedlichem Stichprobenumfang einer auf das Verhältnis der Varianzen v_0	F-Test	f-Verteilung mit $(N - 1, M - 1)$ Freiheitsgraden

6.7 Anwendungsbeispiel: Diagnose von Feuchtesensoren

Hypothesentests werden in der Stochastik zur quantifizierten Prüfung von statistischen Aussagen eingesetzt. In diesem Abschnitt wird gezeigt, dass sie auch für technische Fragestellungen eingesetzt werden können.

Systemfunktionen im Motorsteuergerät erfordern zuverlässige Sensoren und Aktoren. Um ihre Verfügbarkeit und Genauigkeit sicherzustellen, werden Diagnosefunktionen eingesetzt. Die Diagnosefunktionen dienen zur Information des Fahrers über aufgetretene Fehler und zur Aktivierung von Notlauffunktionen, die einen sicheren Betrieb des Fahrzeugs sicherstellen.

In diesem Anwendungsbeispiel wird geprüft, wie im Steuergerät durch den Einsatz von redundanten Sensoren und von einem Hypothesentest eine belastbare Diagnosefunktion implementiert werden kann.

6.7.1 Feuchtesensoren im Motormanagement

Für die Minimierung von Fahrzeugemissionen wird der vom Motor angesaugte Luftmassenstrom gemessen. Eine wesentliche Störgröße bei der Erfassung des Luftmassenstroms ist die Luftfeuchte. Deshalb verfügen einige Luftmassenmesser über ein Modul zur Messung der Luftfeuchte. Kern dieses Moduls sind kapazitive Feuchtesensoren, wie sie zum Beispiel in Bild 6.14 dargestellt sind.

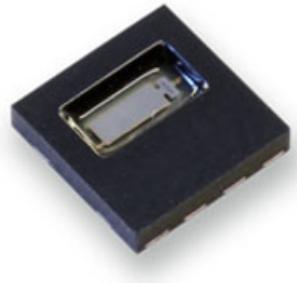


Bild 6.14: Kapazitiver Feuchtesensor

Der Sensor besteht aus einem Kondensator, der mit einem Dielektrikum gefüllt ist, das bei Aufnahme von Luftfeuchtigkeit seine Permittivitätszahl ϵ_R ändert. Es ergibt sich ein weitgehend linearer Zusammenhang zwischen Kapazität des Sensors und relativer Luftfeuchte. Die elektrische Auswertung führt zu einem Sensorsignal, das vom Motorsteuergerät erfasst wird.

6.7.2 Hypothesentest als Diagnosefunktion für Feuchtesensoren

Als Möglichkeit der Sensorüberwachung soll eine redundante Sensoranordnung bewertet werden. Dabei werden zwei identische Sensoren in das Modul integriert. Weichen beide Sensorsignale signifikant voneinander ab, ist zumindest einer der beiden Sensoren defekt.

Zur mathematischen Beschreibung der Diagnoseaufgabe wird das Messergebnis des einen Sensors als x_1 und das Messergebnis des zweiten Sensors mit x_2 bezeichnet. Für die Berechnung wird die Differenz des jeweiligen Messwertes x_n zu seinem Sollverhalten μ_n bewertet. Unter diesen Bedingungen besitzt die Variable

$$z = \frac{x_1 - \mu_1 - (x_2 - \mu_2)}{\sqrt{\sigma_1^2 + \sigma_2^2}} \quad (6.160)$$

eine Standardnormalverteilung. Die beiden Varianzen σ_1^2 und σ_2^2 sind gleich groß und werden als σ^2 bezeichnet. Weiter wird davon ausgegangen, dass die aus der Erprobung der Sensoren bekannte Standardabweichung σ typisch für den Einsatz der Sensoren im Kraftfahrzeug ist. Die Abweichungen der

beiden Messwerte x_1 und x_2 sowie der beiden Mittelwerte μ_1 und μ_2 werden zusammengefasst, zu den Abweichungen Δx und $\Delta \mu$. Damit ergibt sich die standardnormalverteilte Variable

$$z = \frac{x_1 - \mu_1 - (x_2 - \mu_2)}{\sqrt{\sigma_1^2 + \sigma_2^2}} = \frac{\Delta x - \Delta \mu}{\sqrt{2 \cdot \sigma^2}} \quad (6.161)$$

Die Bewertung der Messergebnisse kann damit als Hypothesentest formuliert werden, bei dem geprüft wird, ob beide Sensoren voneinander abweichen.

- Nullhypothese $H_0: \Delta \mu = 0$
- Alternativhypothese $H_1: \Delta \mu \neq 0$

Unter der Annahme, dass die Nullhypothese richtig ist, ist $\Delta \mu = 0$. Damit gilt unter dieser Annahme

$$z = \frac{\Delta x}{\sqrt{2 \cdot \sigma^2}} \quad (6.162)$$

Mit diesen Vorüberlegungen kann der Hypothesentest grafisch dargestellt werden.

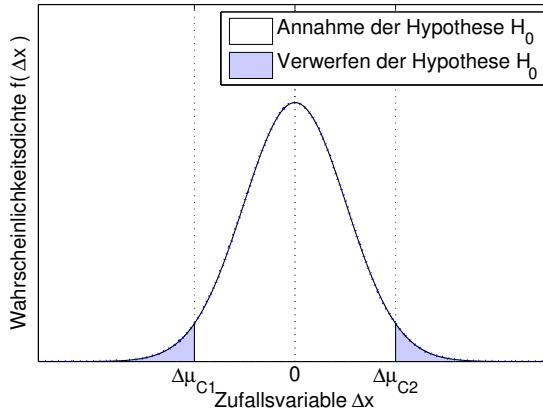


Bild 6.15: Grafische Darstellung des Hypothesentests zur Überwachung zweier Feuchtesensoren

Damit eine über die Stichprobe geschätzte Abweichung Δx mit einer spezifizierten Wahrscheinlichkeit zu der Normalverteilung gehört, muss dieser in dem Intervall $\Delta \mu_{C1} < \Delta x \leq \Delta \mu_{C2}$ liegen. Wird die Wahrscheinlichkeit dafür mit γ bezeichnet, gilt die Gleichung

$$P(\Delta \mu_{C1} < \Delta x \leq \Delta \mu_{C2}) = \gamma \quad (6.163)$$

Zur Berechnung der Grenzen $\Delta \mu_{C1}$ und $\Delta \mu_{C2}$ wird auf die standardnormalverteilte Zufallsvariable z zurückgegriffen. Mit der Definition der Zufallsvariable z in Gleichung (6.162) ist die Wahrscheinlichkeit γ , mit der die Variable z innerhalb des Intervalls $c_1 \dots c_2$ liegt, definiert als

$$\gamma = P(c_1 < z \leq c_2) = F(c_2) - F(c_1) \quad (6.164)$$

Da die Sensoren eine positive oder negative Drift aufweisen können, handelt es sich um einen symmetrischen Test. Damit ergeben sich die Konstanten c_1 und c_2 aus den Bedingungen

$$F(c_1) = \frac{1 - \gamma}{2} = \frac{\alpha}{2} \quad (6.165)$$

und

$$F(c_2) = 1 - \frac{1 - \gamma}{2} = 1 - \frac{\alpha}{2} \quad (6.166)$$

Auflösen nach c_1 und c_2 führt zu

$$c_1 = F^{-1}\left(\frac{\alpha}{2}\right) \quad (6.167)$$

und

$$c_2 = F^{-1}\left(1 - \frac{\alpha}{2}\right) \quad (6.168)$$

Durch Umformungen von Gleichung (6.164) ergibt sich ein Ausdruck für den Annahmebereich der Nullhypothese

$$\gamma = P\left(c_1 < \frac{\Delta x}{\sqrt{2 \cdot \sigma^2}} \leq c_2\right) = P\left(c_1 \cdot \sqrt{2 \cdot \sigma^2} < \Delta x \leq c_2 \cdot \sqrt{2 \cdot \sigma^2}\right) \quad (6.169)$$

Zur numerischen Berechnung der Grenzwerte müssen die Konstanten c_1 und c_2 sowie die Varianz σ^2 bestimmt werden. Aus Kapitel 2 ist bekannt, dass die Wahrscheinlichkeit, mit der ein funktionstüchtiger Sensor als defekt eingestuft wird, klein gehalten werden muss. Deshalb wird ein Signifikanzniveau von $\alpha = 10$ ppm gewählt. Daraus ergeben sich die Grenzen

$$c_1 = F^{-1}\left(\frac{\alpha}{2}\right) = -4.4172 \quad (6.170)$$

und

$$c_2 = F^{-1}\left(1 - \frac{\alpha}{2}\right) = 4.4172 \quad (6.171)$$

Im Rahmen von Freigabeerprobungen wurden Sensoren unterschiedlichen Tests unterzogen. Bei diesen Tests wurden funktionsfähige Sensoren im laufenden Betrieb bewertet. Eine Zusammenfassung der Messergebnisse zeigt Bild 6.16.

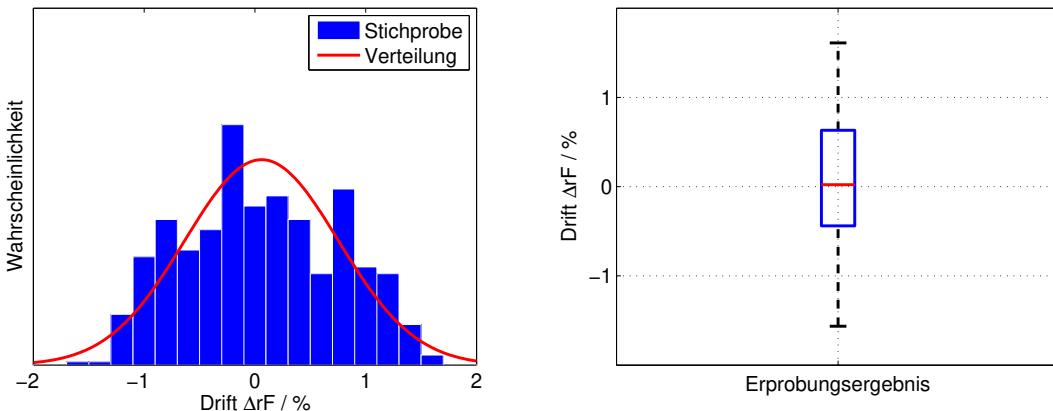


Bild 6.16: Statistische Zusammenfassung aller Driften bei einer Freigabeerprobung

Die Stichprobe weist eine Standardabweichung von $s = 0.6931\%$ auf. Die Messergebnisse werden als repräsentativ für Driften im Fahrzeugbetrieb angenommen. Es wird deshalb von einer Standardabweichung der Grundgesamtheit von $\sigma = 0.6931\%$ ausgegangen. Daraus ergeben sich für die Annahme der Hypothese die Grenzen

$$-4.4172 \cdot \sqrt{2} \cdot 0.6931 < \Delta x \leq 4.4172 \cdot \sqrt{2} \cdot 0.6931 \quad (6.172)$$

beziehungsweise

$$-4.3298\% < \Delta x \leq 4.3298\% \quad (6.173)$$

Liegt die Abweichung Δx bei einer Feuchte von $rF = 85\%$ außerhalb dieser Grenzen, ist der Sensor als defekt einzustufen.

6.7.3 Güte der Diagnosefunktion für Feuchtesensoren

Die Diagnosegrenzen werden unter Annahme der Nullhypothese $\Delta\mu = 0$ mit dem Fehler erster Art festgelegt. Im Betrieb der Diagnosefunktion bleiben die Diagnosegrenzen fest. Liegt der Stichprobenwert in den Grenzen

$$-4.3298\% < \Delta x \leq 4.3298\% \quad (6.174)$$

wird der Sensor als funktionsfähig eingestuft. Der Sensor wird als defekt eingestuft, wenn ein Stichprobenwert Δx unterhalb der Grenze $\Delta x \leq -4.3298$ oder oberhalb der Grenze $\Delta x > 4.3298$ liegt. Da der Stichprobenwert Δx von der wahren Differenz der Sensordriften $\Delta\mu$ abhängt, ist auch die Wahrscheinlichkeit der Einstufung des Sensors als defekter Sensor eine Funktion der Größe $\Delta\mu$. Sie ist definiert als

$$1 - \beta(\Delta\mu) = 1 - P\left(c_1 < \frac{\Delta x - \Delta\mu}{\sqrt{2 \cdot \sigma^2}} \leq c_2\right) \quad (6.175)$$

Da die Variable z mit den dargestellten Annahmen standardnormalverteilt ist, kann die Güte mit den bekannten Grenzen c_1 und c_2 sowie der Standardabweichung $\sigma = 0.6931\%$ berechnet werden aus

$$\begin{aligned} 1 - \beta(\Delta\mu) &= 1 - F\left(\frac{\Delta\mu_{C1} - \Delta\mu}{\sqrt{2 \cdot \sigma^2}}\right) + F\left(\frac{\Delta\mu_{C2} - \Delta\mu}{\sqrt{2 \cdot \sigma^2}}\right) \\ &= 1 - F\left(\frac{-4.3298\% - \Delta\mu}{0.9802}\right) + F\left(\frac{4.3298\% - \Delta\mu}{0.9802}\right) \end{aligned} \quad (6.176)$$

Bild 6.17 stellt die Güte des Hypothesentests mit der Alternativhypothese $\Delta\mu \neq 0$ als Funktion des alternativen Mittelwertes $\Delta\mu$ dar.

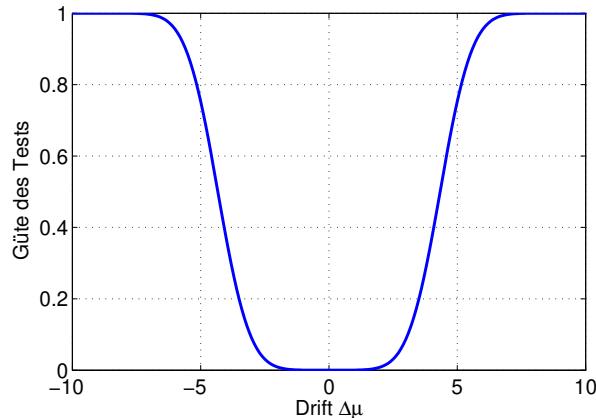


Bild 6.17: Darstellung der Gütfunktion für die Alternative $\Delta\mu \neq 0$

Mit steigendem Abstand $\Delta\mu$ von dem Wert $\Delta\mu = 0$ steigt die Wahrscheinlichkeit für die Einstufung als defekten Sensor. Eine Sensordrift von $\Delta\mu = -5\%$ wird mit einer Wahrscheinlichkeit von 75.2 % richtig diagnostiziert.

6.8 Literatur

- [Krey91] Kreyszig, Erwin: Statistische Methoden und ihre Anwendungen
4., unveränderter Nachdruck der 7. Auflage
Vandenhoeck & Ruprecht, Göttingen, 1991
- [Fahr06] Fahrmeir, Ludwig; Künstler, Rita; Pigeot, Iris; Tutz, Gerhard: Der Weg zur Datenanalyse
6. Auflage
Springer Berlin Heidelberg New York, 2006
- [Ross06] Ross, M. Sheldon: Statistik für Ingenieure und Naturwissenschaftler
3. Auflage
Spektrum Akademischer Verlag, München, 2006
- [Papu01] Papula, Lothar: Mathematik für Ingenieure und Naturwissenschaftler Band 3
4., verbesserte Auflage
Vieweg Teubner, Braunschweig / Wiesbaden, 2008

7 Beschreibende Statistik multivariater Daten

In den bislang dargestellten Datenreihen handelt es sich um eine Zufallsgröße x , die mehrfach ermittelt wird, sodass viele Messwerte x_n einer Größe x vorliegen. Die Größe wird entsprechend als eindimensionale oder univariate Zufallsvariable bezeichnet. In den Kapiteln 3 und 4 werden Methoden zur Beschreibung und Zusammenfassung dieser Daten vorgestellt.

Im Folgenden werden Aufgabenstellungen betrachtet, bei denen der Zusammenhang mehrerer Zufallsgrößen analysiert wird. Derartige Datensätze werden als multivariate Daten bezeichnet. Zur Vereinfachung der Darstellung wird zunächst von einem zweidimensionalen Datensatz ausgegangen, anschließend werden die gewonnenen Erkenntnisse auf mehrere Dimensionen verallgemeinert.

7.1 Darstellung und Charakterisierung von Datensätzen

Die Wahrscheinlichkeit für ein Ereignis A kann von mehreren Einflussgrößen abhängen. Zum Beispiel hängen die Emissionen eines Kraftfahrzeugs, die während eines Testzyklus erzeugt werden, von den unterschiedlichen Fahrzeugen mit unterschiedlichen Sensoren, Aktoren und Regelungsprozessen und von der Kraftstoffqualität ab. Jede einzelne Abhängigkeit kann durch eine Zufallsvariable beschrieben werden, die einen Datentyp und eine Verteilungsfunktion aufweist.

Wie bereits bei den univariaten Daten muss auch bei der zwei- und mehrdimensionalen Betrachtung zwischen den verschiedenen Daten- und Messtypen unterschieden werden. Zunächst wird auf die Darstellung ordinaler oder gruppierter Datensätze eingegangen. Im Anschluss daran wird eine Darstellungsform für stetige Datentypen vorgestellt.

7.1.1 Zweidimensionale gruppierte, ordinale oder diskrete Datensätze

Als Beispiel für einen zweidimensionalen Datensatz mit ordinalen Datentypen zeigt Tabelle 7.1 die Urliste einer Fertigungsstatistik. Für jedes gefertigte Teil wurde der Wochentag, an dem es gefertigt wurde, und die Fertigungsschicht, in der es gefertigt wurde, dokumentiert.

Tabelle 7.1: Fertigungsstatistik eines Produktes als Urliste

Teil Nr.	Wochentag	Fertigungsschicht
1	Mo	1
2	Mo	1
3	Mo	1
:	:	:
15240	Fr	3

Kontingenztafel

Das erste Teil wird am Montag in der ersten Schicht gefertigt, das zweite Teil ebenso und das letzte Teil dieses Beispiels wird am Freitag in der dritten Fertigungsschicht hergestellt. Die Darstellung ist vollständig, aber unübersichtlich. Wie bereits bei der Darstellung eindimensionaler Größen kann die dargestellte Urliste komprimiert werden. Tabelle 7.2 fasst die Fertigungsstatistik für ein Produkt als Funktion des Wochentages und der Fertigungsschicht tabellarisch zusammen.

Tabelle 7.2: Kontingenztafel der Fertigungsstatistik eines Produktes als absolute Häufigkeit

	Mo	Di	Mi	Do	Fr	Summe
Schicht 1	1008	991	1036	971	1109	5115
Schicht 2	1042	1159	1160	1098	1116	5575
Schicht 3	893	906	953	903	895	4550
Summe	2943	3056	3149	2972	3120	15240

Tabelle 7.2 stellt die absolute Häufigkeit eines Ereignisses, nämlich der Fertigung eines Produktes, als Funktion des Wochentages und der Fertigungsschicht dar. Dabei wird allgemein eine Variable als Variable x mit den Ausprägungen x_1, x_2, \dots, x_J bezeichnet. Die einzelnen Ausprägungen sind hier die unterschiedlichen Fertigungsschichten 1 - 3. Eine andere Variable y mit den Ausprägungen y_1, y_2, \dots, y_K repräsentiert entsprechend die Wochentage Mo - Fr. In die Tabelle werden die absoluten Häufigkeiten h_A eingetragen, mit der die Ereignisse eingetroffen sind. Die absolute Häufigkeit eines Bauteils mit den Merkmalen (x_j, y_k) wird mit $h_A(x_j, y_k)$ bezeichnet. Tabelle 7.2 wird Kontingenztafel genannt. Der Name weist auf die Kontingenz, also auf den Zusammenhang zwischen den Größen x und y , hin. Für die Darstellung der Daten in einer Kontingenztafel müssen beide Merkmale gruppiert, ordinal oder diskret sein.

Randhäufigkeit

Die Kontingenztafel wird mit den Reihen- und Spaltensummen der einzelnen Häufigkeiten ergänzt. Zum Beispiel ist die Summe aller in Schicht 1 gefertigter Bauteile 5115 und die Summe aller am Dienstag gefertigten Produkte 3056. Die Summe aller gefertigten Teile ergibt sich zu 15240. Die Reihen- und Spaltensummen werden als absolute Randhäufigkeiten bezeichnet. Die Spaltensummen werden abgekürzt mit

$$h_A(x) = h_A(x, y = \text{beliebig}) = \sum_{k=1}^K h_A(x, y_k) \quad (7.1)$$

und die Reihensummen entsprechend mit

$$h_A(y) = h_A(x = \text{beliebig}, y) = \sum_{j=1}^J h_A(x_j, y) \quad (7.2)$$

bezeichnet. Die sich ergebenden Randhäufigkeiten $h_A(x)$ sind die Häufigkeiten, mit der das jeweilige Merkmal x die Werte x_1, x_2, \dots, x_J annimmt, wenn das Merkmal y beliebig ist und unberücksichtigt bleibt. Entsprechendes gilt für die Reihensummen. Beide Randhäufigkeiten hängen nur noch von einer Variablen ab.

Die Summe aller absoluten Häufigkeiten N ergibt sich aus

$$N = \sum_{k=1}^K \sum_{j=1}^J h_A(x_j, y_k) = \sum_{j=1}^J \sum_{k=1}^K h_A(x_j, y_k) \quad (7.3)$$

Allgemein ergibt sich für zweidimensionale Datensätze damit eine Kontingenztabelle, wie sie in Tabelle 7.3 dargestellt ist.

Tabelle 7.3: Allgemeine Darstellung einer Kontingenztafel mit absoluter Häufigkeit

	y₁	y₂	...	y_K	Summe
x₁	$h_A(x_1, y_1)$	$h_A(x_1, y_2)$...	$h_A(x_1, y_K)$	$h_A(x_1)$
x₂	$h_A(x_2, y_1)$	$h_A(x_2, y_2)$...	$h_A(x_2, y_K)$	$h_A(x_2)$
...
x_J	$h_A(x_J, y_1)$	$h_A(x_J, y_2)$...	$h_A(x_J, y_K)$	$h_A(x_J)$
Summe	$h_A(y_1)$	$h_A(y_2)$...	$h_A(y_K)$	N

Entsprechend der Darstellungen zu univariaten Datensätzen kann alternativ die relative Häufigkeit von Ereignissen dargestellt werden. Dabei werden die einzelnen Häufigkeiten durch die Gesamtanzahl N dividiert.

$$h(x_j, y_k) = \frac{h_A(x_j, y_k)}{N} \quad (7.4)$$

Für das Beispiel aus Tabelle 7.2 ergibt sich die in Tabelle 7.4 dargestellte Kontingenztafel der relativen Häufigkeit.

Tabelle 7.4: Kontingenztafel der relativen Häufigkeit für die Fertigungsstatistik eines Produktes

	Mo	Di	Mi	Do	Fr	Summe
Schicht 1	0.066	0.065	0.068	0.064	0.073	0.336
Schicht 2	0.068	0.076	0.076	0.072	0.073	0.365
Schicht 3	0.059	0.059	0.063	0.059	0.059	0.299
Summe	0.193	0.200	0.207	0.195	0.205	1

Die Randsummen der relativen Häufigkeit werden als relative Randhäufigkeiten bezeichnet. Analog zu den absoluten Randhäufigkeiten $h_A(x)$ ergibt sich für die relativen Randhäufigkeiten $h(x)$

$$h(x) = h(x, y = \text{beliebig}) = \sum_{k=1}^K h(x, y_k) \quad (7.5)$$

und

$$h(y) = h(x = \text{beliebig}, y) = \sum_{j=1}^J h(x_j, y) \quad (7.6)$$

Die allgemeinen Bezeichnungen für relative Häufigkeiten sind analog zur absoluten Häufigkeit in Tabelle 7.5 zusammengefasst.

Tabelle 7.5: Allgemeine Darstellung einer Kontingenztabelle mit relativer Häufigkeit

	y₁	y₂	...	y_K	Summe
x₁	$h(x_1, y_1)$	$h(x_1, y_2)$...	$h(x_1, y_K)$	$h(x_1)$
x₂	$h(x_2, y_1)$	$h(x_2, y_2)$...	$h(x_2, y_K)$	$h(x_2)$
...
x_J	$h(x_J, y_1)$	$h(x_J, y_2)$...	$h(x_J, y_K)$	$h(x_J)$
Summe	$h(y_1)$	$h(y_2)$...	$h(y_K)$	1

Grafische Darstellung

Die grafische Darstellung der Daten kann in dreidimensionalen Säulendiagrammen erfolgen. Bild 7.1 zeigt ein dreidimensionales Säulendiagramm für die Fertigungsstatistik aus Tabelle 7.5 als relative Häufigkeit.

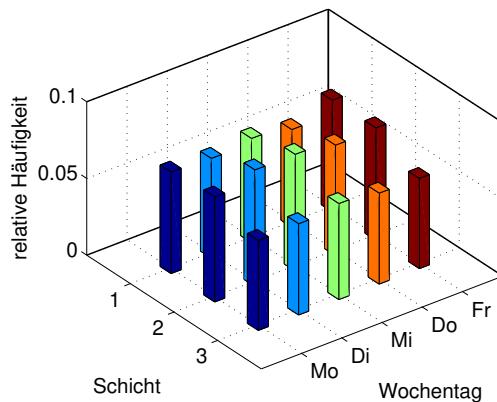


Bild 7.1: Grafische Darstellung der Fertigungsstatistik eines Produktes als relative Häufigkeit

Das dreidimensionale Säulendiagramm wird mit MATLAB mit dem folgenden Programmabschnitt erstellt.

```

1 % Messwerte einlesen
2 load Fertigung.mat;
3
4 % Grafische Darstellung der Fertigungsdaten als dreidimensionales Sä
5   ulendiagramm
6 figure(1);
7 bar3(ZRel, 0.25);
```

Ist die Summenhäufigkeit einzelner Merkmalsausprägungen interessant, kann der Datensatz als gestapeltes Säulendiagramm dargestellt werden. An dieser Darstellung kann direkt abgelesen werden, welche Randhäufigkeit sich für eine bestimmte Merkmalsausprägung ergibt. Bild 7.2 stellt die absolute Randhäufigkeit der Fertigungsschicht dar.

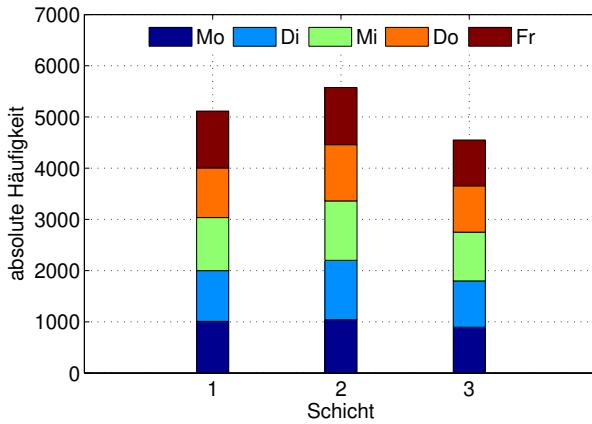


Bild 7.2: Grafische Darstellung der absoluten Randhäufigkeit der Fertigungsschicht eines Produktes

Die Erzeugung eines gestapelten Säulendiagramms mit MATLAB erfolgt in dem folgenden Programmabschnitt.

```

1 % Messwerte einlesen
2 load Fertigung.mat;
3
4 % Grafische Darstellung der Fertigungsdaten als zweidimensionales Sä
5   ulendiagramm
6 figure(1);
7 bar(Z,0.25 , 'stacked');

```

Mit dieser Darstellung kann auch der Begriff der bedingten Wahrscheinlichkeit interpretiert werden. Zum Beispiel ergibt sich die relative Häufigkeit, mit der ein Produkt am Mittwoch gefertigt wurde, unter der Bedingung, dass es in Schicht 3 gefertigt wurde, aus der relativen Aufteilung der Säule Schicht 3. Diese Bedingungen können als bedingte Wahrscheinlichkeit aufgefasst werden, die sich berechnen lässt als

$$h(x|y) = \frac{h(x,y)}{h(x = \text{beliebig}, y)} = \frac{h(x,y)}{h(y)} \quad (7.7)$$

mit $x = \text{Mi}$ und $y = \text{Schicht 3}$. Die bedingte Wahrscheinlichkeit schränkt die Grundmenge möglicher Ereignisse aus $y = \text{Schicht 3}$ ein. Da dabei x beliebig ist, ist die Wahrscheinlichkeit für $y = \text{Schicht 3}$ gerade die Randhäufigkeit $h(y)$. Für das Beispiel ergibt sich

$$h(\text{Mi}|3) = \frac{0.063}{0.297} = 0.2121 \quad (7.8)$$

7.1.2 Zweidimensionale stetige Datensätze

Auch bei stetigen Größen werden die Beobachtungen zunächst in einer Urliste dokumentiert. Tabelle 7.6 zeigt ein Beispiel zur Abhängigkeit von Fertigungsdefekten als Funktion der Umgebungstemperatur. Dabei wurde die Anzahl von Fertigungsdefekten an verschiedenen Tagen in Abhängigkeit der Tagstemperatur untersucht.

Tabelle 7.6: Daten zur Abhängigkeit von Fertigungsdefekten als Funktion der Umgebungstemperatur

Tag	1	2	3	4	5	6
Temperatur T / °C	24.2	22.7	30.5	28.6	25.5	32.0
Anzahl von Defekten D	25	31	36	33	19	24

Tag	7	8	9	10	11	12
Temperatur T / °C	28.6	26.5	25.3	26.0	24.4	24.8
Anzahl von Defekten D	27	25	16	14	22	23

Tag	13	14	15	16	17	18
Temperatur T / °C	24.8	20.6	25.1	21.4	23.7	25.2
Anzahl von Defekten D	20	25	25	23	27	30

Grafische Darstellung

Die Temperatur ist eine stetige Größe, es existiert zumindest physikalisch gesehen keine Quantisierung. Bei diesem stetigen Datentyp versagt die Darstellung mit Kontingenztabellen, da unendlich viele Merkmalsausprägungen existieren und die Kontingenztabelle damit unendlich groß würde. Eine Möglichkeit wäre, die Temperaturdaten zu gruppieren, dann könnte wie in Abschnitt 7.1.1 verfahren werden. Allerdings würden durch die Gruppierung gegebenenfalls wesentliche Informationen verloren gehen.

Eine zweite Möglichkeit, den Zusammenhang zwischen Temperatur T und Anzahl von Defekten D aufzuzeigen, ist ein zweidimensionales Streudiagramm, das in Bild 7.3 dargestellt ist.

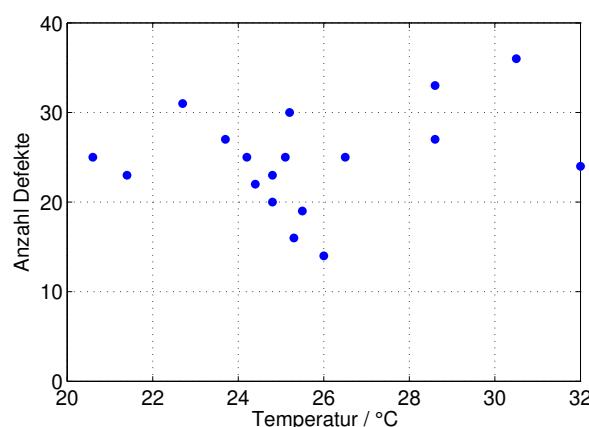


Bild 7.3: Darstellung der Wertepaare aus Tabelle 7.6 in einem Streudiagramm

Jeder Punkt des Streudiagramms ist ein Datenpunkt der Tabelle. Seine Lage wird durch die beiden Merkmale Temperatur und Anzahl von Defekten festgelegt.

Randverteilungen stetiger Datensätze

Bei der Beschreibung univariater Häufigkeiten $h(x)$ wird in Kapitel 3 zur Beschreibung stetiger Daten die Summenhäufigkeit $H(x)$ eingeführt.

$$H(x) = \sum_{\xi=-\infty}^x h(\xi) \quad (7.9)$$

Bei stetigen multivariaten Datensätzen wird die Summenhäufigkeit verwendet, um die Verteilung der Randhäufigkeit zu beschreiben. Dabei wird davon ausgegangen, dass alle Variablen bis auf eine Variable beliebig sind. Zum Beispiel ergibt sich für die Randverteilung $H(x)$ eines zweidimensionalen Datensatzes

$$H(x) = \sum_{\xi=-\infty}^x \sum_{y=-\infty}^{\infty} h(\xi, y) \quad (7.10)$$

Für die Daten aus Tabelle 7.6 ergeben sich die in Bild 7.4 dargestellten Verteilungen der Randhäufigkeiten der Temperatur $H(T)$ und der Anzahl von Defekten $H(D)$.

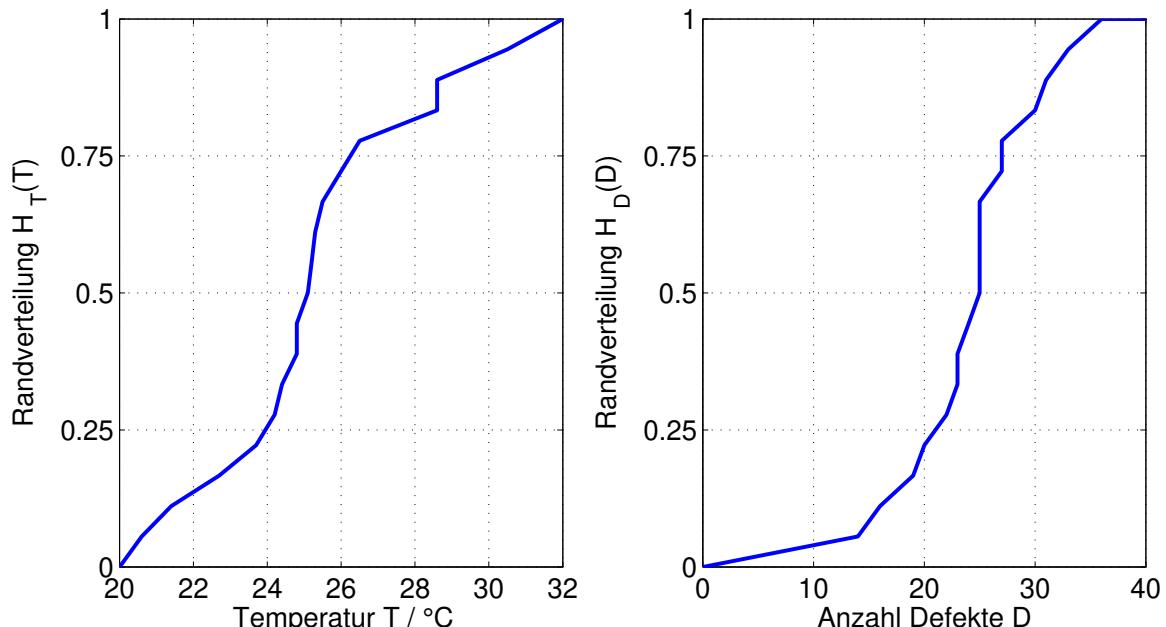


Bild 7.4: Randverteilungen der Temperatur T und der Anzahl von Defekten D

Die Randverteilung $H(T)$ beginnt für sehr kleine Temperaturen bei der Wahrscheinlichkeit 0 und endet für sehr große Werte bei 1, die Aussage gilt sinngemäß für die Randverteilung $H(D)$.

7.1.3 Darstellung und Charakterisierung multivariater Datensätze

Zweidimensionale Datensätze lassen sich wegen der räumlichen Vorstellung noch vergleichsweise einfach darstellen. Auch bei dreidimensionalen Datensätzen ergeben sich noch Möglichkeiten der grafischen Darstellung. Dazu wird ein Datensatz analysiert, bei dem die Ausbeute eines chemischen Prozesses als Funktion der Temperatur und der Katalysatorkonzentration dargestellt wird. Für die Messwertaufnahme wurden alle anderen Parameter konstant gehalten.

Soll die Fertigungsausbeute A in Abhängigkeit von Temperatur T und Katalysatorkonzentration K dargestellt werden, kann das in Form eines dreidimensionalen Streudiagramms erfolgen. Dabei wird die räumliche Darstellung in die Ebene projiziert.

Tabelle 7.7: Urliste einer Messreihe eines chemischen Prozesses

Nr.	Temperatur T / °C	Katalysatorkonzentration K / %	Ausbeute A / %
1	130	0.3	67.47
2	140	0.5	84.27
3	120	0.1	54.78
4	120	0.1	54.13
5	120	0.5	73.82
6	130	0.3	66.18
7	140	0.5	83.05
8	140	0.1	61.86
9	140	0.1	61.33
10	120	0.5	71.20
11	140	0.1	60.41
12	130	0.3	69.08
13	120	0.1	51.06
14	120	0.5	84.95
15	120	0.5	71.31
16	120	0.1	53.67
17	140	0.5	83.50
18	120	0.5	71.87
19	140	0.1	61.78
20	130	0.3	66.23

Bild 7.5 stellt die Ausbeute A in Abhängigkeit der Temperatur T und Katalysatorkonzentration K als Streudiagramm dar.

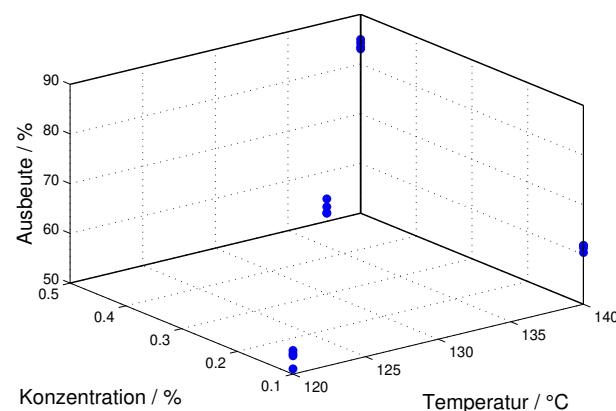


Bild 7.5: Ausbeute in Abhängigkeit der Temperatur und der Katalysatorkonzentration als Streudiagramm

Ein dreidimensionales Streudiagramm mit MATLAB wird über den folgenden Programmabschnitt erstellt.

```

1 % Messwerte einlesen
2 load ChemischeIndustrie.mat;
3
4 % Grafische Darstellung der Messdaten als Streudiagramm
5 figure(1);
6
7 scatter3(values (: ,1) ,values (: ,2) ,values (: ,3) , 'ob' , 'filled' );

```

Bereits die Darstellung von dreidimensionalen Datensätzen führt zu Messpunkten im Streudiagramm, deren Lagen wegen der Projektion nicht mehr eindeutig erkennbar sind. Steigt die Dimension des Datensatzes auf einen Wert größer drei, ist auch eine quasi-räumliche Darstellung der Daten nicht mehr möglich, sodass hier andere Wege der grafischen Darstellung gefunden werden müssen.

Eine einfache Darstellungsmöglichkeit mehrdimensionaler Datensätze besteht darin, jeweils für zwei Größen ein Streudiagramm zu bilden. Es ergibt sich eine Matrix von Streudiagrammen, bei der der Zusammenhang zwischen zwei Größen dargestellt ist. Alle übrigen Größen werden nicht eingeschränkt, sind also beliebig. Auf der Hauptdiagonalen der Matrix sind die Größen bezeichnet und die verwendeten Einheiten sind dargestellt. Die Achsenbeschriftung befindet sich jeweils am Rand der Matrix.

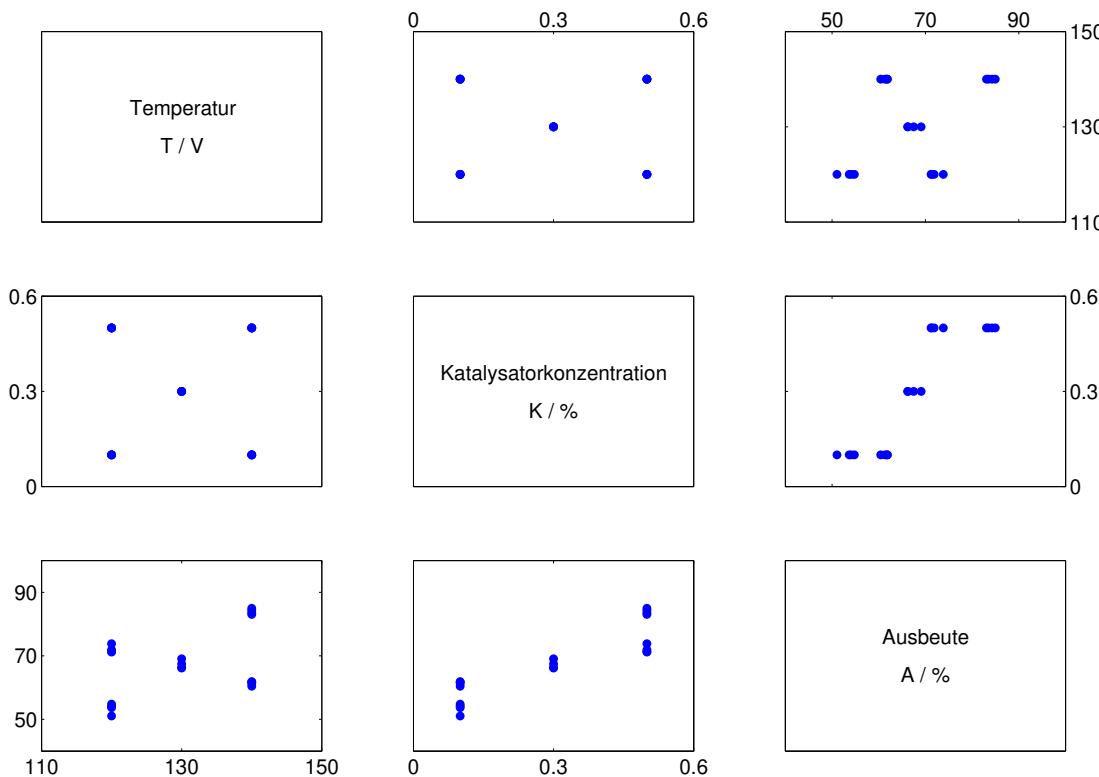


Bild 7.6: Streudiagramm-Matrix der Messwerte eines chemischen Prozesses

Bild 7.6 stellt die Streudiagramm-Matrix der Messwerte eines chemischen Prozesses mit den Werten aus Tabelle 7.7 dar. Die Matrix ist symmetrisch zur Hauptdiagonalen. An der grafischen Darstellung kann abgelesen werden, welche Kombinationen von Temperatur und Katalysatorkonzentration verwendet wurden. Es wird außerdem deutlich, dass die Ausbeute von der Temperatur und der Katalysatorkonzentration abhängig ist. Sowohl eine Erhöhung der Temperatur als auch der Katalysatorkonzentration

könnte in dem untersuchten Parameterbereich somit zur Steigerung der Ausbeute verwendet werden.

Diese Art der Darstellung kann durch die relative Häufigkeitsverteilung der einzelnen Stichprobengrößen erweitert werden. Die Häufigkeitsverteilungen werden auf der Hauptdiagonale platziert. Dadurch werden in dem Diagramm mehr Informationen dargestellt, allerdings wirkt die Darstellung weniger übersichtlich und die Zuordnung der Daten zu den Größen ist weniger deutlich. Bild 7.7 stellt diese Variante der Streudiagramm-Matrix dar.

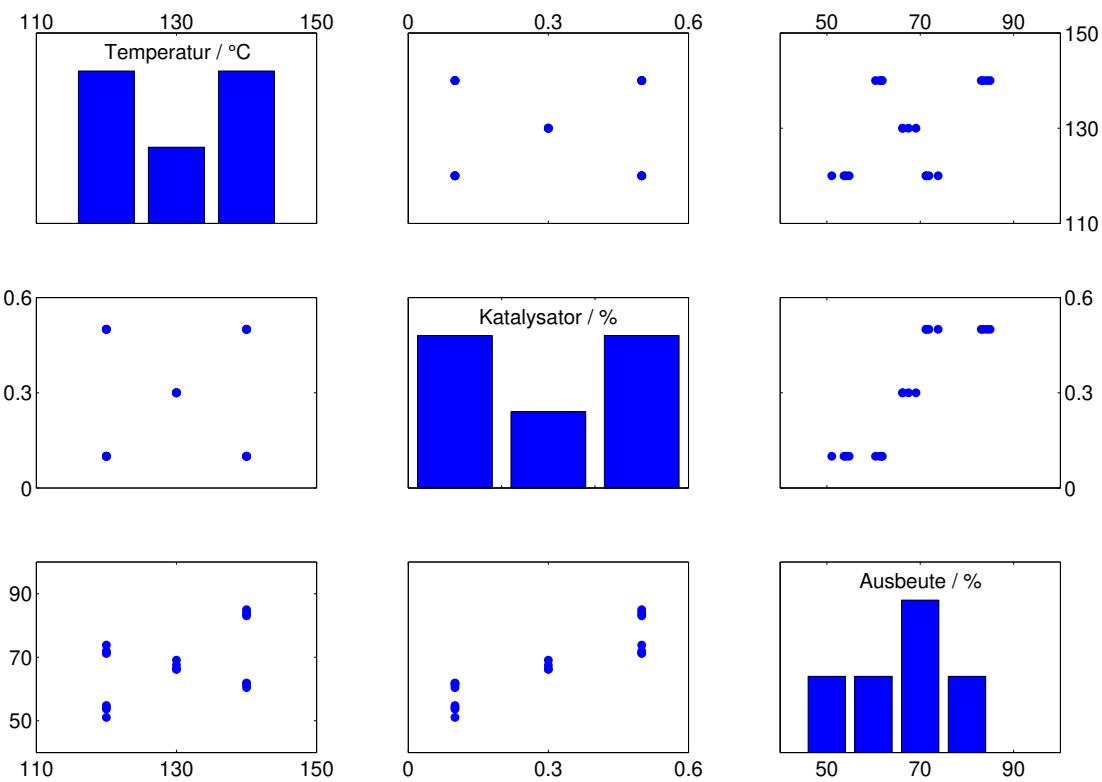


Bild 7.7: Streudiagramm-Matrix der Messwerte eines chemischen Prozesses mit Häufigkeitsverteilung der Stichprobengrößen

Die paarweise Streudiagramm-Matrix kann für m-dimensionale Stichproben entsprechend erweitert werden. In MATLAB können sie durch Anwendung eines sogenannten subplot realisiert werden. Alternativ kann die Streudiagramm-Matrix mit dem MATLAB-Befehl gplotmatrix dargestellt werden.

```

1 % Messwerte einlesen
2 load ChemischeIndustrie.mat;
3
4 % Grafische Darstellung der Messdaten als Streudiagramm
5 figure(2));
6
7 gplotmatrix(values);

```

7.1.4 MATLAB- und Python-Befehle Darstellung multivariater Datensätze

Der Vollständigkeit halber sind in Tabelle 7.8 MATLAB-Befehle zur Darstellung multivariater Datensätze zusammengefasst.

Tabelle 7.8: Zusammenfassung von MATLAB-Befehlen zur Darstellung multivariater Datensätze

MATLAB-Befehl	Funktionsbeschreibung
bar3(X)	Darstellung eines dreidimensionalen Datensatzes als räumliches Balkendiagramm
scatter3(X)	Darstellung eines dreidimensionalen Datensatzes als räumliches Streudiagramm
gplotmatrix(X)	Darstellung eStreudiagramm-Matrix für die Daten X

In erreicht, der eigentliche Befehl für ein Balkendiagramm oder ein Streudiagramm ist identisch zu der zweidimensionalen Darstellung. Der Befehl gplotmatrix ist in der Bibliothek pandas.plotting verfügbar. Alternativ können diese Grafiken über eine Matrix von Diagrammen erzeugt werden. Tabelle 7.9 sind die entsprechenden Python-Befehle aus der Bibliothek matplotlib.pyplot zur Darstellung multivariater Datensätze zusammengefasst. Dabei wird die räumliche Darstellung über den Parameter projection = '3d' erreicht, der eigentliche Befehl für ein Balkendiagramm oder ein Streudiagramm ist identisch zu der zweidimensionalen Darstellung. Der Befehl gplotmatrix ist in der Bibliothek pandas.plotting verfügbar. Alternativ können diese Grafiken über eine Matrix von Diagrammen erzeugt werden.

Tabelle 7.9: Zusammenfassung von Python-Befehlen zur Darstellung multivariater Datensätze

Python-Befehl	Funktionsbeschreibung
bar(X)	Darstellung eines dreidimensionalen Datensatzes als räumliches Balkendiagramm
scatter(X)	Darstellung eines dreidimensionalen Datensatzes als räumliches Streudiagramm
scatter_matrix(X)	Darstellung eStreudiagramm-Matrix für die Daten X

7.2 Kenngrößen multivariater Stichproben

Wie bei univariaten Stichproben können multivariate Datensätze durch Kenngrößen zusammenfassend beschrieben werden. Dabei wird analog zu der univariaten Berechnung vorgegangen. Im Folgenden wird der arithmetische Mittelwert als Lagekenngröße vorgestellt. Die Streuung der Größen und die Abhängigkeiten der einzelnen Größen untereinander werden mit der Kovarianz oder der Korrelation gekennzeichnet.

7.2.1 Arithmetischer Mittelwertsvektor einer Stichprobe

In Kapitel 3 wird der arithmetische Mittelwert einer univariaten Stichprobe x_1, \dots, x_N definiert zu

$$\bar{x} = \frac{x_1 + \dots + x_N}{N} = \frac{1}{N} \cdot \sum_{n=1}^N x_n \quad (7.11)$$

Für die Beschreibung einer M-dimensionalen Stichprobe mit M eindimensionalen Zufallsvariablen $\underline{x}_1, \dots, \underline{x}_M$ wird die Vektorschreibweise angewendet. Die Matrix \mathbf{X} aller Messwerte ergibt sich aus den N Messwerten jeder Größe.

$$\mathbf{X} = \begin{pmatrix} \underline{x}_1 & \dots & \underline{x}_M \end{pmatrix} = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1M} \\ x_{21} & x_{22} & \cdots & x_{2M} \\ \vdots & \vdots & & \vdots \\ x_{N1} & x_{N2} & \cdots & x_{NM} \end{pmatrix} \quad (7.12)$$

Der arithmetische Mittelwert einer M-dimensionalen Stichprobe wird damit auf die Berechnung M einzelner Mittelwerte zurückgeführt.

$$\underline{\bar{x}}^T = \begin{pmatrix} \bar{x}_1 & \dots & \bar{x}_M \end{pmatrix} = \begin{pmatrix} \frac{1}{N} \cdot \sum_{n=1}^N x_{n1} & \dots & \frac{1}{N} \cdot \sum_{n=1}^N x_{nM} \end{pmatrix} \quad (7.13)$$

Für die Stichprobe eines chemischen Prozesses aus Tabelle 7.7 ergibt sich ein Vektor der Mittelwerte von

$$\underline{\bar{x}}^T = \begin{pmatrix} 130.00 & 0.30 & 67.60 \end{pmatrix} \quad (7.14)$$

In MATLAB wird der Mittelwertsvektor der Messreihe mit folgender Befehlssequenz berechnet:

```

1 % Messwerte einlesen
2 load ChemischeIndustrie.mat;
3
4 % Berechnung der Kovarianzmatrix \
5 xquer = mean(values);
```

Bei einer mittleren Temperatur von 130 °C und einer mittleren Katalysatorkonzentration von 0.3 % ist eine Ausbeute von 67.6 % zu erwarten.

7.2.2 Kovarianzmatrix einer Stichprobe

Um einen Zusammenhang mehrerer Zufallsvariablen quantitativ zu beschreiben, kann auf Basis der vorliegenden Stichprobe die Kovarianz berechnet werden. Zum besseren Verständnis und aus Gründen der Darstellung wird zunächst wieder die Kovarianz für eine zweidimensionale Stichprobe eingeführt. Die dabei gewonnenen Erkenntnisse werden dann auf mehrdimensionale Stichproben erweitern.

Kovarianz einer zweidimensionalen Stichprobe

Die zweidimensionale Stichprobe besteht aus N Wertepaaren der Form

$$\begin{pmatrix} \underline{x} & \underline{y} \end{pmatrix} = \begin{pmatrix} x_1 & y_1 \\ \vdots & \vdots \\ x_N & y_N \end{pmatrix} \quad (7.15)$$

Um eine Kennzahl zu berechnen, die den Zusammenhang zwischen den einzelnen Datenpaaren angibt, werden zunächst die jeweiligen Mittelwerte der x-Werte der Stichprobe

$$\bar{x} = \frac{1}{N} \cdot (x_1 + x_2 + \dots + x_N) = \frac{1}{N} \cdot \sum_{n=1}^N x_n \quad (7.16)$$

und der y-Werte der Stichprobe

$$\bar{y} = \frac{1}{N} \cdot (y_1 + y_2 + \dots + y_N) = \frac{1}{N} \cdot \sum_{n=1}^N y_n \quad (7.17)$$

berechnet. Für das Datenpaar mit dem Index n ist damit die Abweichung vom Mittelwert gegeben durch

$$\Delta x_n = x_n - \bar{x} \quad (7.18)$$

beziehungsweise

$$\Delta y_n = y_n - \bar{y} \quad (7.19)$$

Damit wird der zugrunde liegende Datensatz zentriert, was in Bild 7.8 verdeutlicht wird.

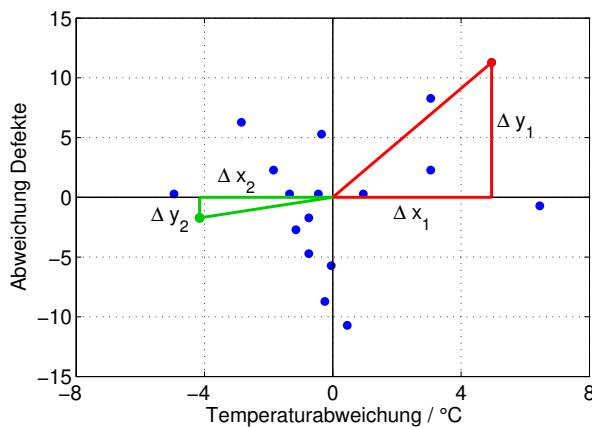


Bild 7.8: Darstellung der Abweichung der Wertepaare aus Tabelle 7.6 von ihrem jeweiligen Mittelwert in einem Streudiagramm

Für große Werte von x_n ist Δx_n positiv, während Δx_n für kleine Werte von x_n negativ ist. Diese Aussagen gelten für die Werte von y_n entsprechend. Wenn also große x-Werte gewöhnlich mit großen y-Werten und kleine x-Werte gewöhnlich mit kleinen y-Werten auftreten, dann sind die Vorzeichen von Δx_n und Δy_n gleich. Damit wird aber das Produkt der Abweichungen vom Mittelwert größer als null.

$$\Delta y_n = y_n - \bar{y} \quad (7.20)$$

Für diesen Fall wird auch die Summe aller Produkte

$$\sum_{n=1}^N ((x_n - \bar{x}) \cdot (y_n - \bar{y})) > 0 \quad (7.21)$$

größer als null sein. Liegen bei einer Stichprobe große x-Werte gewöhnlich mit kleinen y-Werten und kleine x-Werte mit großen y-Werten zusammen, dann sind die Vorzeichen von Δx_n und Δy_n gewöhnlich unterschiedlich. Damit wird das Produkt der Abweichung vom Mittelwert kleiner null sein. Entsprechendes gilt für deren Summe. Je größer der Betrag der Summe ist, desto klarer ist die Aussage eines linearen Zusammenhangs der Zufallsgrößen x und y.

Aus der Summe aller Produkte in Gleichung (7.21) ergibt sich durch die Normierung mit $N - 1$ die Kovarianz der Größen x und y.

$$s_{xy} = \frac{1}{N-1} \cdot \sum_{n=1}^N ((x_n - \bar{x}) \cdot (y_n - \bar{y})) \quad (7.22)$$

Der Wert der Kovarianz ist entsprechend den obigen Überlegungen positiv, wenn x und y tendenziell einen gleichsinnigen linearen Zusammenhang aufweisen. Dagegen ist die Kovarianz negativ, wenn x und y einen gegensinnig linearen Zusammenhang besitzen. Bei einer Kovarianz von $s_{xy} = 0$ sind die Größen x und y voneinander unabhängig.

Kovarianzmatrix einer multivariaten Stichprobe

Durch die Darstellung multivariater Größen in Streudiagramm-Matrizen wird die mehrdimensionale Darstellung in eine zweidimensionale Darstellung transformiert. Ganz analog wird bei der Kovarianz

verfahren. Es ergibt sich die Kovarianzmatrix \mathbf{S} .

$$\begin{aligned} \mathbf{S} &= \begin{pmatrix} s_1^2 & \dots & s_{1M} \\ \vdots & \ddots & \vdots \\ s_{M1} & \dots & s_M^2 \end{pmatrix} \\ &= \begin{pmatrix} \frac{1}{N-1} \cdot \sum_{n=1}^N (x_{n1} - \bar{x}_1)^2 & \dots & \frac{1}{N-1} \cdot \sum_{n=1}^N (x_{n1} - \bar{x}_1) \cdot (x_{nM} - \bar{x}_M) \\ \vdots & & \vdots \\ \frac{1}{N-1} \cdot \sum_{n=1}^N (x_{nM} - \bar{x}_M) \cdot (x_{n1} - \bar{x}_1) & \dots & \frac{1}{N-1} \cdot \sum_{n=1}^N (x_{nM} - \bar{x}_M)^2 \end{pmatrix} \end{aligned} \quad (7.23)$$

Gleichung (7.23) kann in Vektorschreibweise überführt werden. Mit der Matrix \mathbf{X} von Stichprobenwerten

$$\mathbf{X} = \begin{pmatrix} \underline{x}_1 & \dots & \underline{x}_M \end{pmatrix} = \begin{pmatrix} x_{11} & x_{12} & \dots & x_{1M} \\ x_{21} & x_{22} & \dots & x_{2M} \\ \vdots & \vdots & & \vdots \\ x_{N1} & x_{N2} & \dots & x_{NM} \end{pmatrix} \quad (7.24)$$

und dem Vektor der Mittelwerte

$$\underline{\bar{x}}^T = \begin{pmatrix} \bar{x}_1 & \dots & \bar{x}_M \end{pmatrix} = \begin{pmatrix} \frac{1}{N} \cdot \sum_{n=1}^N x_{n1} & \dots & \frac{1}{N} \cdot \sum_{n=1}^N x_{nM} \end{pmatrix} \quad (7.25)$$

kann Gleichung (7.23) umgeschrieben werden zu

$$\begin{aligned} \mathbf{S} &= \begin{pmatrix} \frac{1}{N-1} \cdot \sum_{n=1}^N (x_{n1} - \bar{x}_1)^2 & \dots & \frac{1}{N-1} \cdot \sum_{n=1}^N (x_{n1} - \bar{x}_1) \cdot (x_{nM} - \bar{x}_M) \\ \vdots & & \vdots \\ \frac{1}{N-1} \cdot \sum_{n=1}^N (x_{nM} - \bar{x}_M) \cdot (x_{n1} - \bar{x}_1) & \dots & \frac{1}{N-1} \cdot \sum_{n=1}^N (x_{nM} - \bar{x}_M)^2 \end{pmatrix} \\ &= \frac{1}{N-1} \cdot \begin{pmatrix} x_{11} - \bar{x}_1 & \dots & x_{N1} - \bar{x}_1 \\ \vdots & & \vdots \\ x_{1M} - \bar{x}_M & \dots & x_{NM} - \bar{x}_M \end{pmatrix} \cdot \begin{pmatrix} x_{11} - \bar{x}_1 & \dots & x_{1M} - \bar{x}_M \\ \vdots & & \vdots \\ x_{N1} - \bar{x}_1 & \dots & x_{NM} - \bar{x}_M \end{pmatrix} \\ &= \frac{1}{N-1} \cdot \left(\mathbf{X} - \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix} \cdot \underline{\bar{x}}^T \right)^T \cdot \left(\mathbf{X} - \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix} \cdot \underline{\bar{x}}^T \right) \end{aligned} \quad (7.26)$$

Auf der Hauptdiagonalen der Matrix \mathbf{S} liegen die Varianzen der einzelnen Zufallsgrößen. Die Nebenelemente entsprechen den Kovarianzen.

Als Beispiel wird der Datensatz des chemischen Prozesses aus Tabelle 7.7 aufgegriffen. Für die Stichprobe berechnet sich die Kovarianzmatrix zu

$$\mathbf{S} = \begin{pmatrix} 84.21 & 0 & 41.75 \\ 0 & 0.034 & 1.74 \\ 41.75 & 1.74 & 29.67 \end{pmatrix} \quad (7.27)$$

Mit MATLAB kann die Kovarianzmatrix direkt mit dem Befehl cov durchgeführt werden.

```

1 % Messwerte einlesen
2 load ChemischeIndustrie.mat;
3
4 % Berechnung der Kovarianzmatrix
5 S = cov(values);

```

In der Kovarianzmatrix aus Gleichung (7.27) ist zu erkennen, dass die Kovarianz zwischen der Temperatur T und der Katalysatorkonzentration K den Wert 0 hat. Die Größen sind somit unabhängig voneinander. Dies ist ein Ergebnis der statistischen Versuchsplanung, die bei der Erfassung des Datensatzes durchgeführt wurde.

7.2.3 MATLAB- und Python-Befehle zur Berechnung von Kenngrößen multivariater Stichproben

Tabelle 7.10 fasst die MATLAB-Befehle zur Berechnung von Kenngrößen multivariater Stichproben zusammen.

Tabelle 7.10: MATLAB-Befehle zur Berechnung von Kenngrößen multivariater Stichproben

MATLAB-Befehl	Funktionsbeschreibung
<code>mean(X)</code>	Mittelwerte der Spaltenvektoren von dem Datensatz X , jede Spalte repräsentiert eine Größe
<code>cov(X)</code>	Kovarianzmatrix S zum Datensatz X , jede Spalte repräsentiert eine Größe

Tabelle 7.11 fasst die MATLAB-Befehle zur Berechnung von Kenngrößen multivariater Stichproben zusammen.

Tabelle 7.11: Python-Befehle der Bibliothek numpy zur Berechnung von Kenngrößen multivariater Stichproben

Python-Befehl	Funktionsbeschreibung
<code>np.mean(X)</code>	Mittelwerte der Spaltenvektoren von dem Datensatz X , jede Spalte repräsentiert eine Größe
<code>np.cov(X)</code>	Kovarianzmatrix S zum Datensatz X , jede Spalte repräsentiert eine Größe

7.3 Anwendungsbeispiel: Schwindung beim Spritzgießen

Spritzgießen ist ein Verfahren, das in der Kunststoffverarbeitung zur Fertigung von Formteilen in großer Stückzahl eingesetzt wird. Dazu wird mit einer Spritzgießmaschine der jeweilige Kunststoff in einer Spritzeinheit plastifiziert und in ein Spritzgießwerkzeug eingespritzt. Der Hohlraum des Werkzeugs bestimmt die Form und die Oberflächenstruktur des fertigen Teils [Wiki11]. Mit Spritzgussprozessen lassen sich auch komplexe Werkstücke mit hoher Qualität fertigen. Bild 7.9 zeigt das Kunststoffgehäuse eines Luftmassenmessers als Beispiel für ein komplexes Kunststoffteil, an das extreme Anforderungen hinsichtlich Maßhaltigkeit gestellt werden.



Bild 7.9: Kunststoffgehäuse eines Luftmassenmessers der Robert Bosch GmbH [BOSC11]

Die Werkzeugmaße müssen so ausgelegt werden, dass Formteile mit den gewünschten späteren Endmaßen gefertigt werden können. Dabei muss die Schwindung des Bauteils berücksichtigt werden. Die Schwindung ist zwar in erster Linie eine Werkstoffeigenschaft, sie wird darüber hinaus aber auch bestimmt durch die Gestalt und Wanddicke des Spritzgussteils sowie durch die Verarbeitungsbedingungen. Das Zusammenwirken dieser verschiedenen Faktoren macht eine exakte Vorhersage der Schwindung meist sehr schwierig. Zur Ermittlung von praxisrelevanten Schwindungsmaßen hat sich ein Testkästchen bewährt, das in Bild 7.10 dargestellt ist. Ausgewertet wird meist die Länge A als Maß für die Schwindung des Kästchenbodens. [BASF11]

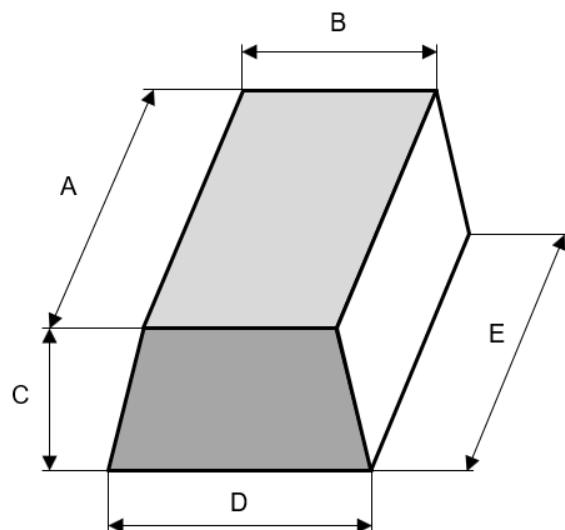


Bild 7.10: Testkörper zur Ermittlung von praxisrelevanten Schwindungsmaßen

Der Spritzgussprozess wird bei einer Werkzeugtemperatur T mit einem speziellen Druckprofil durch-

geführt, das in Bild 7.11 schematisch dargestellt ist. Die Einspritzung des geschmolzenen Kunststoffes erfolgt im Zeitraum von $0 \dots T_E$ bei sehr hohen Drücken p_E . Ist das Spritzgusswerkzeug mit Kunststoff gefüllt, wird der Kunststoff für einen Zeitraum $T_N \dots T_A$ einem definierten Nachdruck p_N ausgesetzt. Anschließend wird der Kunststoff abgekühlt und entformt.

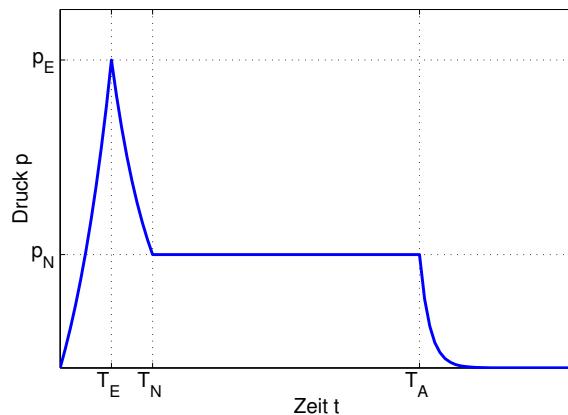


Bild 7.11: Schematisierter Druckverlauf während eines Spritzgussvorgangs

In einem Versuch wird die Schwindung des Bauteils unter unterschiedlichen Randbedingungen analysiert. Es ergibt sich der in Tabelle 7.12 dargestellte Datensatz.

Tabelle 7.12: Urliste eines Versuchs zur Bewertung der Schwindung beim Spritzgießen

Wand-dicke D / mm	Werkzeug-temperatur T / °C	Nach-druck p_N / bar	Schwin-dung S / %	Wand-dicke D / mm	Werkzeug-temperatur T / °C	Nach-druck p_N / bar	Schwin-dung S / %
1.5	31.2	508.0	1.20	8	60.1	740.5	2.39
5	28.7	509.4	1.83	1.5	91.4	757.8	1.42
8	30.8	490.1	2.42	5	91.1	755.7	2.07
1.5	58.0	502.1	1.42	8	89.5	741.8	2.62
5	60.0	502.4	2.08	1.5	29.2	997.3	0.70
8	59.9	489.9	2.64	5	29.4	988.1	1.35
1.5	90.0	492.6	1.67	8	27.0	978.0	1.90
5	89.4	510.8	2.30	1.5	59.5	1009.9	0.92
8	92.2	498.7	2.89	5	60.2	994.8	1.59
1.5	26.3	753.9	0.92	8	60.6	1003.3	2.13
5	30.9	750.9	1.60	1.5	92.9	1002.3	1.19
8	31.8	743.6	2.17	5	89.3	1000.2	1.81
1.5	61.5	744.4	1.20	8	91.2	990.0	2.39
5	61.2	754.4	1.83				

Um Abhängigkeiten der mehrdimensionalen Stichprobe eindeutig erkennen zu können, wird sie in einer Streudiagramm-Matrix dargestellt. Bild 7.12 stellt die Streudiagramm-Matrix zur Urliste aus Tabelle 7.12 dar. Die Matrix ist symmetrisch zur Hauptdiagonalen. An der grafischen Darstellung kann abgelesen werden, welche Kombinationen von Dicke D, Temperatur T und Nachdruck p_N zur Untersuchung der Schwindung verwendet wurden. Es wird außerdem deutlich, dass die Schwindung maßgeblich von der Wanddicke D abhängig ist.

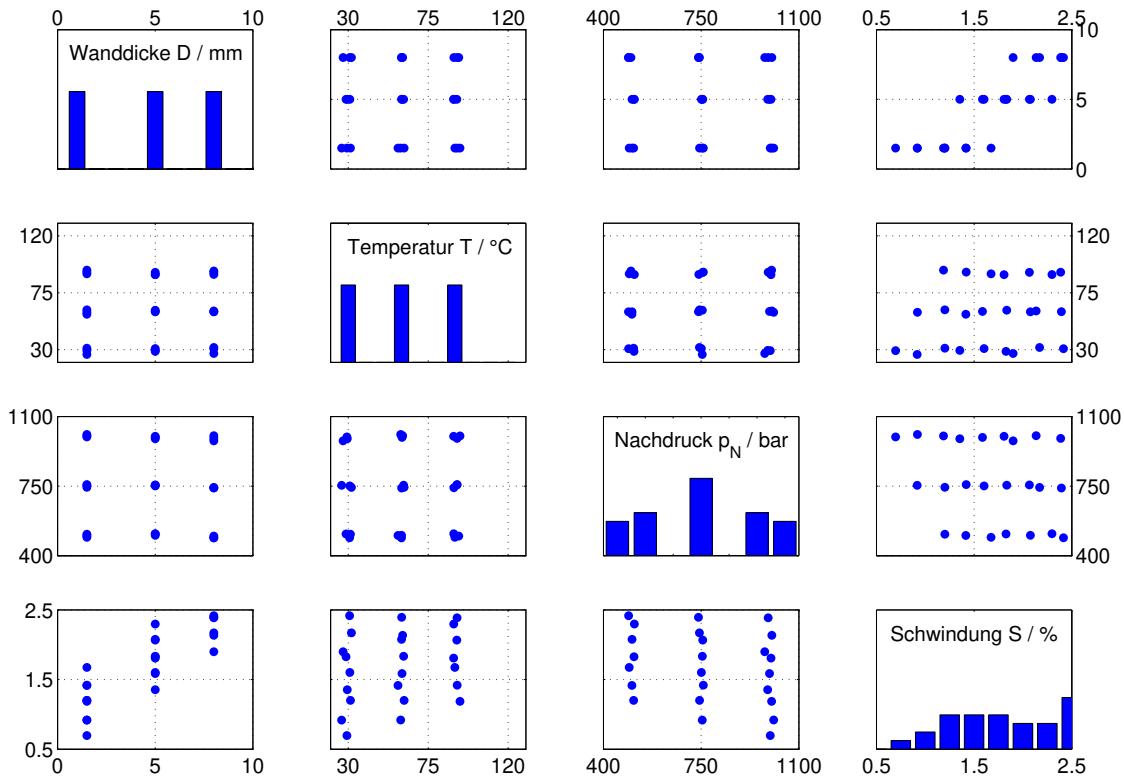


Bild 7.12: Berechnung der Kennwerte der multivariaten Stichprobe aus Tabelle 7.12 erfolgt mit MATLAB durch

```

1 % Messwerte einlesen
2 load spritzguss.mat;
3
4 % Grafische Darstellung
5 f = figure(1);
6 set(f, 'Position', [100 100 1200 800]);
7 namen = {'Dicke d / mm' 'Temperatur T/ C' 'Druck p_N / bar' 'Schwindung S / %'};
8 gplotmatrix([d' T' p' s'],[],[],[],[],[],'off','hist',namen)
9
10 % Berechnung der Kenngrö ß en
11 datamean = mean([d' T' p' s'])
12 datacov = cov([d' T' p' s'])
13 datacorr = corr([d' T' p' s'])

```

Es ergibt sich ein Vektor der Mittelwerte von

$$\bar{x} = \begin{pmatrix} 4.8333 & 60.1222 & 748.5519 & 1.8019 \end{pmatrix} \quad (7.28)$$

Die Kovarianzmatrix \mathbf{S} berechnet sich zu

$$S = \begin{pmatrix} 7.3269 & 0.3788 & -11.2641 & 1.3661 \\ 0.3788 & 652.3241 & 28.1773 & 5.1260 \\ -11.2641 & 28.1773 & 42564 & -44.4499 \\ 1.3661 & 5.1260 & -44.4499 & 0.3365 \end{pmatrix} \quad (7.29)$$

Die Daten der Kovarianzmatrix lassen sich wegen einer fehlenden Normierung nur bedingt interpretieren. Zum Beispiel ist die Varianz des Druckes aufgrund der hohen Zahlenwerte sehr hoch, auch die mit

dem Druck verbundenen Kovarianzen haben vergleichsweise große Werte. Deshalb wird zusätzlich die Korrelationsmatrix \mathbf{R} berechnet.

$$R = \begin{pmatrix} 1 & 0.0055 & -0.0202 & 0.8700 \\ 0.0055 & 1 & 0.0053 & 0.3460 \\ -0.0202 & 0.0053 & 1 & -0.3714 \\ 0.8700 & 0.3460 & -0.3714 & 1 \end{pmatrix} \quad (7.30)$$

Die größte Korrelation zur Schwindung hat die Wandstärke D. Wegen des positiven Vorzeichens steigt die Schwindung mit steigender Wandstärke. Im Gegensatz dazu liegt zwischen Nachdruck und Schwindung eine negative Korrelation vor, mit steigendem Nachdruck nimmt die Schwindung ab.

Uneinheitliche Schwindung an unterschiedlichen Orten im Bauteil führt dazu, dass sich ein Spritzgussteil verzieht. Um den Verzug im Bauteil zu verringern, muss bei der Konstruktion darauf geachtet werden, dass alle Teile eine vergleichbare Wandstärke aufweisen.

Ergänzend ist unten ein Python-Beispiel zur Auswertung der Stichprobe aufgeführt.

```

1 Bibliotheken importieren
2 from scipy.io import loadmat
3 import numpy as np
4 import matplotlib.pyplot as plt
5 import pandas as pd
6
7 Laden der Daten und Initialisieren der Variablen
8 values= loadmat('spritzguss')
9 d = values[ 'd' ]
10 T = values[ 'T' ]
11 p = values[ 'p' ]/1000
12 S = values[ 's' ]
13 X = np.append(np.append(np.append(d,T, axis=0),p, axis=0),S, axis=0)
14
15 Kumulative Randhäufigkeiten berechnen
16 dsort = np.append(np.append(0.0,np.sort(d)),10)
17 Tsort = np.append(np.append(20,np.sort(T)),130)
18 psort = np.append(np.append(0.4,np.sort(p)), 1.1)
19 Ssort = np.append(np.append(0.5,np.sort(S)), 3)
20 H=np.append([0,0],np.cumsum(1/len(T)*np.ones(np.shape(T)))) )
21
22 Kumulative Randhäufigkeiten darstellen
23 f1 = plt.figure(1, figsize=(12, 8))
24 axes1 = f1.subplots(2,2)
25
26 axes1[0,0].step(dsort,H, color='b')
27 axes1[0,0].grid(True, ls='--')
28 axes1[0,0].set_xlabel('Wanddicke d / mm')
29 axes1[0,0].set_ylabel('Kumulative Randhäufigkeit H_d(d)')
30
31 axes1[0,1].step(Tsort, H,color='b')\newline axes1[0,1].grid(True, ls='--')
32 axes1[0,1].set_xlabel('Temperatur T / °C')
33 axes1[0,1].set_ylabel('Kumulative Randhäufigkeit H\$\\_T\\$(T) ')
34
35 axes1[1,0].step(psort, H,color='b')
36 axes1[1,0].grid(True, ls='--')
37 axes1[1,0].set_xlabel('Nachdruck p / kbar')
```

```

38 axes1[1,0].set_ylabel('Kumulative Randhäufigkeit H_p(p)')
39
40 axes1[1,1].step(Ssort, H, color='b')
41 axes1[1,1].grid(True, ls='--')
42 axes1[1,1].set_xlabel('Schwindung S / %')
43 axes1[1,1].set_ylabel('Kumulative Randhäufigkeit H_S(S)')
44
45 Streudiagramm als Matrix
46 df = pd.DataFrame(np.transpose(X), columns=['d / mm', 'T / C', 'p / kbar',
47   'S / %'])
48 axes2 = pd.plotting.scatter_matrix(df, alpha=1, figsize=(12, 8))
49
50 Kennwerte berechnen
51 mX = np.mean(X, axis=1)
52 cX = np.cov(X)

```

7.4 Literatur

- [Krey91] Kreyszig, Erwin: Statistische Methoden und ihre Anwendungen
4., unveränderter Nachdruck der 7. Auflage
Vandenhoeck & Ruprecht, Göttingen, 1991
- [Fahr96] Fahrmeir, Ludwig; Hamerle, Alfred; Tutz, Gerhard: Multivariate statistische Verfahren
2., überarbeitete Auflage
Walter de Gruyter & Co., Berlin
- [Ross06] Ross, M. Sheldon: Statistik für Ingenieure und Naturwissenschaftler
3. Auflage
Spektrum Akademischer Verlag, München, 2006
- [Hart07] Hartung, Joachim; Elpelt, Bärbel: Multivariate Statistik
7., unveränderte Auflage
R. Oldenbourg Verlag, München / Wien
- [Papu01] Papula, Lothar: Mathematik für Ingenieure und Naturwissenschaftler Band 3
4., verbesserte Auflage
Vieweg Teubner, Braunschweig / Wiesbaden, 2008
- [Wiki11] Wikipedia: Spritzgießen <http://de.wikipedia.org>, 22.04.2011
- [BASF11] BASF: Verpackungsportal <http://www.packaging.bASF.com>, 22.04.2011
- [BOSC11] Robert Bosch GmbH: Heißfilmluftmassenmesser HFM7, Stuttgart, 2011

8 Multivariate Wahrscheinlichkeitstheorie

In Kapitel 4 wird ausgehend von der eindimensionalen Zufallsvariablen und deren Verteilung die univariante Wahrscheinlichkeitstheorie behandelt. Dabei werden Kenngrößen von Verteilungen berechnet und diese Erkenntnisse auf diskrete und stetige Verteilungen angewandt. Wie im vorangegangen Kapitel 7 gezeigt wird, treten in der Praxis jedoch oftmals zwei- oder mehrdimensionale Aufgabestellungen auf. Auch bei den Stichprobenfunktionen und den Schätzmethoden für Parameter, die in den folgenden Abschnitten behandelt werden, handelt es sich um mehrdimensionale Aufgabenstellungen. Aus diesem Grund werden in diesem Kapitel die Grundlagen für den Umgang mit multivariaten Problemstellungen gelegt. Dabei werden die wesentlichen Kenngrößen von mehrdimensionalen Verteilungen berechnet und auf diskrete und stetige Verteilungen angewandt. Zum Abschluss werden exemplarisch die Multinomial-Verteilung als diskrete und die Normalverteilung als stetige multivariate Verteilung vorgestellt.

8.1 Gemeinsame Verteilungs- und Dichtefunktionen

In der Praxis werden bei komplexeren Aufgabestellungen gleichzeitig mehrere Einflussgrößen beobachtet und analysiert. Dabei entstehen M-dimensionale Größen der Form

$$\underline{x}^T = \begin{pmatrix} x_1 & \dots & x_M \end{pmatrix} \quad (8.1)$$

Der Vektor \underline{x} wird als Zufallsvektor oder M-dimensionale Zufallsvariable bezeichnet, wenn die Komponenten x_1, \dots, x_M eindimensionale Zufallsvariablen sind. In Kapitel 7 wird gezeigt, wie derartige Stichproben durch Kenngrößen beschrieben werden können.

Im eindimensionalen Fall gilt die Beziehung

$$F(x) = P(\xi \leq x) \quad (8.2)$$

Bei multivariante Fragestellungen und Verfahren wird die Wahrscheinlichkeit mit der Verteilungsfunktion

$$F(\underline{x}) = F(x_1, \dots, x_M) = P(\xi_1 \leq x_1, \dots, \xi_M \leq x_M) = P(\xi \leq \underline{x}) \quad (8.3)$$

beschrieben. Dabei wird analog zu den univariaten Wahrscheinlichkeitsverteilungen aus Kapitel 4 zwischen diskreten und stetigen Verteilungen unterschieden.

8.1.1 Diskrete Verteilungen

Ein Zufallsvektor \underline{x} heißt diskret, wenn nur Werte aus einer abzählbaren Menge angenommen werden können. Die Wahrscheinlichkeitsfunktion einer diskreten Größe \underline{x} ist gegeben durch

$$f(\underline{x}) = \begin{cases} P(\xi = \underline{x}) \\ 0 \quad sonst \end{cases} \quad (8.4)$$

Durch Summation der einzelnen Glieder ergibt sich die Verteilungsfunktion $F(\underline{x})$ eines diskreten Zufallsvektors aus

$$F(\underline{x}) = F(x_1, \dots, x_M) = \sum_{\xi_1 < x_1} \dots \sum_{\xi_M < x_M} f(\xi_1, \dots, \xi_M) \quad (8.5)$$

Die diskrete multivariate Verteilung soll anhand eines Zufallsexperiments verdeutlicht werden, das aus Gründen der Übersichtlichkeit von zwei Zufallsgrößen abhängt.

Beispiel: Würfelexperiment

Als Beispiel wird das Würfeln mit zwei unterscheidbaren Würfeln aufgegriffen. Für das Experiment werden die beiden Zufallsgrößen x und y definiert. Die Variable x repräsentiert das Ergebnis des ersten

Würfels und die Variable y die Zahl, die der zweite Würfel anzeigt. Da die möglichen 36 Zahlenpaare alle gleichwahrscheinlich sind, hat jedes die Wahrscheinlichkeit von $1/36$. Bild 8.1 zeigt die Wahrscheinlichkeitsfunktion für das Experiment.

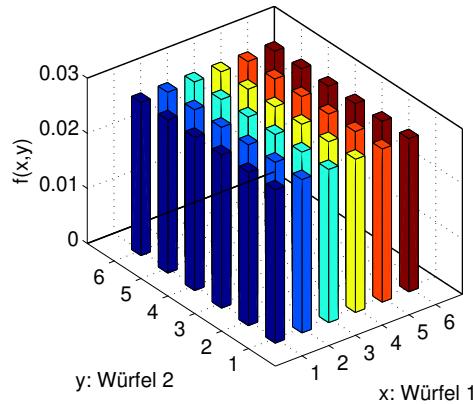


Bild 8.1: Grafische Darstellung der Wahrscheinlichkeitsfunktion $f(x,y)$

Durch Summation der einzelnen Wahrscheinlichkeiten ergibt sich die Verteilungsfunktion $F(x,y)$.

$$F(x,y) = \sum_{\xi=1}^x \sum_{\psi=1}^y f(\xi, \psi) \quad (8.6)$$

Sie ist in Bild 8.2 dargestellt.

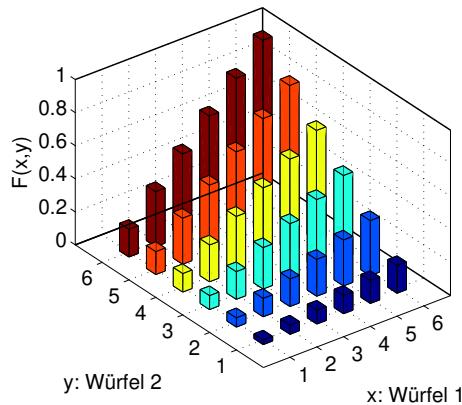


Bild 8.2: Grafische Darstellung der Verteilungsfunktion $F(x,y)$

Die Verteilungsfunktion wird benötigt, um zum Beispiel die Frage zu beantworten, mit welcher Wahrscheinlichkeit beim ersten Würfel eine Zahl kleiner gleich 2 und beim zweiten Würfel eine Zahl kleiner gleich 4 gewürfelt wird. Die Wahrscheinlichkeit hierfür ergibt sich aus Gleichung (8.6) zu

$$F(x \leq 4, y \leq 2) = \sum_{x_j=1}^4 \sum_{y_k=1}^2 f(x_j, y_k) = \frac{2 \cdot 4}{36} = 22\% \quad (8.7)$$

Das Ergebnis deckt sich mit dem Wert, der aus dem Diagramm in Bild 8.2 entnommen werden kann.

8.1.2 Stetige Verteilungen

Ein Zufallsvektor \underline{x} heißt stetig, wenn es eine Dichtefunktion

$$f(\underline{x}) = f(x_1, \dots, x_M) \geq 0 \quad (8.8)$$

mit der Verteilungsfunktion

$$F(\underline{x}) = F(x_1, \dots, x_M) = \int_{-\infty}^{x_M} \dots \int_{-\infty}^{x_1} f(\xi_1, \dots, \xi_M) d\xi_1 \dots d\xi_M = \int_{-\infty}^{\underline{x}} f(\underline{\xi}) d\underline{\xi} \quad (8.9)$$

gibt. Dabei muss $f(\underline{x})$ in der ganzen Ebene definiert, nicht negativ und beschränkt sein. Die zu einem Bereich $a_1 < x_1 \leq b_1, \dots, a_M < x_M \leq b_M$ gehörige Wahrscheinlichkeit ist dann durch

$$P(a_1 < x_1 \leq b_1, \dots, a_M < x_M \leq b_M) = \int_{a_M}^{b_M} \dots \int_{a_1}^{b_1} f(\xi_1, \dots, \xi_M) d\xi_1 \dots d\xi_M \quad (8.10)$$

gegeben.

Beispiel: Fertigung von Passstiften

Die stetige multivariate Verteilung soll anhand eines Beispiels verdeutlicht werden. Hierzu wird die Fertigung von Passstiften in einer automatisierten Fertigungseinrichtung betrachtet. Die Passstifte werden durch ihren Durchmesser D und ihre Länge L definiert. Die Fertigung ist auf einen Sollwert des Durchmessers von $D = 5 \text{ mm}$ und eine Länge von $L = 19 \text{ mm}$ eingestellt.

Durch den Verschleiß des Schneidewerkzeuges variieren die tatsächlichen Werte der Passstifte in einem Bereich von $\pm 1 \text{ mm}$ um den spezifizierten Sollwert. Durch den gleichmäßigen Verschleiß des Schneidewerkzeugs kann der Durchmesser D und die Länge L der gefertigten Passstifte mit einer multivariaten Gleichverteilung beschrieben werden. Jeder Wert in den Intervallen $4 \text{ mm} < D \leq 6 \text{ mm}$ und $18 \text{ mm} < L \leq 20 \text{ mm}$ kommt mit gleicher Wahrscheinlichkeit vor. Es soll ermittelt werden, wie viel Auschuss die Fertigungseinrichtung liefert, wenn eine Toleranzspanne von $\pm 5\%$ für den Durchmesser D und die Länge L zugelassen wird.

Entsprechend der Beschreibung der Fertigungseinrichtung wird die Dichteverteilung $f(D, L)$ beschrieben durch

$$f(D, L) = \begin{cases} \frac{1}{(b_1 - a_1) \cdot (b_2 - a_2)} & \text{für } a_1 < D \leq b_1, a_2 < L \leq b_2 \\ 0 & \text{sonst} \end{cases} \quad (8.11)$$

Bild 8.3 zeigt die durch Gleichung (8.11) definierte Dichteverteilung $f(D, L)$.

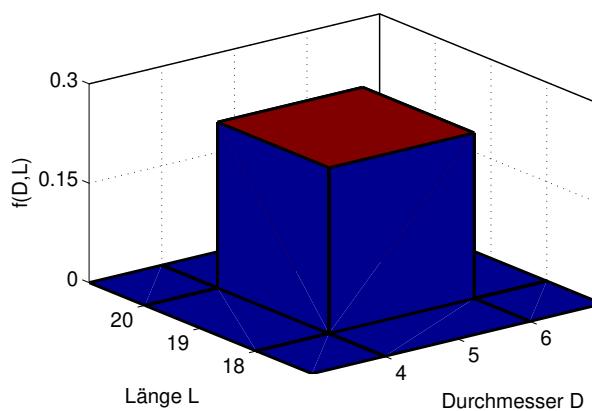


Bild 8.3: Grafische Darstellung der Dichtefunktion $f(D, L)$

Mit Gleichung (8.9) kann die Verteilungsfunktion $F(D, L)$ aufgestellt werden.

$$F(D, L) = \int_{18}^L \int_4^D \frac{1}{(b_1 - a_1) \cdot (b_2 - a_2)} d\delta d\lambda = \frac{(L - 4) \cdot (D - 18)}{(6 - 4) \cdot (20 - 18)} = \frac{(L - 4) \cdot (D - 18)}{4} \quad (8.12)$$

Daraus folgt die in Bild 8.4 dargestellte Verteilungsfunktion $F(D, L)$.

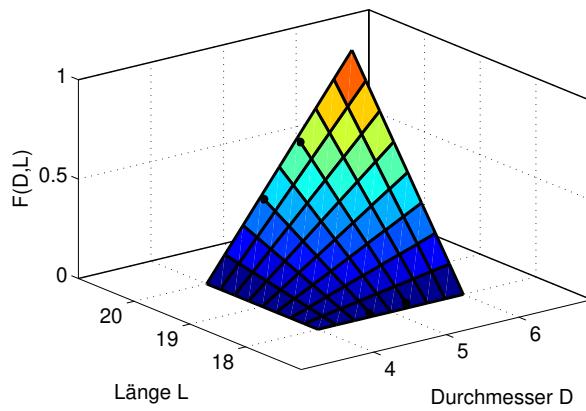


Bild 8.4: Grafische Darstellung der Verteilungsfunktion $F(D, L)$

Mithilfe der Verteilungsfunktion aus Gleichung (8.12) wird berechnet, mit welcher Wahrscheinlichkeit P bei der untersuchten Fertigung die produzierten Passstifte in einem spezifizierten Toleranzbereich von $\pm 5\%$ um den spezifizierten Durchmessers von $D = 5 \text{ mm}$ und die spezifizierte Länge von $L = 19 \text{ mm}$ liegen.

Statt der räumlichen Darstellung in Bild 8.4 wird zur Ermittlung der Wahrscheinlichkeit ein Kontur-Plot verwendet, der die Werte die Funktionswerte auf eine Ebene projiziert. Bild 8.5 zeigt die Verteilungsfunktion aus Bild 8.4 als Kontur-Plot.

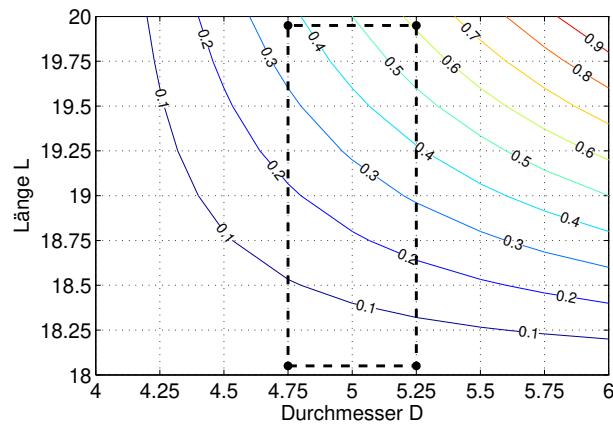


Bild 8.5: Kontur-Plot der Verteilungsfunktion $F(D, L)$

Die Wahrscheinlichkeit P , mit der die gefertigten Passstifte in dem definierten Toleranzbereich liegen, folgt damit zu

$$P = F(19.95, 5.25) - F(18.05, 5.25) - F(19.95, 4.75) + F(18.05, 4.75) = 23.75\% \quad (8.13)$$

Daraus ergibt sich ein prozentualer Ausschuss von

$$A = 1 - P = 1 - 0.2375 = 76.25\% \quad (8.14)$$

Der Ausschussanteil der Fertigung liegt bei 76.25 %, lediglich 23.75 % der gefertigten Passstifte entsprechen den spezifizierten Qualitätsanforderungen.

8.1.3 Randverteilungen

Bei der Auswertung von multivariaten Stichproben werden Randhäufigkeiten definiert und ausgerechnet. Dieser Randhäufigkeit entspricht bei multivariaten Verteilungen die Randverteilung. Jeder M -dimensionalen Verteilung lassen sich M eindimensionale Randverteilungen zuordnen. Um den Begriff der Randverteilung zu erläutern, erfolgt die Betrachtung am Beispiel des Würfeln mit zwei unterschiedbaren Würfeln aus Abschnitt 8.1.1.

Für das Würfelbeispiel wird die Fragestellung untersucht, mit welcher Wahrscheinlichkeit der erste Würfel die Zahl 4 aufweist. Der zweite Würfel bleibt bei dieser Fragestellung unbeachtet.

$$f_x(4) = P(x = 4, y = \text{beliebig}) \quad (8.15)$$

Die Berechnung erfolgt im diskreten Fall durch die Addition aller Wahrscheinlichkeiten $f(x,y)$, bei der das positive Ereignis $x = 4$ eintritt.

$$f_x(4) = \sum_{y=1}^6 f(4,y) = \frac{6}{36} = \frac{1}{6} = 16.7\% \quad (8.16)$$

Genauso könnte die Frage nach der Wahrscheinlichkeit gestellt werden, mit der der zweite Würfel den Wert 2 aufweist, während der erste Würfel unberücksichtigt bleibt.

$$f_y(2) = P(x = \text{beliebig}, y = 2) \quad (8.17)$$

Die Funktionen aus Gleichung (8.15) und Gleichung (8.17) stellen eine eindimensionale Wahrscheinlichkeitsverteilung dar. Sie werden als Randverteilung der Variablen x beziehungsweise y bezüglich der gegebenen zweidimensionalen Verteilung bezeichnet. Durch Summation ergeben sich die zugehörigen Verteilungsfunktionen der Randverteilungen

$$F_x(x) = P(\xi \leq x, y = \text{beliebig}) = \sum_{\xi=1}^x f_x(\xi) \quad (8.18)$$

und

$$F_y(y) = P(x = \text{beliebig}, \psi \leq y) = \sum_{\psi=1}^y f_y(\psi) \quad (8.19)$$

Für das Würfelbeispiel aus Abschnitt 8.1.1 ergeben sich die in Tabelle 8.1 aufgelisteten Werte für die Wahrscheinlichkeiten $f(x,y)$ der Wertepaare und die entsprechenden Werte der Randverteilungen.

Tabelle 8.1: Wahrscheinlichkeitsfunktionen $f(x)$ und $f(y)$ der Randverteilungen

Es soll die Frage untersucht werden, mit welcher Wahrscheinlichkeit bei dem ersten Würfel eine Zahl kleiner oder gleich 5 gewürfelt wird, wenn Würfel 2 unberücksichtigt bleibt. Diese Wahrscheinlichkeit berechnet sich zu

$$F_x(5) = P(x \leq 5, y = \text{beliebig}) = \frac{6 \cdot 5}{36} = 83.3\% \quad (8.20)$$

Für den stetigen Fall folgt äquivalent für die Wahrscheinlichkeitsdichte

$$f_x(x) = \int_{-\infty}^{\infty} f(x, y) dy \quad (8.21)$$

und

$$f_y(y) = \int_{-\infty}^{\infty} f(x, y) dx \quad (8.22)$$

Als Bestimmungsgleichung für die Verteilungsfunktion ergeben sich

$$F_x(x) = \int_{-\infty}^x f_x(\xi) d\xi \quad (8.23)$$

und

$$F_y(y) = \int_{-\infty}^y f_y(\psi) d\psi \quad (8.24)$$

Die Randverteilungen einer stetigen Verteilung sind ebenfalls stetig.

Für den Fall multivariater Datensätze mit mehr als zwei Dimensionen ergeben sich M eindimensionale Randverteilungen, die jeweils nur von einer Zufallsvariable abhängen. Mit ihnen können oftmals mehrdimensionale Fragestellungen in mehrere univariate Teilprobleme zerlegt werden.

8.2 Kenngrößen multivariater Wahrscheinlichkeitsverteilungen

In der deskriptiven Statistik in Kapitel 7 werden multivariate Datensätze beschrieben. Dazu werden für die vorliegenden Stichproben Häufigkeitsverteilungen bestimmt und empirische Kenngrößen berechnet. Im Folgenden werden Verteilungen durch ihren Mittelwert und ihre Kovarianzmatrix beschrieben.

8.2.1 Arithmetischer Mittelwert als Lagekennwert einer multivariaten Verteilung

Die Lage einer mehrdimensionalen Zufallsgröße \underline{x} kann über den Vektor der arithmetischen Mittelwerte beschrieben werden. Er ist definiert durch den Erwartungswertvektor von \underline{x}

$$\underline{\mu}^T = E(\underline{x}^T) = \begin{pmatrix} E(x_1) & \cdots & E(x_M) \end{pmatrix} = \begin{pmatrix} \mu_1 & \cdots & \mu_M \end{pmatrix} \quad (8.25)$$

Die Berechnung der einzelnen Mittelwerte erfolgt mit dem in Abschnitt 4.1.3 vorgestellten univariaten Ausdruck für stetige Verteilungen

$$\mu = \int_{-\infty}^{\infty} x \cdot f(x) dx \quad (8.26)$$

Die Mittelwerte μ_m werden über die Randverteilungen $f_{x_m}(x_m)$ der Zufallsvariablen x_m bestimmt. Sie sind erwartungsgemäß von den übrigen Variablen $x_k \neq x_m$ unabhängig. Damit ergibt sich der Mittelwert einer stetigen multivariaten Verteilung zu

$$\underline{\mu}^T = E(\underline{x}^T) = \left(\int_{-\infty}^{\infty} x_1 \cdot f_{x_1}(x_1) dx_1 \quad \cdots \quad \int_{-\infty}^{\infty} x_M \cdot f_{x_M}(x_M) dx_M \right) \quad (8.27)$$

8.2.2 Kovarianzmatrix als Streuungskennwert einer multivariaten Verteilung

Die Streuung einer multivarianten Zufallsgröße wird über die Kovarianzmatrix bestimmt. Die Kovarianzmatrix ist ähnlich wie die Varianz bei univariaten Zufallsgrößen über den Erwartungswert definiert.

$$\begin{aligned} \Sigma &= \begin{pmatrix} \sigma_1^2 & \cdots & \sigma_{1M} \\ \vdots & & \vdots \\ \sigma_{M1} & \cdots & \sigma_M^2 \end{pmatrix} \\ &= \begin{pmatrix} E((x_1 - \mu_1)^2) & \cdots & \sigma_{1M} = E((x_1 - \mu_1) \cdot (x_M - \mu_M)) \\ \vdots & & \vdots \\ \sigma_{M1} = E((x_M - \mu_M) \cdot (x_1 - \mu_1)) & \cdots & E((x_M - \mu_M)^2) \end{pmatrix} \\ &= E \left(\left(X - \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix} \cdot \underline{\mu}^T \right)^T \cdot \left(X - \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix} \cdot \underline{\mu}^T \right) \right) \end{aligned} \quad (8.28)$$

Wegen der Kommutativität des Erwartungswertes ist die Kovarianzmatrix symmetrisch zur Hauptdiagonalen. Mit dem arithmetischen Mittelwert und der Kovarianzmatrix ist eine Verteilung hinsichtlich ihrer Lage und Streuung definiert.

8.3 Unabhängige Zufallsvariablen

In vielen Fällen vereinfacht sich die Lösung multivariater Aufgabenstellungen, wenn von unabhängigen Zufallsvariablen ausgegangen werden kann. In der Praxis ist dies oftmals zumindest in guter Näherung gegeben. Aus diesem Grund werden im Folgenden Eigenschaften unabhängiger Variablen näher untersucht.

8.3.1 Verteilungen unabhängiger Zufallsvariablen

In Abschnitt 2.4.2 wird die Unabhängigkeit von Ereignissen diskutiert. Es wird gezeigt, dass bei unabhängigen Ereignissen die folgende Beziehung gilt.

$$P(A \cap B) = P(A) \cdot P(B) = P(B) \cdot P(A) \quad (8.29)$$

Auf gleiche Weise kann die Unabhängigkeit bei Zufallsvariablen ausgedrückt werden. Zwei Zufallsvariablen x und y einer zweidimensionalen Verteilung mit der gemeinsamen Verteilungsfunktion $F(x,y)$ sind voneinander unabhängig, wenn für alle möglichen Kombinationen von x und y die Beziehung gilt

$$F(x,y) = F_x(x) \cdot F_y(y) \quad (8.30)$$

Ist diese Bedingung nicht für alle Fälle erfüllt, sind die Zufallsvariablen abhängig. In Analogie zu Gleichung (8.30) kann die Bedingung für die Unabhängigkeit von zwei Variablen nach den Rechenregeln zur Integralrechnung auch durch die Beziehung

$$f(x,y) = f_x(x) \cdot f_y(y) \quad (8.31)$$

ausgedrückt werden. Dies kann entsprechend auf den multivariaten Fall verallgemeinert werden. Die Bedingung für die Unabhängigkeit kann hierbei durch die Beziehung

$$F(x_1, \dots, x_M) = \prod_{m=1}^M F_m(x_m) \quad (8.32)$$

beziehungsweise

$$f(x_1, \dots, x_M) = \prod_{m=1}^M f(x_m) \quad (8.33)$$

bewertet werden.

Beispiel: Ziehen von verschiedenfarbigen Kugeln

An einem Beispiel soll geprüft werden, ob es sich bei den verwendeten Zufallsvariablen um unabhängige Zufallsvariablen handelt. Hierzu wird angenommen, dass in einer Urne zehn Kugeln liegen. Vier Kugeln haben die Farbe Blau, die restlichen Kugeln sind rot. Es werden nacheinander zwei Kugeln gezogen, die zuerst gezogene Kugel wird nicht zurückgelegt.

Im Folgenden werden die Zufallsvariablen x und y betrachtet. Die Zufallsvariable x repräsentiert die Zahl der blauen Kugeln beim ersten Zug, die Zufallsvariable y die Anzahl blauer Kugeln beim zweiten Zug. Die Werte der Wahrscheinlichkeitsfunktionen können anhand der Häufigkeiten berechnet werden. Diese sind in Tabelle 8.2 zusammengefasst.

Tabelle 8.2: Wahrscheinlichkeiten des Zufallsexperimentes

Tabelle 8.2: Wahrscheinlichkeitsfunktionen $f(x)$ und $f(y)$ der Randverteilungen

		Zug 1	
		blaue Kugel (x = 1)	rote Kugel (x = 0)
Zug 2	blaue Kugel (y = 1)	$f(1,1) = \frac{4}{10} \cdot \frac{3}{9} = \frac{2}{15}$	$f(0,1) = \frac{6}{10} \cdot \frac{4}{9} = \frac{4}{15}$
	rote Kugel (y = 0)	$f(1,0) = \frac{4}{10} \cdot \frac{6}{9} = \frac{4}{15}$	$f(0,0) = \frac{6}{10} \cdot \frac{5}{9} = \frac{1}{3}$

Zur Überprüfung der Unabhängigkeit müssen die Werte der Randverteilungen berechnet werden. Die Wahrscheinlichkeit, dass beim ersten Zug keine blaue Kugel gezogen wird, berechnet sich zu

$$f_x(x = 0) = \frac{4}{15} + \frac{1}{3} = 0.6 \quad (8.34)$$

und die Wahrscheinlichkeit, dass beim zweiten Zug keine blaue Kugel gezogen wird, ergibt sich zu

$$f_y(y = 0) = \frac{4}{15} + \frac{1}{3} = 0.6 \quad (8.35)$$

Damit ergibt sich bei unabhängigen Zufallsvariablen die Wahrscheinlichkeit, zweimal hintereinander eine rote Kugel zu ziehen, durch die Multiplikation der Wahrscheinlichkeiten der Randverteilungen aus Gleichung (8.34) und Gleichung (8.35) zu

$$f_x(x = 0) \cdot f_y(y = 0) = 0.6 \cdot 0.6 = 0.36 \quad (8.36)$$

Der Vergleich mit Tabelle 4.4 zeigt, dass die Wahrscheinlichkeit für dieses Ereignis 0.33 beträgt. Damit gilt zumindest für ein Wertepaar die Beziehung

$$f(x,y) \neq f_x(x) \cdot f_y(y) \quad (8.37)$$

Die Variablen sind demnach abhängig. Die Ursache liegt darin, dass die Wahrscheinlichkeit, beim zweiten Zug eine blaue Kugel zu ziehen, durch den ersten Zug verändert wird. Würde die Kugel des ersten Zuges zurück in die Urne gelegt, gäbe es keine Veränderung der Wahrscheinlichkeit gegenüber der Ausgangsposition. Die Zufallsgrößen wären in diesem Fall unabhängig.

8.3.2 Kovarianz unabhängiger Zufallsvariablen

Handelt es sich bei den Zufallsvariablen x und y um unabhängige Zufallsvariablen, gilt für deren Kovarianz σ_{xy}

$$\sigma_{xy} = E((x - \mu_x) \cdot (y - \mu_y)) = 0 \quad (8.38)$$

Um diese Beziehung zu beweisen, muss Gleichung (8.38) weiter umgeformt werden. Ausmultiplizieren des Terms führt zu

$$\begin{aligned}\sigma_{xy} &= E(x \cdot y) - E(x) \cdot \mu_y - E(y) \cdot \mu_x + \mu_x \cdot \mu_y = E(x \cdot y) - \mu_x \cdot \mu_y - \mu_x \cdot \mu_y + \mu_x \cdot \mu_y \\ &= E(x \cdot y) - \mu_x \cdot \mu_y = E(x \cdot y) - E(x) \cdot E(y) = 0\end{aligned}\quad (8.39)$$

Damit die Kovarianz der Zufallsvariablen x und y zu null wird, muss demnach gelten

$$E(x \cdot y) = E(x) \cdot E(y) \quad (8.40)$$

Diese Bedingung wird für diskrete und stetige Zufallsvariable berechnet. Aus Kapitel 4 ist bekannt, dass sich der Erwartungswert für eine diskrete Zufallsvariable z berechnet aus

$$E(z) = \sum_{j=-\infty}^{\infty} z_j \cdot f(z_j) = \sum_{j=-\infty}^{\infty} z_j \cdot P(z = z_j) \quad (8.41)$$

Wird die Zufallsvariable z aus Gleichung (8.41) durch das Produkt der Zufallsvariablen x und y ersetzt, ergibt sich

$$E(x \cdot y) = \sum_{j=-\infty}^{\infty} \sum_{k=-\infty}^{\infty} x_j \cdot y_k \cdot f(x_j, y_k) = \sum_{j=-\infty}^{\infty} \sum_{k=-\infty}^{\infty} x_j \cdot y_k \cdot P(x = x_j \cap y = y_k) \quad (8.42)$$

Da es sich bei den Zufallsvariablen x und y um unabhängige Zufallsgrößen handelt, kann mit Gleichung (2.65) die vorige Gleichung (8.42) umgestellt werden zu

$$\begin{aligned}E(x \cdot y) &= \sum_{j=-\infty}^{\infty} \sum_{k=-\infty}^{\infty} x_j \cdot y_k \cdot P(x = x_j \cap y = y_k) = \sum_{j=-\infty}^{\infty} x_j \cdot P(x = x_j) \cdot \sum_{k=-\infty}^{\infty} y_k \cdot P(y = y_k) \\ &= \sum_{j=-\infty}^{\infty} x_j \cdot f_{x_j}(x_j) \cdot \sum_{k=-\infty}^{\infty} y_k \cdot f_{y_k}(y_k) = E(x) \cdot E(y)\end{aligned}\quad (8.43)$$

Die Umrechnung hat gezeigt, dass die Beziehung aus Gleichung (8.40) für unabhängige diskrete Zufallsvariablen erfüllt ist und damit die Kovarianz zu Null wird. Analog kann dies mit den Regeln zum Erwartungswert aus Kapitel 4 und unter Berücksichtigung von Gleichung (8.31) auch für stetige Zufallsvariablen gezeigt werden.

$$\begin{aligned}E(x \cdot y) &= \int_{-\infty}^{\infty} z \cdot f(z) dz = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} x \cdot y \cdot f(x, y) dx dy = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} x \cdot y \cdot f_x(x) \cdot f_y(y) dx dy \\ &= \int_{-\infty}^{\infty} x \cdot f_x(x) dx \cdot \int_{-\infty}^{\infty} y \cdot f_y(y) dy = E(x) \cdot E(y)\end{aligned}\quad (8.44)$$

Damit gilt allgemein, dass die Kovarianz unabhängiger Zufallsgrößen null ist.

8.4 Funktionen von Zufallsvariablen

Zur Beschreibung von Zufallsexperimenten kann es notwendig sein, eine Funktion von Zufallsvariablen abzubilden. Dabei wird jedem möglichen Wert des Vektors von Zufallsvariablen \underline{x} über die Funktion $g(\underline{x})$ ein reeller Wert zugeordnet. Hierbei sind vier grundlegende Rechenoperationen von Bedeutung, mit denen komplexe Funktionen analysiert werden können:

1. Summe von Zufallszahlen
2. Differenz von Zufallszahlen
3. Produkt von Zufallszahlen
4. Quotient von Zufallszahlen

Die Herleitung beschränkt sich der Übersicht wegen auf zwei stetige Zufallsvariablen. Außerdem wird an dieser Stelle der Zusammenhang für die Summe von Zufallszahlen hergeleitet. Die Herleitungen zu den übrigen Rechenoperationen sind ähnlich. Die dabei erlangten Erkenntnisse lassen sich auf Funktionen diskreter Zufallsvariablen und auch auf mehrere Zufallsvariablen erweitern.

8.4.1 Summe von Zufallsvariablen

Im Folgenden soll die Summenfunktion der unabhängigen Zufallsvariablen x und y untersucht werden. Die Verteilungsfunktion der neuen Zufallsvariablen der Form

$$z = x + y \quad (8.45)$$

berechnet sich zu

$$P(\zeta \leq z) = F_z(z) = \iint_{x+y \leq z} f(x,y) dx dy \quad (8.46)$$

Wahrscheinlichkeitsdichte der Summe von unabhängigen Zufallszahlen

Wie bereits bei der linearen Transformation muss zur Umformung des Integrals eine entsprechende Koordinatentransformation durchgeführt werden. Hierbei wird

$$y = z - x \quad (8.47)$$

gesetzt. Diese Koordinatentransformation erfordert eine Umrechnung des Integranden. Sie lässt sich nach dem Transformationssatz durch die Berechnung der Jacobi-Determinanten durchführen. Es ergibt sich

$$dxdy = \left| \begin{pmatrix} \frac{\partial x}{\partial x} & \frac{\partial x}{\partial y} \\ \frac{\partial y}{\partial x} & \frac{\partial y}{\partial y} \end{pmatrix} \right| dxdz = \left| \begin{pmatrix} \frac{\partial x}{\partial x} & \frac{\partial x}{\partial z} \\ \frac{\partial (z-x)}{\partial x} & \frac{\partial (z-x)}{\partial z} \end{pmatrix} \right| dxdz = \left| \begin{pmatrix} 1 & 0 \\ -1 & 1 \end{pmatrix} \right| dxdz = |1| dxdz \quad (8.48)$$

Damit folgt Gleichung (8.46) zu

$$F_z(z) = \int_{-\infty}^z \int_{-\infty}^{\infty} f(x, \zeta - x) dx d\zeta = \int_{-\infty}^z f_z(\zeta) d\zeta \quad (8.49)$$

Die Wahrscheinlichkeitsdichte ergibt sich aus dem inneren Integral zu

$$f_z(z) = \int_{-\infty}^{\infty} f(x, z - x) dx \quad (8.50)$$

Im Fall unabhängiger Zufallsvariablen kann sie als Produkt zweier Wahrscheinlichkeitsdichten dargestellt werden. Das Integral entspricht in diesem Fall der Faltung der Wahrscheinlichkeitsdichten.

$$f_z(z) = \int_{-\infty}^{\infty} f(x, z-x) dx = \int_{-\infty}^{\infty} f_x(x) \cdot f_y(z-x) dx = f_x(x) * f_y(z-x) \quad (8.51)$$

Diese Erkenntnis kann auf M unabhängige Zufallsvariablen erweitert werden. Die Verteilungsfunktion der Summe von M unabhängigen Zufallsvariablen

$$z = \sum_{m=1}^M x_m \quad (8.52)$$

ergibt sich aus der Faltung der M Verteilungsfunktionen $f_{xm}(x_m)$.

Kenngrößen für die Summe von Zufallsvariablen

Zur Berechnung des Mittelwertes μ_z wird die Definition über den Erwartungswert herangezogen. Mit der Linearität des Erwartungswert-Operators folgt

$$\mu_z = E(z) = E(x + y) = E(x) + E(y) = \mu_x + \mu_y \quad (8.53)$$

Der Gesamtmittelwert setzt sich nach Gleichung (8.53) aus der Summe der Einzelmittelwerte zusammen. Nach den Rechenregeln zum Erwartungswert errechnet sich die Varianz einer Zufallsvariable z zu

$$\sigma_z^2 = E(z^2) - E(z)^2 \quad (8.54)$$

Mit z als Summe der Zufallsvariablen x und y ergibt sich

$$E(z^2) = E(x^2 + 2 \cdot x \cdot y + y^2) = E(x^2) + 2 \cdot E(x \cdot y) + E(y^2) \quad (8.55)$$

und

$$(E(z))^2 = (E(x + y))^2 = (E(x) + E(y))^2 = (E(x))^2 + 2 \cdot E(x) \cdot E(y) + (E(y))^2 \quad (8.56)$$

Werden die beiden Ausdrücke aus Gleichung (8.55) und Gleichung (8.56) in die Gleichung (8.54) der Varianz σ_z^2 eingesetzt, ergibt sich

$$\sigma_z^2 = (E(x^2) - (E(x))^2) + (E(y^2) - (E(y))^2) + 2 \cdot (E(x \cdot y) - E(x) \cdot E(y)) \quad (8.57)$$

In den ersten Klammern der Gleichung stehen die Varianzen der Zufallsvariablen x und y. Der letzte Ausdruck kann zusammengefasst dargestellt werden als

$$E(x \cdot y) - E(x) \cdot E(y) = \sigma_{xy} \quad (8.58)$$

Die Varianz der Summe zweier Zufallsvariablen ergibt sich damit allgemein zu

$$\sigma_z^2 = \sigma_x^2 + \sigma_y^2 + 2 \cdot \sigma_{xy} \quad (8.59)$$

Die Größe σ_{xy} ist die Kovarianz der Zufallsvariablen x und y.

Sind die Variablen x und y unabhängig voneinander, ist die Kovarianz null. Damit ergibt sich die Varianz der Summe unabhängiger Zufallsvariablen x und y aus der Summe

$$\sigma_z^2 = \sigma_x^2 + \sigma_y^2 \quad (8.60)$$

Diese Regeln kann auf M unabhängige Zufallsvariablen x_m mit den Mittelwerten μ_m und σ_m erweitert werden. Die Summe der Zufallsvariablen

$$z = \sum_{m=1}^M x_m \quad (8.61)$$

besitzt demzufolge einen Mittelwert bei

$$\mu_z = \sum_{m=1}^M \mu_m \quad (8.62)$$

und weist eine Varianz von

$$\sigma_z^2 = \sum_{m=1}^M \sigma_m^2 \quad (8.63)$$

auf.

8.4.2 Zusammenfassung Funktionen von Zufallszahlen

Analog zu der Herleitung für die Summe zweier Zufallszahlen kann bei den übrigen Rechenoperationen vorgegangen werden. Hierzu muss zunächst eine entsprechende Zufallsvariable definiert werden. Durch eine Variablentransformation ergibt sich die Verteilungsfunktion der neuen Zufallsvariable z . Mithilfe des Erwartungswertoperators lassen sich Ausdrücke für den Mittelwert μ_z und der Varianz σ_z^2 berechnen. Für unabhängige Zufallsvariable sind in Tabelle 8.3 die Dichtefunktion $f(z)$, der Mittelwert μ_z und die Varianz σ_z^2 der verschiedenen Rechenoperationen zusammengestellt.

Tabelle 8.3: Zusammenfassung der Funktionen von unabhängigen Zufallsvariablen

Rechenoperation	Dichtefunktion $f(z)$	Mittelwert μ_z	Varianz σ_z^2
Addition	$f_{x+y}(z) = \int_{-\infty}^{\infty} f_x(x) \cdot f_y(z-x) dx$	$\mu_z = \mu_x + \mu_y$	$\sigma_z^2 = \sigma_x^2 + \sigma_y^2$
Subtraktion	$f_{x-y}(z) = \int_{-\infty}^{\infty} f_x(x) \cdot f_y(x-z) dx$	$\mu_z = \mu_x - \mu_y$	$\sigma_z^2 = \sigma_x^2 + \sigma_y^2$
Multiplikation	$f_{x \cdot y}(z) = \int_{-\infty}^{\infty} \left \frac{1}{x} \right \cdot f_x(x) \cdot f_y\left(\frac{z}{x}\right) dx$	$\mu_z = \mu_x \cdot \mu_y$	$\sigma_z^2 = \mu_x^2 \cdot \sigma_y^2 + \mu_y^2 \cdot \sigma_x^2$
Division	$f_{x/y}(z) = \int_{-\infty}^{\infty} \left \frac{x}{z^2} \right \cdot f_x(x) \cdot f_y\left(\frac{x}{z}\right) dx$	$\mu_z = \mu_x \cdot \mu_{1/y}$	$\sigma_z^2 = \mu_x^2 \cdot \sigma_{1/y}^2 + \mu_{1/y}^2 \cdot \sigma_x^2$

Die in diesem Kapitel vorgestellten Funktionen zweier unabhängiger Zufallszahlen lassen sich durch wiederholte Anwendung auf Funktionen mehrerer Zufallsvariablen erweitern.

Im Fall abhängiger Zufallsvariablen kann die Dichtefunktion nicht auf diese Art berechnet werden. Die Rechenregeln für Mittelwert und Standardabweichung bleiben jedoch für die Summe und die Differenz von Zufallsvariablen erhalten, und es ergeben sich die in Tabelle 8.4 dargestellten Zusammenhänge.

Tabelle 8.4: Zusammenfassung der Funktionen von abhängigen Zufallsvariablen

Rechenoperation	Mittelwert μ_z	Varianz σ_z^2
Addition	$\mu_z = \mu_x + \mu_y$	$\sigma_z^2 = \sigma_x^2 + \sigma_y^2 + 2 \cdot \sigma_{xy}$
Subtraktion	$\mu_z = \mu_x - \mu_y$	$\sigma_z^2 = \sigma_x^2 + \sigma_y^2 - 2 \cdot \sigma_{xy}$

Durch Anwenden dieser Funktionen lässt sich eine multivariate Fragestellung auf eine univariate Fragestellung abbilden. Sie sind insbesondere bei der Toleranzrechnung von Bedeutung.

8.5 Zentraler Grenzwertsatz

Nach dem zentralen Grenzwertsatz besitzt eine Zufallsvariable, die sich aus der Summe von unabhängigen Zufallsvariablen ergibt, eine Normalverteilung. Zum Beweis wird von den Zufallsvariablen x_1, \dots, x_M ausgegangen, die alle denselben Mittelwert μ und dieselbe Varianz σ^2 aufweisen. Die Zufallsvariable y , die sich aus der Summe unabhängiger Zufallsvariablen ergibt

$$y = x_1 + x_2 + \dots + x_M = \sum_{m=1}^M x_m \quad (8.64)$$

hat nach Gleichung (8.62) den Mittelwert $\mu_y = M \cdot \mu$ und nach Gleichung (8.63) die Varianz $\sigma_y^2 = M \cdot \sigma^2$. Sind die Variablen x_1, \dots, x_M normalverteilt, ist auch die Variable y normalverteilt und die Zufallsvariable

$$z = \frac{y - \mu_y}{\sigma_y} = \frac{y - M \cdot \mu}{\sigma \cdot \sqrt{M}} \quad (8.65)$$

weist eine Standardnormalverteilung auf.

Sind die Variablen nicht normalverteilt, so ist die Zufallsvariable z aus Gleichung (8.65) für eine große Anzahl M von Summanden asymptotisch standardnormalverteilt. Diese wichtige Beziehung wird als zentraler Grenzwertsatz der Wahrscheinlichkeitsrechnung bezeichnet. Er ist der wesentliche Grund für die große Bedeutung der Normalverteilung in der Wahrscheinlichkeitstheorie. Statt eines Beweises wird der zentrale Grenzwertsatz grafisch motiviert. Als Basis werden Variablen x_m mit einer Gleichverteilung in einem Bereich von $-0.2 \dots 0.2$ verwendet. Sie besitzen eine Gleichverteilung mit einem Mittelwert $\mu = 0$ und einer Varianz $\sigma^2 = 0.0133$. Die Wahrscheinlichkeitsverteilung der Summe von unabhängigen Zufallsvariablen ergibt sich nach Gleichung (8.51) aus der Faltung der einzelnen Wahrscheinlichkeitsdichten.

$$f(y) = f(x_1) * f(x_2) * \dots * f(x_M) \quad (8.66)$$

Die Wahrscheinlichkeitsdichte $f(y)$ hängt davon ab, wie viele Faltungen stattgefunden haben. Bild 8.6 stellt die Wahrscheinlichkeitsdichten für unterschiedliche Anzahlen M von Zufallsvariablen dar.

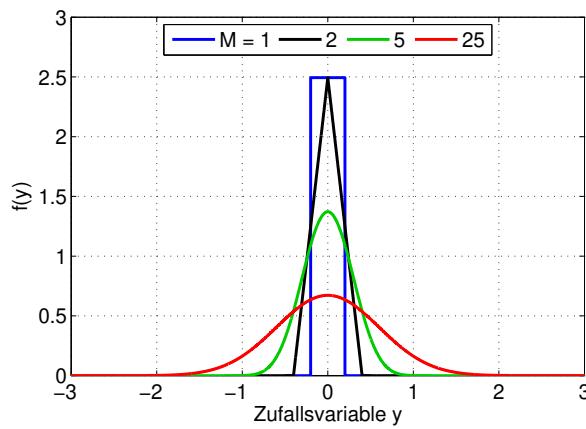


Bild 8.6: Wahrscheinlichkeitsdichte der Zufallsvariablen y bei unterschiedlicher Anzahl summierter Zufallsvariablen

Die Wahrscheinlichkeitsdichte $f(y)$ wird mit steigendem M breiter und nähert sich in ihrer Form der Standardnormalverteilung zunehmend an. In Bild 8.7 wird für $M = 25$ die Verteilung der Zufallsvariable

$$z = \frac{y - M \cdot \mu}{\sigma \cdot \sqrt{M}} \quad (8.67)$$

mit einer Standardnormalverteilung verglichen.

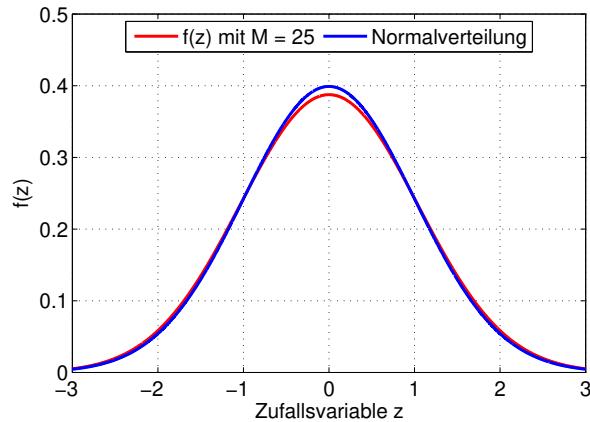


Bild 8.7: Wahrscheinlichkeitsdichte der Zufallsvariablen z im Vergleich zur Standardnormalverteilung

Es zeigt sich eine gute grafische Übereinstimmung der beiden Wahrscheinlichkeitsdichten, sodass das Beispiel den zentralen Grenzwertsatz grafisch bestätigt. Die Güte der Übereinstimmung ist von der Anzahl der Stichproben M abhängig. Wie groß die Anzahl erforderlicher Stichproben M ist, um eine ausreichend gute Näherung der Standardnormalverteilung zu erhalten, kann nicht pauschal beantwortet werden. Die Anzahl notwendiger Stichproben ist von der Wahrscheinlichkeitsdichte $f_{xm}(x_m)$ der Zufallsvariablen x_m und der Aufgabenstellung abhängig. In dem Beispiel zeigt sich für $M = 25$ eine gute Übereinstimmung, in der Literatur wird der Wert $M = 30$ für eine gute Approximation angegeben. Oft reichen aber auch bereits erheblich weniger Stichproben aus, um mit der Normalverteilung näherungsweise rechnen zu können.

8.6 Spezielle multivariate Verteilungen

Abschließend werden mit der Multinomial-Verteilung für diskrete Variablen und der multivariaten Normalverteilung für stetige Variablen die beiden wichtigsten multivariaten Verteilungen eingeführt.

8.6.1 Multinomial-Verteilung

Die Multinomial-Verteilung ergibt sich als Verallgemeinerung der Binomial-Verteilung. Wenn für einen Zufallsprozess M sich gegenseitig ausschließende Ausgänge möglich sind und der Zufallsprozess N-mal unabhängig wiederholt wird, können die Wahrscheinlichkeiten mithilfe der Multinomial-Verteilung berechnet werden. Dabei hat jedes mögliche Ereignis x_1, \dots, x_M eine Auftretenswahrscheinlichkeit von p_1, \dots, p_M . Die Wahrscheinlichkeitsfunktion eines multinomialverteilten Zufallsvektors

$$\underline{x} = (x_1, \dots, x_M) \quad (8.68)$$

ist dabei definiert durch

$$f(\underline{x}) = f(x_1, \dots, x_M) = \frac{N!}{\prod_{m=1}^{M-1} x_m! \cdot (N - \sum_{m=1}^{M-1} x_m)!} \cdot \prod_{m=1}^{M-1} p_m^{x_m} \cdot \left(1 - \sum_{m=1}^{M-1} p_m\right)^{N - \sum_{m=1}^{M-1} x_m} \quad (8.69)$$

Die durch Gleichung (8.69) definierte Multinomial-Verteilung für zwei Dimensionen mit $p_1=1/2$, $p_2=1/3$ und $p_3=1/6$ bei $N = 10$ Wiederholungen ist in Bild 8.8 zu sehen.

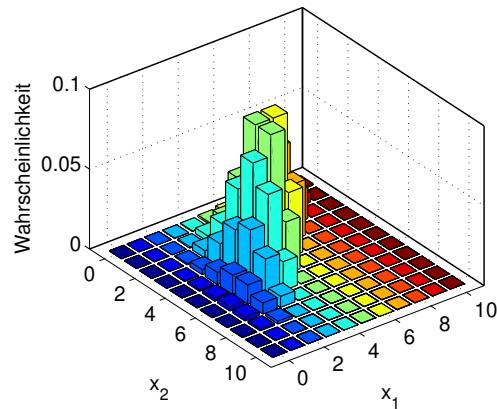


Bild 8.8: Grafische Darstellung der zweidimensionalen Multinomial-Verteilung mit $p_1=1/2$, $p_2=1/3$ und $p_3=1/6$

Für den Spezialfall von $M = 2$ führt Gleichung (8.69) zu der aus Kapitel 4 bekannten Binomial-Verteilung, die durch die Gleichung

$$f(x_1) = \frac{N!}{x_1 \cdot (N - x_1)!} \cdot p_1^{x_1} \cdot (1 - p_1)^{N - x_1} \quad (8.70)$$

beschrieben ist. Die Wahrscheinlichkeit für eine durch den Vektor \underline{x} gekennzeichnete Wertekombination berechnet sich durch

$$P(\underline{x}) = P(x_1, \dots, x_M) = \frac{N!}{\prod_{m=1}^M x_m!} \cdot \prod_{m=1}^M p_m^{x_m} \quad (8.71)$$

Das Zufallsexperiment wird N-mal ausgeführt. Damit ist die Wahrscheinlichkeit für eine Anzahl von Ereignissen

$$P\left(\sum_{m=1}^M x_m \neq N\right) = 0 \quad (8.72)$$

Die Verteilungsfunktion der Multinomial-Verteilung ergibt sich durch Summation über die Wahrscheinlichkeitsfunktion zu

$$P(\underline{\xi} \leq \underline{x}) = F(\underline{x}) = F(x_1, \dots, x_M) = \sum_{\xi_M=1}^{x_M} \dots \sum_{\xi_1=1}^{x_1} f(\xi_1, \dots, \xi_M) \quad (8.73)$$

Die Randverteilungen der Multinomial-Verteilung entsprechen den Binomialverteilungen der entsprechenden Variablen. Die Anwendung der Multinomial-Verteilung soll anhand eines Beispiels verdeutlicht werden.

Beispiel: Schaltschrankfertigung

Als Beispiel für eine Multinomial-Verteilung wird eine Schaltschrankfertigung betrachtet. In einer Woche werden an vier Montageplätzen jeweils ein Schaltschrank hergestellt und geprüft. Es ist bekannt, dass die Schaltschränke mit einer Wahrscheinlichkeit von $p_1 = 0.85$ vollständig funktionstüchtig sind, mit einer Wahrscheinlichkeit von $p_2 = 0.1$ während der Prüfung geringfügig nachgearbeitet werden müssen und mit einer Wahrscheinlichkeit von $p_3 = 0.05$ in der folgenden Woche nachgearbeitet werden müssen und nicht termingerecht ausgeliefert werden können.

Im Folgenden wird die Verteilung der Schaltschrankqualität beschrieben. Die Zufallsvariable x_1 beschreibt die Anzahl vollständig funktionstüchtiger Schaltschränke, die Zufallsvariable x_2 die Anzahl der geringfügig nachzuarbeitenden Schaltschränke und die Zufallsvariable x_3 die Anzahl der nicht termingerecht auslieferbaren Schaltschränke.

Die Gleichung zur Bestimmung der Wahrscheinlichkeiten für die einzelnen möglichen Ausgänge des Zufallsexperimentes lassen sich mit Gleichung (8.71) mit der Nebenbedingung

$$\sum_{m=1}^M x_m = N \quad (8.74)$$

berechnen. Die Wahrscheinlichkeit für eine bestimmte Kombination errechnet sich zu

$$\begin{aligned} P(x_1, x_2, x_3) &= \frac{N!}{x_1! \cdot x_2! \cdot x_3!} \cdot p_1^{x_1} \cdot p_2^{x_2} \cdot p_3^{x_3} \\ &= \frac{N!}{x_1! \cdot x_2! \cdot (4 - (x_1 + x_2))!} \cdot 0.85^{x_1} \cdot 0.1^{x_2} \cdot 0.05^{4-(x_1+x_2)} \end{aligned} \quad (8.75)$$

Mit den Zufallsgrößen x_1 , x_2 und der Beziehung $x_3 = N - x_1 - x_2$ ist es ausreichend, zwei der drei möglichen Ergebnisse der Zufallsexperimente anzugeben. Die dritte Zufallsvariable ist linear abhängig. Die Wahrscheinlichkeiten nach Gleichung (8.75) sind in Tabelle 8.5 zusammengefasst. Zusätzlich zu den Auftretenswahrscheinlichkeiten sind in Tabelle 8.5 noch die Zeilen- und Spaltensummen eingetragen, die den Randverteilungen der Zufallsvariablen x_1 und x_2 entsprechen. Diese genügen der Binomialverteilung, die bereits aus Kapitel 4 bekannt ist.

Tabelle 8.5: Wahrscheinlichkeiten des Zufallsexperimentes Schaltschrankbau

		Zufallsvariable x_2					$f_{x_1}(x_{x_1})$
		0	1	2	3	4	
Zufallsvariable	0	0.000	0.000	0.000	0.000	0.000	0.000
	1	0.000	0.003	0.005	0.003	0	0.011
	2	0.011	0.043	0.043	0	0	0.097
	3	0.123	0.246	0	0	0	0.369
	4	0.522	0	0	0	0	0.522
$f_{x_2}(x_2)$		0.656	0.292	0.048	0.003	0.000	1

Die Wahrscheinlichkeiten für das Eintreten einer bestimmten Anzahl x_1 von vollständig funktionsfähigen und einer bestimmten Anzahl x_2 von geringfügig nachzuarbeitenden Schaltschränken lassen sich grafisch darstellen. Dies ist in Bild 8.9 zu sehen.

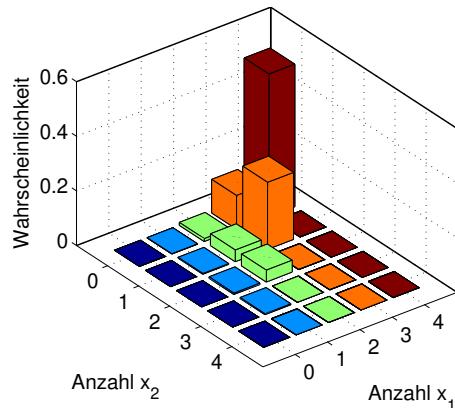


Bild 8.9: Grafische Darstellung der Wahrscheinlichkeiten aus Tabelle 8.5

Die Wahrscheinlichkeit, dass alle Schaltschränke termingerecht ausgeliefert werden können, ergibt sich aus den Fällen, in denen die Summe der beiden Zufallsvariablen x_1 und x_2 bereits 4 ergibt. Nach Tabelle 8.5 ergibt sich die Wahrscheinlichkeit

$$P(x_3 = 0) = 0 + 0.003 + 0.043 + 0.246 + 0.522 = 0.8140 \quad (8.76)$$

Die Berechnung erfolgte dabei mit MATLAB durch den folgenden Programmabschnitt.

```

1 % Definition der Kennwerte
2 p = [0.85 0.1 0.05];
3 N = 4;
4
5 % Wahrscheinlichkeit für eine termingerechte Auslieferung aller Schalschränke
6 p = mnpdf([4 0 0; 3 1 0; 2 2 0; 1 3 0; 0 4 0], p);
7 P = sum(p);

```

In Python ergibt sich analog der folgende Programmausschnitt.

```

1 %% Definition der Kennwerte %%
2 p = [0.85, 0.1, 0.05]
3 N = 4
4
5 %% Punkte für termingerechte Lieferung generieren %%
6 x = ([[4, 0, 0], [3, 1, 0], [2, 2, 0], [1, 3, 0], [0, 4, 0]]);
7
8 %% Wahrscheinlichkeit für eine termingerechte Auslieferung aller Schalschränke %%
9 f = multinomial.pmf(x, N, p)
10 P = np.sum(f)

```

8.6.2 Multivariate Normalverteilung

Bei der multivariaten Normalverteilung handelt es sich um die Verallgemeinerung der eindimensionalen Normalverteilung. Sie tritt als Grenzwert von Summen unabhängiger mehrdimensionaler Zufallsvariablen auf. Dadurch kann die multivariate Normalverteilung dort angewandt werden, wo mehrdimensionale zufällige Größen als Überlagerung von vielen voneinander unabhängigen Einzeleffekten angesehen werden können.

Die univariate Normalverteilung wird in Kapitel 4 beschrieben durch die Gleichung

$$f(x) = \frac{1}{\sigma \cdot \sqrt{2 \cdot \pi}} \cdot e^{-\frac{1}{2} \left(\frac{x - \mu}{\sigma} \right)^2} \quad (8.77)$$

Die multivariate Normalverteilung wird durch die beiden Verteilungsparameter μ und Σ definiert, die den Parametern μ und σ der univariaten Normalverteilung entsprechen. Eine Verallgemeinerung der in Gleichung (8.77) angegebene univariate Normalverteilung für M -dimensionale Stichproben der Form

$$\underline{x}^T = \begin{pmatrix} x_1 & \dots & x_M \end{pmatrix} \quad (8.78)$$

führt mit dem Mittelwertsvektor μ und der Kovarianzmatrix Σ zu der Dichtefunktion der multivariaten Normalverteilung

$$f(\underline{x}) = \frac{1}{\frac{1}{M} \cdot |\Sigma|^{\frac{1}{2}} \cdot (2 \cdot \pi)^{\frac{M}{2}}} \cdot e^{-\frac{1}{2} \cdot (\underline{x} - \mu)^T \cdot \Sigma^{-1} \cdot (\underline{x} - \mu)} \quad (8.79)$$

Wie bei der univariaten Normalverteilung liegt auf Grund der Symmetrie das Maximum der Verteilung an dem Punkt, der durch den Mittelwertsvektor μ beschrieben ist. Bild 8.10 zeigt die Dichtefunktion einer zweidimensionalen Standardnormalverteilung der Zufallsvariablen x und y mit einem Mittelwertsvektor von

$$\underline{\mu}^T = \begin{pmatrix} \mu_x & \mu_y \end{pmatrix} = \begin{pmatrix} 0 & 0 \end{pmatrix} \quad (8.80)$$

und einer Kovarianzmatrix von

$$\Sigma = \begin{pmatrix} \sigma_x^2 & \sigma_{xy} \\ \sigma_{xy} & \sigma_y^2 \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \quad (8.81)$$

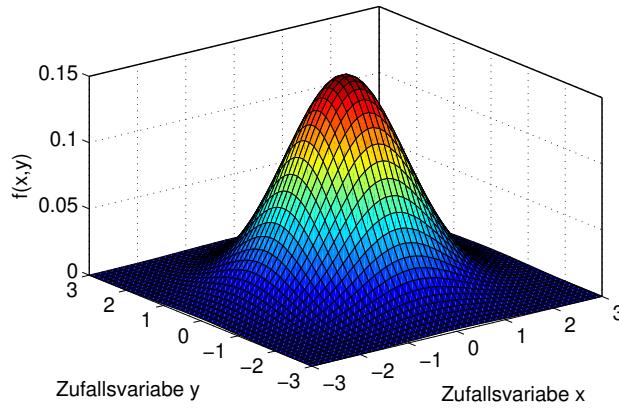


Bild 8.10: Räumliche Darstellung der Dichtefunktion der multivariaten Standardnormalverteilung

Da es bei den räumlichen Darstellungen nur schwer möglich ist, den genauen Wert der Dichtefunktion für eine bestimmte Wertekonstellation abzulesen, wird stattdessen ein Kontur-Plot verwendet, der die Werte der Dichtefunktion auf eine Ebene projiziert. Bild 8.11 zeigt die multivariate Standardnormalverteilung als Kontur-Plot.

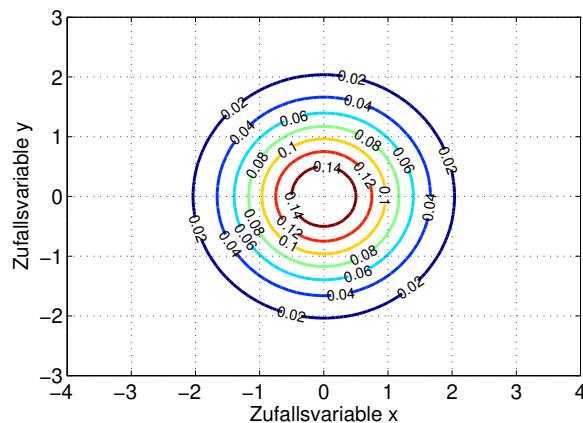


Bild 8.11: Kontur-Plot der Dichtefunktion der multivariaten Standardnormalverteilung

Die Verteilungsfunktion der Standardnormalverteilung berechnet sich über die Integration zu

$$F(X) = \int_{-\infty}^{X} \frac{1}{\frac{1}{|\Sigma|^{\frac{1}{2}}} \cdot (2\pi)^{\frac{1}{2}}} \cdot e^{-\frac{1}{2} \cdot (\Xi - M)' \cdot \Sigma^{-1} \cdot (\Xi - M)} d\Xi \quad (8.82)$$

Bild 8.12 zeigt Verteilungsfunktion der Standardnormalverteilung als räumliche Darstellung und als Kontur-Plot.

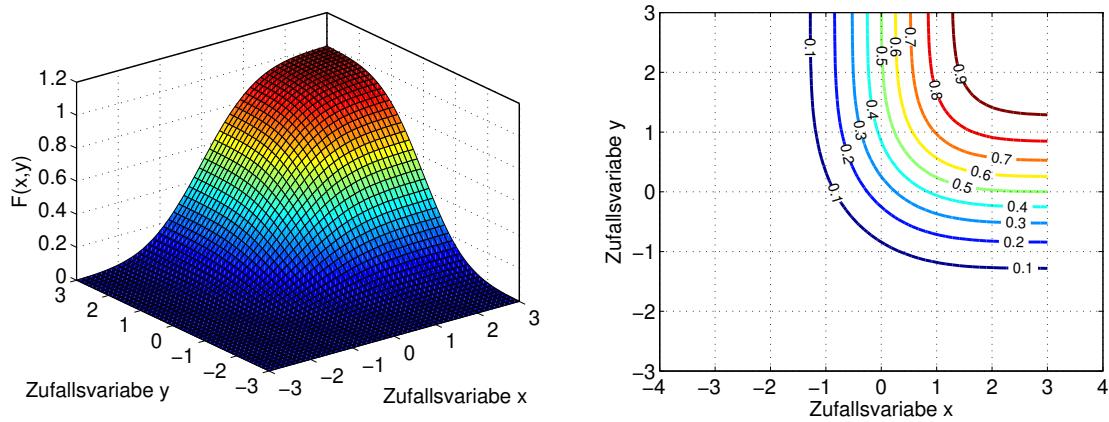


Bild 8.12: Verteilungsfunktion der multivariaten Standardnormalverteilung

Die Randverteilungen der multivariaten Normalverteilung entsprechen der univariaten Normalverteilung, in diesem Fall der univariaten Standardnormalverteilung.

Für die multivariate Standardnormalverteilung sind die Zufallsgrößen x und y unabhängig voneinander, da die Kovarianz der beiden Größen den Wert 0 annimmt.

$$\Sigma = \begin{pmatrix} \sigma_x^2 & \sigma_{xy} \\ \sigma_{xy} & \sigma_y^2 \end{pmatrix} = \begin{pmatrix} \sigma_x^2 & 0 \\ 0 & \sigma_y^2 \end{pmatrix} \quad (8.83)$$

Sind die beiden Zufallsgrößen wie bei der Standardnormalverteilung unabhängig voneinander, sind die Hauptachsen der Ellipse im Kontur-Plot parallel zu den beiden xy-Achsen ausgerichtet. Mit steigender Kovarianz verändert sich der Winkel zwischen den Hauptachsen und den Diagrammachsen, er steigt mit wachsender Kovarianz von 0 ($\sigma_{xy} = 0$) auf maximal $\pm 45^\circ$ ($\sigma_{xy} = \pm\infty$) an. Die Randverteilungen bleiben dabei unverändert. Um dies zu verdeutlichen, wird in Bild 8.13 und Bild 8.14 die multivariate Normalverteilung mit unterschiedlichen Kovarianzmatrizen dargestellt. Für die Darstellung in Bild 8.13 wurde eine zweidimensionale Normalverteilung mit einem Mittelwertsvektor von

$$\underline{\mu}^T = (\mu_x \quad \mu_y) = (0 \quad 0) \quad (8.84)$$

und einer Kovarianzmatrix von

$$\Sigma = \begin{pmatrix} 1 & 0.8 \\ 0.8 & 1 \end{pmatrix} \quad (8.85)$$

gewählt. Die beiden Zufallsvariablen besitzen demnach eine Kovarianz von $\sigma_{xy} = 0.8$. Zur Visualisierung werden sowohl die räumliche Darstellung als auch der Kontur-Plot verwendet.

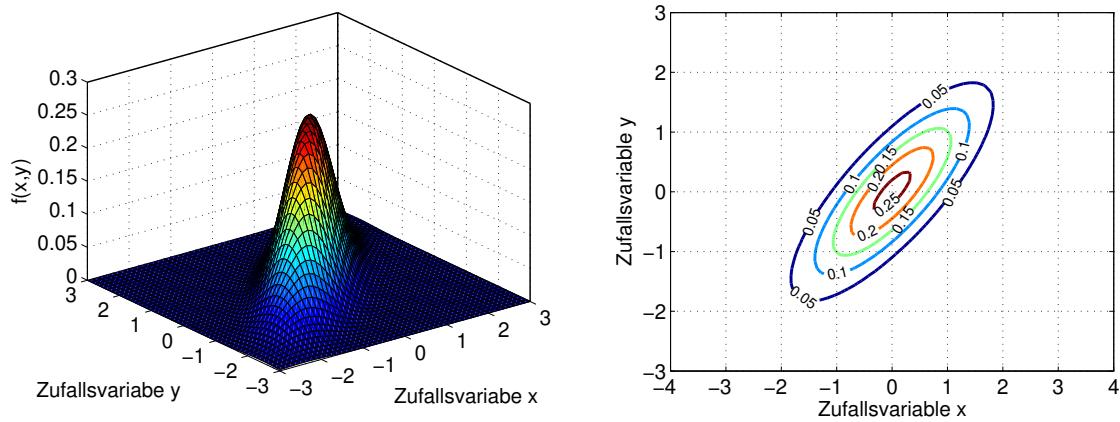


Bild 8.13: Dichtefunktion der multivariaten Standardnormalverteilung mit $\sigma_{xy} = 0.8$

Zum Vergleich zeigt Bild 8.14 die Dichtefunktion für eine Kovarianz von $\sigma_{xy} = -0.8$.

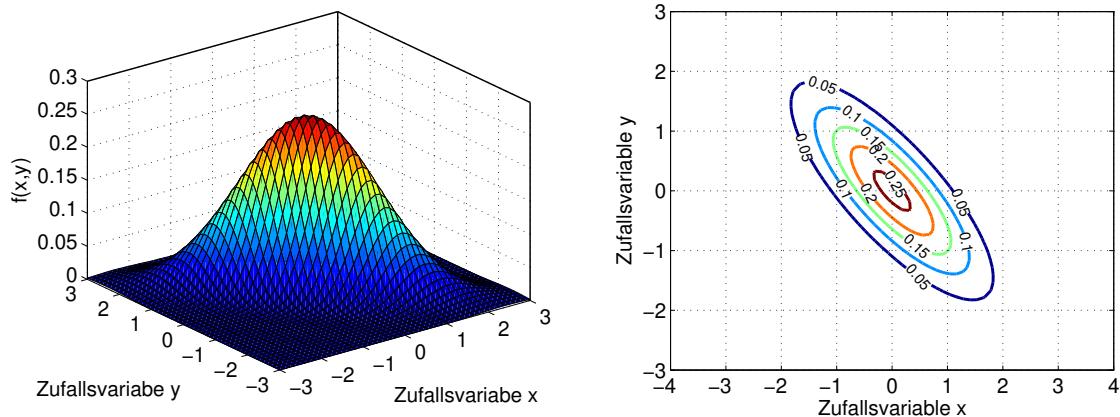


Bild 8.14: Dichtefunktion der multivariaten Standardnormalverteilung mit $\sigma_{xy} = -0.8$

Die Drehung der Dichtefunktion ist entgegengesetzt zu der vorigen Dichtefunktion in Bild 8.13. Anhand der grafischen Darstellung kann abgelesen werden, ob die beteiligten Zufallsgrößen abhängig voneinander sind und welches Vorzeichen die Kovarianz besitzt.

In den bisherigen Beispielverteilungen in Bild 8.11, Bild 8.13 oder Bild 8.14 waren die Zufallsgrößen x und y stets standardnormalverteilt. In der Praxis ist dies ohne Zentrierung der Verteilung nur selten der Fall, meist liegt zusätzlich eine Standardabweichung σ ungleich 1 vor.

In Bild 8.15 wird gezeigt, wie sich die multivariate Verteilungsfunktion ändert, wenn die Verteilung der Zufallsvariable x durch eine Normalverteilung mit einem Mittelwert von $\mu_x = 0$ und einer Varianz von $\sigma_{xy} = 1.8$ beschrieben werden kann. Die Zufallsvariable y wird als standardnormalverteilt beibehalten. Beide Zufallsgrößen sind voneinander unabhängig.

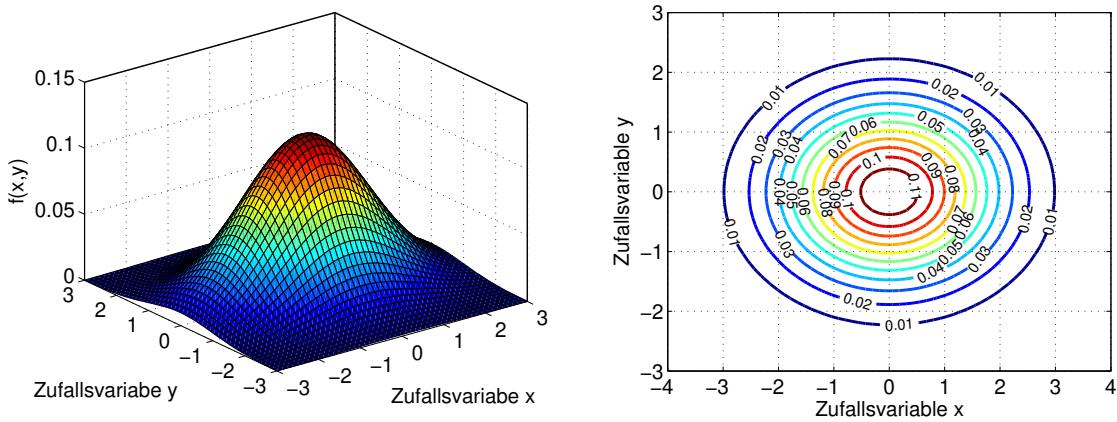


Bild 8.15: Dichtefunktion der multivariaten Standardnormalverteilung mit $\sigma_x^2 = 1, \sigma_y^2 = 0.5, \sigma_{xy} = 0$

Es ist zu erkennen, dass die Verteilung in Richtung der Zufallsvariablen x aufgrund der größeren Streuung breiter geworden ist.

Bild 8.16 zeigt die Änderung der multivariaten Verteilungsfunktion, wenn die Verteilung der Zufallsvariablen y mit einer Normalverteilung mit einem Mittelwert von $\mu_y = 0$ und einer Varianz von $\sigma_y^2 = 0.5$ angenommen wird. Zur Veranschaulichung wird für die Darstellung die Verteilung der Zufallsvariablen x wieder als standardnormalverteilt angenommen. Beide Zufallsgrößen sind voneinander unabhängig.

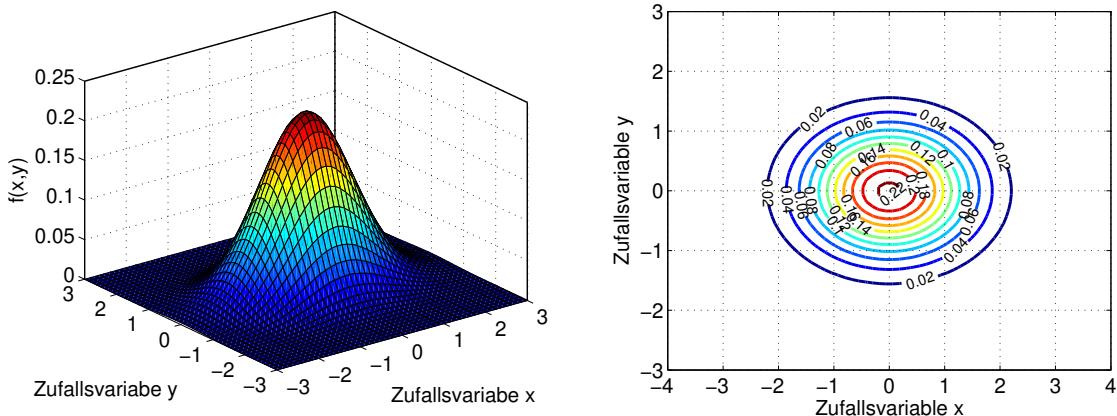


Bild 8.16: Dichtefunktion der multivariaten Standardnormalverteilung mit $\sigma_x^2 = 1, \sigma_y^2 = 0.5, \sigma_{xy} = 0$

In der Grafik ist gut zu erkennen, dass sich die Verteilung in Richtung der Zufallsvariablen y verschmälert hat. Die multivariate Normalverteilung wird an einem Beispiel vertieft.

Beispiel: Qualitätsbewertung von Temperaturwiderständen

Als Beispiel für die multivariate Normalverteilung wird die Fertigung eines temperaturabhängigen Widerstandes untersucht. Er besitzt einen Widerstandswert von

$$R(T) = R_{20} \cdot (1 + \alpha \cdot (T - T_{20})) \quad (8.86)$$

Es ist bekannt, dass der Widerstandswert R bei 20°C durch eine Normalverteilung mit einem Mittelwert von $\mu_R = 1000 \Omega$ und einer Standardabweichung von $\sigma_R = 5 \Omega$ beschrieben werden kann. Ebenso ist bekannt, dass auch der Temperaturkoeffizient α des Widerstandes durch eine Normalverteilung abgebildet werden kann. Der mittlere Wert des Temperaturkoeffizienten liegt bei $\mu = 3.85 \cdot 10^{-3}/K$, die Standardabweichung beträgt $\sigma_C = 0.51 \cdot 10^{-3}/K$. Die Kovarianz der beiden Größen beträgt $\sigma_{RC} = 1.63 m\Omega/K$. Mit den gegebenen Angaben ergibt sich der Mittelwertsvektor

$$\underline{\mu}^T = \begin{pmatrix} \mu_R & \mu_C \end{pmatrix} = \begin{pmatrix} 1000 \Omega & 3.85 \cdot 10^{-3}/K \end{pmatrix} \quad (8.87)$$

und die Kovarianzmatrix

$$\Sigma = \begin{pmatrix} \sigma_R^2 & \sigma_{RC} \\ \sigma_{RC} & \sigma_C^2 \end{pmatrix} = \begin{pmatrix} 25\Omega^2 & 1.63m\Omega/K \\ 1.63\Omega/K & 0.2510^{-6}/K^2 \end{pmatrix} \quad (8.88)$$

Durch die beiden Parameter μ und Σ ist die multivariate Normalverteilung ausreichend spezifiziert. Die dazugehörige Dichteverteilung ist in Bild 8.17 zu sehen.

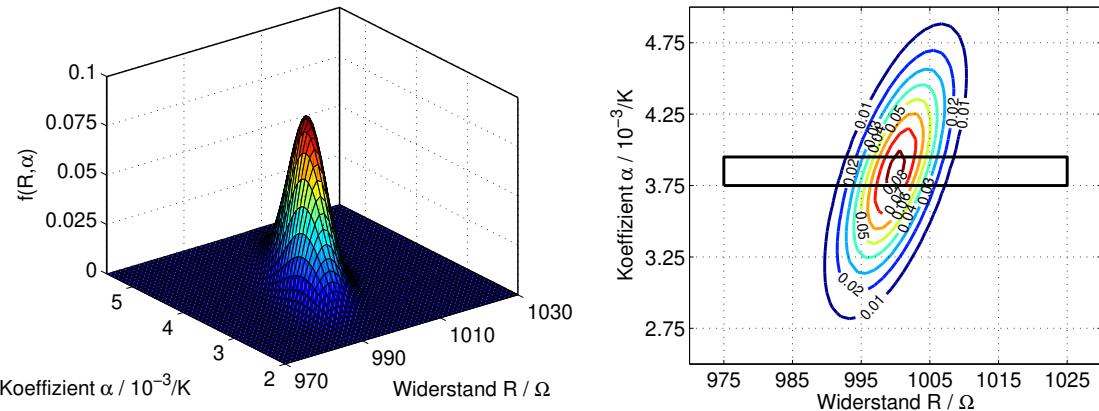


Bild 8.17: Dichtefunktion der Temperaturwiderstände

Bei der vorliegenden Widerstandsproduktion wird bewertet, wie viel Prozent der produzierten Temperaturwiderstände Ausschussware sind, wenn mit einem Kunden A eine Toleranz des Widerstandswertes und des Temperaturkoeffizienten von jeweils $\pm 2.5\%$ vereinbart wird. Der sich ergebende Bereich, in dem die Widerstände für Kunde A liegen dürfen, ist exemplarisch in dem Kontur-Plot in Bild 8.17 eingezzeichnet. Mit einem Kunden B wird stattdessen eine erweiterte Toleranzgrenze von $\pm 5\%$ vereinbart. Die Berechnung erfolgt mithilfe der Verteilungsfunktion, die in Bild 8.18 abgebildet ist.

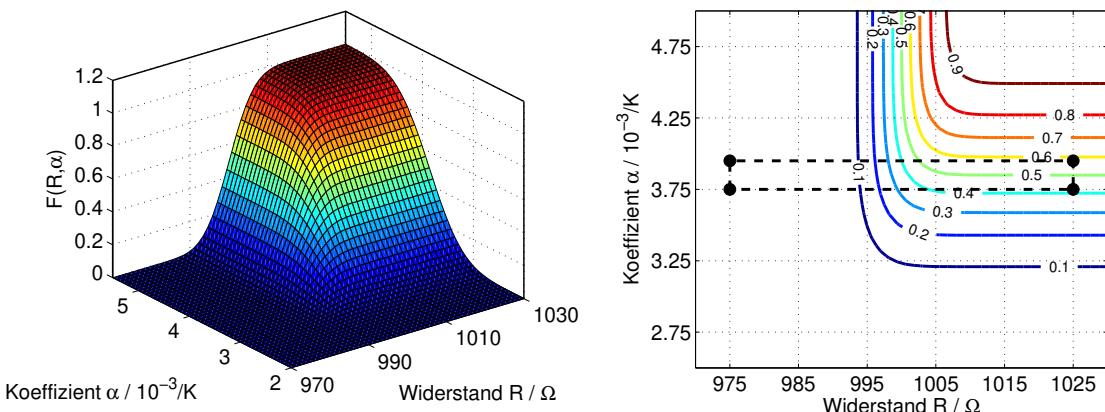


Bild 8.18: Verteilungsfunktion der Temperaturwiderstände

Die gesuchte Wahrscheinlichkeit für den Ausschussanteil bei Kunde A kann als die Fläche außerhalb des in Bild 8.17 eingezeichneten Quadrates verstanden werden. Mithilfe der Verteilungsfunktion aus Bild 8.18 kann die Wahrscheinlichkeit p für eine Fertigung innerhalb der Toleranzen berechnet werden zu

$$\begin{aligned} p &= F(1025\Omega, 3.95 \cdot 10^{-3}/K) - F(975\Omega, 3.95 \cdot 10^{-3}/K) \\ &\quad - F(1025\Omega, 3.75 \cdot 10^{-3}/K) + F(975\Omega, 3.75 \cdot 10^{-3}/K) \\ &= 15.85\% \end{aligned} \quad (8.89)$$

Der Ausschussanteil ergibt sich somit für Kunde A zu

$$A_A = 1 - p = 84.15\% \quad (8.90)$$

Da für Kunde B Toleranzgrenzen von $R = 950 \dots 1050\Omega$ und $\alpha = 3.66 \dots 4.04 \cdot 10^{-3}/K$ vereinbart wurden, verringert sich der Ausschussanteil der gefertigten Widerstände auf

$$A_B = 70.39\% \quad (8.91)$$

Die Berechnung erfolgte dabei mit MATLAB durch den folgenden Programmabschnitt.

```

1 % Definition der Kennwerte
2 Mu = [1000 3.85e-3];
3 Sigma = [25 1.63e-3; 1.63e-3 0.25e-6];
4 pF_A = mvncdf([1025 3.95e-3; 975 3.95e-3; 1025 3.75e-3; 975 3.75e-3], Mu,
    Sigma);
5 pF_B = mvncdf([1050 4.04e-3; 950 4.04e-3; 1050 3.66e-3; 950 3.66e-3], Mu,
    Sigma);
6
7 % Prozentualer Ausschuss von Kunde A
8 Ausschuss_A = 1 - (pF_A(1) - pF_A(2) - pF_A(3) + pF_A(4));
9 Ausschuss_B = 1 - (pF_B(1) - pF_B(2) - pF_B(3) + pF_B(4));

```

Mit Python kann die Aufgabe mit dem folgenden Programmabschnitt gelöst werden.

```

1 %%Bibliotheken importieren%%
2 from scipy.stats import multivariate_normal
3 import numpy as np
4
5 %%Werte aus Aufgabe übernehmen%%
6 muR = 1000
7 varR = 25
8 muAlpha = 3.85e-3
9 varAlpha = 0.25e-6
10 covRAlpha = 1.63e-3
11
12 %%Definition der entsprechenden Verteilung%%
13 rv = multivariate_normal([muR, muAlpha], [[varR, covRAlpha],
14 [covRAlpha, varAlpha]])
15
16 %%Berechnung der gesuchten Wahrscheinlichkeit%%
17 PA = rv.cdf([1025, 3.95e-3]) - rv.cdf([975, 3.95e-3]) - rv.cdf([1025, 3.75e
18 -3]) + rv.cdf([975, 3.75e-3])
19 PB = rv.cdf([1050, 4.04e-3]) - rv.cdf([950, 4.04e-3]) - rv.cdf([1050, 3.66e
20 -3]) + rv.cdf([950, 3.66e-3])
21
22 %%Ausgabe%%
23 print(' ')
24 print('Ausschuss Kunde A: ', 1-PA)
25 print('Ausschuss Kunde B: ', 1-PB)

```

8.7 Literatur

- [Krey91] Kreyszig, Erwin: Statistische Methoden und ihre Anwendungen
4., unveränderter Nachdruck der 7. Auflage
Vandenhoeck & Ruprecht, Göttingen, 1991
- [Fahr96] Fahrmeir, Ludwig; Hamerle, Alfred; Tutz, Gerhard: Multivariate statistische Verfahren
2., überarbeitete Auflage
Walter de Gruyter & Co., Berlin
- [Ross06] Ross, M. Sheldon: Statistik für Ingenieure und Naturwissenschaftler
3. Auflage
Spektrum Akademischer Verlag, München, 2006
- [Hart07] Hartung, Joachim; Elpelt, Bärbel: Multivariate Statistik
7., unveränderte Auflage
R. Oldenbourg Verlag, München / Wien
- [Papu01] Papula, Lothar: Mathematik für Ingenieure und Naturwissenschaftler Band 3
4., verbesserte Auflage
Vieweg Teubner, Braunschweig / Wiesbaden, 2008

9 Varianzanalyse

Die Verfahren zur Datenanalyse multivariater Daten sind von dem Merkmalstyp des Eingangssignals und der Zielgröße abhängig. Korrelationsfunktionen und Regressionsfunktionen beschreiben die Abhängigkeit von Zielgrößen als Funktion kontinuierlicher oder diskreter Eingangsgrößen. Leider versagt diese mathematische Beschreibung bei ordinalen oder gruppierenden Eingangsgrößen. Zum Beispiel lässt sich die Frage, ob Spritzgussteile aus unterschiedlichen Formnestern A - D gleiche Abmessungen besitzen, nicht mit Regressionsfunktionen beantworten. Mithilfe eines Hypothesentests könnte bewertet werden, ob die Teile aus zwei Formnestern dieselbe Geometrie haben. Die vorgestellten Hypothesentests versagen jedoch bei mehr als zwei Gruppen. Die Varianzanalyse (ANOVA, Analysis of Variance) schließt diese Lücke und bewertet den Einfluss einer oder mehrerer ordinaler Eingangsgrößen auf eine diskrete oder stetige Ausganggröße.

Die Varianzanalyse wird mit einem Beispiel eingeführt. Bei der Fertigung von Bolzen muss sichergestellt werden, dass unterschiedliche Fertigungseinrichtungen zu beliebigen Zeitpunkten dieselbe Qualität liefern. Als kritisches Qualitätsmerkmal der Bolzen wird der Bolzendurchmesser d gemessen, der normalen Fertigungsschwankungen unterliegt. Bei der Auswertung jeweils einer Stichprobe pro Fertigungscharge variiert der Mittelwert der Stichproben. Deshalb ist ein direkter Vergleich der Stichprobenmittelwerte von Charge zu Charge nicht aussagekräftig. Bild 9.1 verdeutlicht diesen Zusammenhang für vier Fertigungschargen 1 - 4.

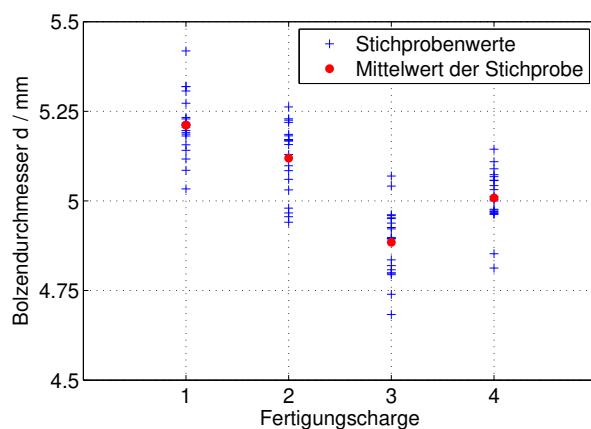


Bild 9.1: Streuung der Stichprobenwerte bei der Fertigung von Bolzen und unterschiedlichen Fertigungschargen

Um zu entscheiden, ob die unterschiedlichen Chargen einen signifikanten Unterschied besitzen, muss versucht werden, den Einfluss der unterschiedlichen Chargen von der typischen Varianz des Prozesses zu trennen. Dies ist eine Aufgabe der einfaktoriellen Varianzanalyse. Die einfaktorielle Varianzanalyse prüft, ob zumindest eine Stichprobe signifikant von den anderen abweicht.

Bei der Optimierung von Fertigungsprozessen werden teilweise auch ganz gezielt Fertigungsparameter geändert. In diesem Fall wird der Einfluss unterschiedlicher Parameteränderung auf die Zielgröße untersucht. Sollen M Einflussparameter gleichzeitig untersucht werden, müssen die von den Einflussfaktoren hervorgerufenen Varianzen und die normale Fertigungsvarianz voneinander getrennt werden. Dies ist eine Aufgabe der M -faktoriellen Varianzanalyse, die im Folgenden als ein- und zweifaktorielle Varianzanalyse hergeleitet und dann auf eine mehrfaktorielle Varianzanalyse verallgemeinert wird.

Diese Aufgabenstellungen werden bei der klassischen Varianzanalyse unter der Annahme bearbeitet, dass die Gruppen von Zahlen aus normalverteilten Grundgesamtheiten entstammen, die alle dieselbe Varianz σ^2 besitzen. Die Varianz σ^2 muss dabei nicht bekannt sein. Diese Annahme ist durch geeignete Hypothesentests vor Durchführen einer Varianzanalyse zu prüfen. Sind die Annahmen nicht erfüllt, müssen parameterfreie Tests durchgeführt werden.

9.1 Einfaktorielle Varianzanalysen

Bei einer einfaktoriellen Varianzanalyse wird der Einfluss einer ordinalen Eingangsgröße auf eine Zielgröße untersucht. Zur Darstellung der einfaktoriellen Varianzanalyse wird ein Fertigungsprozess betrachtet. In regelmäßigen Abständen wird auf Basis von Stichproben bewertet, ob sich die Zielwerte der vorliegenden Charge signifikant von den Werten der bisher gefertigten Chargen unterscheiden.

9.1.1 Datenstruktur und Modellansatz der einfaktoriellen Varianzanalyse

Für die Bewertung werden J Fertigungslose untersucht, für jedes Fertigungslos sind die notwendigen Fertigungsparameter bekannt. Aus jedem Fertigungslos werden N Stichprobenwerte bestimmt.

Tabelle 9.1: Nomenklatur zur Bezeichnung der Stichprobenwerte in Gruppen

1.Fertigungslos	$\cdot j \cdot$	J.Fertigungslos
$x_{11} \cdot x_{1n} \cdot x_{1N}$	$x_{j1} \cdot x_{jn} \cdot x_{jN}$	$x_{J1} \cdot x_{Jn} \cdot x_{JN}$

Der erste Index j der Größe x_{jn} bezeichnet dabei die Nummer j des Fertigungsloses, das allgemein als Gruppe bezeichnet wird. Jede der J Gruppen besteht aus N Stichprobenwerten. Geprüft werden soll, ob hinsichtlich des Mittelwertes der Zielgröße bei den J Gruppen signifikante Unterschiede bestehen, die durch die unterschiedlichen Fertigungsparameter hervorgerufen wurden oder ob nur zufallsbedingte Streuungen vorliegen. Bestehen nur zufallsbedingte Unterschiede, ist es gleichgültig, mit welchen Fertigungsparametern die Bauteile gefertigt wurden. Weist ein Fertigungsparameter einen signifikanten Einfluss auf den Zielwert auf, kann er zum Beispiel zur Optimierung der Fertigungsausbeute oder zur Qualitätsverbesserung verwendet werden.

Es wird vorausgesetzt, dass die J Gruppen von Stichproben aus J normalverteilten Grundgesamtheiten stammen, die alle dieselbe Varianz σ^2 besitzen. Der einzelne Stichprobenwert x_{jn} kann dann als Summe eines konstanten Mittelwertes μ , einer Abweichung zwischen den einzelnen Fertigungslosen α_j und einer zufälligen, normalverteilten Abweichung ϵ_{jn} dargestellt werden. Dabei wird davon ausgegangen, dass die Variablen μ , α_j und ϵ_{jn} voneinander unabhängig sind.

$$x_{jn} = \mu + \alpha_j + \epsilon_{jn} \quad (9.1)$$

Damit kann die Varianzanalyse als Test der Hypothese aufgefasst werden, dass die J Gruppen aus derselben Grundgesamtheit mit dem Mittelwert μ stammen. Trifft diese Hypothese zu, besteht kein signifikanter Einfluss der Fertigungsparameter und die Abweichungen zwischen den einzelnen Fertigungslosen α_j sind null.

$$\alpha_1 = \alpha_2 = \alpha_3 = \dots = \alpha_J = 0 \quad (9.2)$$

Die Gegenhypothese ist entsprechend, dass zumindest die Fertigungsparameter einer Stichprobe einen signifikanten Einfluss besitzen und damit gilt

$$\alpha_j \neq 0 \quad (9.3)$$

Für $J = 2$ Gruppen wird diese Aufgabe bei dem Hypothesentest zum Test auf gleichen Mittelwert bei unbekannter Varianz in Kapitel 6 behandelt. Für $J > 2$ Gruppen wird die Aufgabe mit der Varianzanalyse gelöst. Hierbei wird davon ausgegangen, dass eine große Streuung zwischen den Gruppen bei

gleichzeitig geringer Streuung innerhalb der Gruppen auf eine große Signifikanz des zu untersuchenden Prozessparameters hinweist. Damit wird der Hypothesentest abgelehnt, wenn die Streuung der Mittelwerte der Stichproben untereinander größer ist als die Streuung innerhalb der Stichproben.

In dem Modellansatz aus Gleichung (9.1) wird ein einzelner Stichprobenwert als Summe eines konstanten Mittelwertes μ , einer Abweichung zwischen den einzelnen Fertigungslosen α_j und einer zufälligen, normalverteilten Abweichung ϵ_{jn} dargestellt. Beide Größen sind definitionsgemäß voneinander unabhängig. Aus Abschnitt 8.4.1 ist bekannt, dass die Varianz der Summe von unabhängigen Zufallsgrößen als Summe der einzelnen Varianzen berechnet wird. Damit ergibt sich direkt aus dem Ansatz in Gleichung (9.1) der Modellansatz der Varianzen aus den Stichprobenwerten der J Fertigungslose zu

$$s_x^2 = s_\mu^2 + s_\alpha^2 + s_\epsilon^2 = s_\alpha^2 + s_\epsilon^2 \quad (9.4)$$

Ziel des folgenden Abschnitts ist es, diese Varianzen herzuleiten.

9.1.2 Berechnung der einzelnen Varianzen über die standardisierten Quadratsummen

Mathematisch gesehen beruht die Berechnung der einzelnen Varianzen auf der Berechnung von Quadratsummen, die mit ihrer entsprechenden Anzahl von Freiheitsgraden normiert werden. Sie werden auch als standardisierte Quadratsummen bezeichnet. Bei der einfaktoriellen Varianzanalyse sind dies die standardisierten Quadratsummen M und M.

Die Quadratsumme q_x , die die Abweichungen aller Stichprobenwerte von dem Gesamtmittelwert \bar{x} beschreibt

$$q_x = \sum_{j=1}^J \sum_{n=1}^N (x_{jn} - \bar{x})^2 \quad (9.5)$$

kann in Anteile zerlegt, die der entsprechenden Variationsursache entsprechen.

$$\begin{aligned} q_x &= \sum_{j=1}^J \sum_{n=1}^N (x_{jn} - \bar{x})^2 = \sum_{j=1}^J \sum_{n=1}^N (x_{jn} - \bar{x}_j + \bar{x}_j - \bar{x})^2 \\ &= \sum_{j=1}^J \sum_{n=1}^N (x_{jn} - \bar{x}_j)^2 + 2 \cdot (x_{jn} - \bar{x}_j) \cdot (\bar{x}_j - \bar{x}) + (\bar{x}_j - \bar{x})^2 \\ &= N \cdot \sum_{j=1}^J (\bar{x}_j - \bar{x})^2 + \sum_{n=1}^N \sum_{j=1}^J (x_{jn} - \bar{x}_j)^2 + 2 \cdot \sum_{n=1}^N \sum_{j=1}^J (x_{jn} - \bar{x}_j) \cdot (\bar{x}_j - \bar{x}) \end{aligned} \quad (9.6)$$

Dabei berechnet sich der Mittelwert der j-ten Gruppe zu

$$\bar{x}_j = \frac{1}{N} \cdot \sum_{n=1}^N x_{jn} \quad (9.7)$$

und der Gesamtmittelwert der Stichprobenwerte aller Gruppen zu

$$\bar{x} = \frac{1}{J \cdot N} \cdot \sum_{j=1}^J \sum_{n=1}^N x_{jn} = \frac{1}{J} \cdot \sum_{j=1}^J \bar{x}_j \quad (9.8)$$

Der erste Summand in Gleichung (9.6)

$$q_\alpha = N \cdot \sum_{j=1}^J (\bar{x}_j - \bar{x})^2 \quad (9.9)$$

beschreibt die Streuung zwischen den Gruppen, indem er die Mittelwerte der Stichproben \bar{x}_j von den unterschiedlichen Gruppen mit dem Gesamtmittelwert aller Stichproben \bar{x} vergleicht. Die zweite Summe beschreibt die Streuung innerhalb der Gruppen

$$q_\varepsilon = \sum_{j=1}^J \sum_{n=1}^N (x_{jn} - \bar{x}_j)^2 \quad (9.10)$$

Der dritte Summand kann umgerechnet werden zu

$$\begin{aligned} q_0 &= 2 \cdot \sum_{j=1}^J \sum_{n=1}^N (x_{jn} - \bar{x}_j) \cdot (\bar{x}_j - \bar{x}) = 2 \cdot \sum_{j=1}^J \sum_{n=1}^N (x_{jn} \cdot \bar{x}_j - x_{jn} \cdot \bar{x} - \bar{x}_j \cdot \bar{x}_j + \bar{x}_j \cdot \bar{x}) \\ &= 2 \cdot \sum_{j=1}^J \sum_{n=1}^N x_{jn} \cdot \bar{x}_j - 2 \cdot N \cdot \sum_{j=1}^J \bar{x}_j \cdot \bar{x} - 2 \cdot \sum_{j=1}^J \sum_{n=1}^N x_{jn} \cdot \bar{x}_j + 2 \cdot N \cdot \sum_{j=1}^J \bar{x}_j \cdot \bar{x} = 0 \end{aligned} \quad (9.11)$$

Die Quadratsumme aus Gleichung (9.6) kann bei der einfaktoriellen Varianzanalyse damit in zwei Gruppen zerlegt werden, die die Streuung von Gruppe zu Gruppe beziehungsweise die Streuung innerhalb der Gruppen repräsentiert.

$$q_x = q_\alpha + q_\varepsilon \quad (9.12)$$

Analog zu der Berechnung der Stichprobenvarianz in 3.3.2 müssen die berechneten Quadratsummen q und q auf ihre entsprechende Anzahl von Freiheitsgraden v normiert werden, um die gesuchten Varianzen zu erhalten. Bei der Berechnung der Gesamtstreuung wird nur ein Mittelwert benötigt. Damit ergibt sich die standardisierte Quadratsumme durch Bezug auf die Anzahl der Stichproben $J \cdot N$ abzüglich eines Freiheitsgrades für einen festgelegten Parameter. Die Gesamtvarianz errechnet sich damit zu

$$M_x = s_x^2 = \frac{q_x}{v_x} = \frac{q_\alpha + q_\varepsilon}{J \cdot N - 1} \quad (9.13)$$

Die Berechnung der standardisierten Quadratsumme für den Einfluss der einzelnen J Gruppen ergibt sich analog durch Bezug auf die Anzahl der Gruppen J abzüglich eines Freiheitsgrades für den erforderlichen Mittelwert.

$$M_\alpha = s_\alpha^2 = \frac{q_\alpha}{v_\alpha} = \frac{N \cdot \sum_{j=1}^J (\bar{x}_j - \bar{x})^2}{J - 1} \quad (9.14)$$

Für den Einfluss der Reststreuung müssen J Mittelwerte bekannt sein, sodass $J \cdot N - J$ Freiheitsgrade vorliegen.

$$M_\varepsilon = s_\varepsilon^2 = \frac{q_\varepsilon}{v_\varepsilon} = \frac{\sum_{j=1}^J \sum_{n=1}^N (x_{jn} - \bar{x}_j)^2}{J \cdot N - J} \quad (9.15)$$

Die berechneten Streuungskennwerte der Einzeleinflüsse können damit auf ihre Signifikanz hin untersucht werden.

9.1.3 Signifikanzbewertung der einzelnen Einflüsse

Die Bewertung der Varianzen der einzelnen Einflüsse erfolgt über das Verhältnis der im Abschnitt 9.1.2 berechneten Varianzen. Dabei wird das Verhältnis der Streuung zwischen den Gruppen s_α^2 zu der Streuung innerhalb der Gruppen s_ε^2 bewertet. Mit den Grundlagen aus 5.4.3 ist bekannt, dass die Zufallsvariable

$$v_0 = \frac{s_\alpha^2}{s_\varepsilon^2} = \frac{\frac{q_\alpha}{J-1}}{\frac{q_\varepsilon}{J \cdot (N-1)}} \quad (9.16)$$

eine F-Verteilung mit den Freiheitsgraden $(J - 1, J \cdot (N - 1))$ aufweist. Liegt ein signifikanter Einfluss vor, wird die Varianz s^2 groß und der Quotient v_0 steigt an. Ist der Einfluss nicht signifikant, wird der Quotient v_0 klein sein. Die Entscheidung auf Signifikanz beruht deshalb auf einen Vergleich des Quotienten der Varianzen v_0 mit einer Grenze c , die sich aus der F-Verteilung mit $(J - 1, J \cdot (N - 1))$ Freiheitsgraden und dem Signifikanzniveau α ergibt.

$$P(v \leq c) = \gamma = 1 - \alpha \quad (9.17)$$

Damit ergibt sich, dass im Fall $v v_0 \leq c$ die Hypothese der gleichen Mittelwerte $\alpha_1 = \alpha_2 = \dots = \alpha_J = 0$ angenommen wird, ist $v_0 > c$ wird die Hypothese verworfen. Muss die Nullhypothese verworfen werden, kann auf Basis der vorliegenden Stichprobenwerte davon ausgegangen werden, dass die Gruppen einen signifikanten Unterschied besitzen.

Alternativ kann der p-Wert, der zu der Variable v_0 gehört, mit dem Signifikanzniveau $\alpha = 1 - \gamma$ verglichen werden. Bild 9.2 veranschaulicht den p-Wert bei der einfaktoriellen Varianzanalyse grafisch.

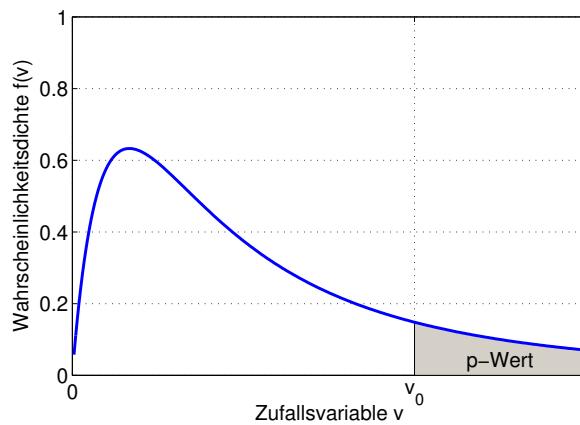


Bild 9.2: p-Wert bei der einfaktoriellen Varianzanalyse

Liegt der p-Wert oberhalb der Grenze α , kann die Hypothese der gleichen Mittelwerte nicht verworfen werden. Der entsprechende Einflussfaktor wäre entsprechend nicht signifikant.

$$p = 1 - F(v_0) = 1 - F\left(\frac{s_\alpha^2}{s_\epsilon^2}\right) \quad (9.18)$$

Die Varianzanalyse wird typischerweise als ANOVA-Tabelle dargestellt. Dabei steht ANOVA für Analysis of Variance. Tabelle 9.2 zeigt den allgemeinen Aufbau einer einfaktoriellen ANOVA-Tabelle.

Tabelle 9.2: Zusammenfassung der einfaktoriellen Varianzanalyse als ANOVA-Tabelle

Streuungsquelle	Quadratsumme	Freiheitsgrade	Standardisierte Quadratsumme	Wert der Testvariable	p-Wert
Zwischen den Gruppen	q_α	$J - 1$	$s_\alpha^2 = \frac{q_\alpha}{J - 1}$	$v_0 = \frac{s_\alpha^2}{s_\epsilon^2}$	$P(v > v_0)$
Innerhalb der Gruppen	q_ϵ	$J \cdot (N - 1)$	$s_\epsilon^2 = \frac{q_\epsilon}{J \cdot (N - 1)}$		
Gesamtstreuung	q_x	$J \cdot N - 1$			

Das Vorgehen zu diesem Test ist in Tabelle 9.3 dargestellt.

Tabelle 9.3: Test der Hypothese, dass die normalverteilten Grundgesamtheiten gleicher Varianz, aus denen J Gruppen stammen, alle denselben Mittelwert besitzen

Nr.	Prozessschritt	
1	Wahl eines Signifikanzniveaus α	
2	Bestimmung des zugehörigen Parameters c aus der inversen F-Verteilung mit $(J - 1, J \cdot N - J)$ Freiheitsgraden $F(c) = 1 - \alpha$	
3	Berechnung der J Mittelwerte der Gruppen und des Mittelwertes der gesamten Stichprobe $\bar{x}_j = \frac{1}{N} \cdot \sum_{n=1}^N x_{jn}$ $\bar{x} = \frac{1}{J \cdot N} \cdot \sum_{j=1}^J \sum_{n=1}^N x_{jn} = \frac{1}{J} \cdot \sum_{j=1}^J \bar{x}_j$	
4	Berechnung der Varianz zwischen den Mittelwerten der Gruppe und des Mittelwertes der gesamten Stichprobe $s_\alpha^2 = \frac{N \cdot \sum_{j=1}^J (\bar{x}_j - \bar{x})^2}{J - 1}$ $s_\varepsilon^2 = \frac{q_\varepsilon}{v_\varepsilon} = \frac{\sum_{j=1}^J \sum_{n=1}^N (x_{jn} - \bar{x}_j)^2}{J \cdot N - J}$	
5	Bestimmung des Annahmebereichs $\frac{s_\alpha^2}{s_\varepsilon^2} < c$	Berechnung des p-Values mit der f-Verteilung $p = 1 - F\left(\frac{s_\alpha^2}{s_\varepsilon^2}\right)$
6	Für $s_\alpha^2/s_\varepsilon^2 < c$ wird die Hypothese angenommen, für $s_\alpha^2/s_\varepsilon^2 \geq c$ wird die Hypothese verworfen	Für $p < 1 - \alpha$ wird die Hypothese angenommen, für $p \geq 1 - \alpha$ wird die Hypothese verworfen

Beispiel: Kondensatorfertigung

Das Vorgehen wird anhand eines Beispiels verdeutlicht, in dem eine Kondensatorfertigung untersucht wird. Auf $J = 3$ Fertigungseinrichtungen mit gleichen Fertigungsparametern werden 47 nF - Kondensatoren gefertigt. Mithilfe einer einfaktoriellen Varianzanalyse soll festgestellt werden, ob die Fertigungsrichtung einen signifikanten Einfluss auf den Kapazitätswert besitzt. Dazu werden je Fertigungseinrichtung $N = 4$ Kondensatoren untersucht, und es ergeben sich die in Tabelle 9.4 gezeigten Messwerte.

Tabelle 9.4: Stichprobe zur Überprüfung der Fertigungseinrichtungen in der Kondensatorenfertigung

C / nF	Fertigungseinrichtung 1	Fertigungseinrichtung 2	Fertigungseinrichtung 3
Stichproben	46.60	47.56	48.20
	48.20	47.24	47.56
	43.74	40.24	45.30
	46.60	46.60	47.88
Gruppenmittelwert	46.29	45.41	47.24
Gesamtmittelwert		46.31	

Die Gruppenmittelwerte und der Mittelwert der gesamten Stichprobe sind bereits in die Tabelle eingetragen. Die Quadratsumme zwischen den Gruppen ist

$$q_\alpha = N \cdot \sum_{j=1}^J (\bar{x}_j - \bar{x})^2 = 4 \cdot (0.02^2 + 0.9^2 + 0.93^2) = 6.665 \quad (9.19)$$

Die Quadratsumme für die Varianz innerhalb der Gruppen ergibt sich aus

$$q_\varepsilon = \sum_{j=1}^J \sum_{n=1}^N (x_{jn} - \bar{x}_j)^2 = (46.60 - 46.29)^2 + \dots + (47.88 - 47.24)^2 = 51.656 \quad (9.20)$$

Damit berechnet sich die Gesamtquadratsumme zu

$$q_x = q_\alpha + q_\varepsilon = 6.665 + 51.656 = 58.321 \quad (9.21)$$

Mit den berechneten Zahlenwerten lässt sich der Quotient v_0 der Stichprobe bestimmen zu

$$v_0 = \frac{s_\alpha^2}{s_\varepsilon^2} = \frac{\frac{q_\alpha}{J-1}}{\frac{q_\varepsilon}{J \cdot (N-1)}} = \frac{\frac{6.665}{3-1}}{\frac{51.656}{3 \cdot (4-1)}} = 0.58062 \quad (9.22)$$

Mit dem Signifikanzniveau $\alpha = 0.05$, den Freiheitsgraden $(J-1) = 2$ beziehungsweise $J \cdot (N-1) = 9$ und der inversen F-Verteilung ergibt sich für

$$F(c) = 1 - \alpha = 0.95 \quad (9.23)$$

die kritische Grenze $c = 4.26$. Der Vergleich mit v_0 zeigt, dass $v_0 < c$ ist. Die Hypothese, dass alle Mittelwerte gleich sind, wird deshalb bestätigt. Aufgrund der vorliegenden Stichprobe kann also angenommen werden, dass die unterschiedlichen Fertigungseinrichtungen keinen Einfluss auf den Kapazitätswert haben. Der Kapazitätswert schwankt nur zufällig zwischen den verschiedenen Fertigungseinrichtungen, der Unterschied der Fertigungseinrichtungen ist also nicht signifikant.

Die Signifikanz des Einflusses der verschiedenen Fertigungseinrichtungen kann auch durch die Berechnung des p-Wertes geprüft werden.

$$p = 1 - F\left(\frac{s_\alpha^2}{s_\varepsilon^2}\right) = 57.92\% \quad (9.24)$$

Da die Wahrscheinlichkeit p mit 57.92% über dem gewählten Signifikanzniveau von $\alpha = 5\%$ liegt, kann die Nullhypothese nicht verworfen werden. Dies stimmt mit der Einschätzung aus dem Vergleich der Größe v_0 mit der berechneten Grenze c überein.

Für das Beispiel ergibt sich mit den berechneten Daten die in Tabelle 9.5 dargestellte ANOVA-Tabelle. Darin sind nochmals die wesentlichen Ergebnisse zusammengefasst, die für die Bewertung der Signifikanz des Einflusses der Fertigungseinrichtung benötigt werden.

Tabelle 9.5: Bewertung der Fertigungseinrichtungen als ANOVA-Tabelle

Streuungsquelle	Quadratsumme	Freiheitsgrade	Standardisierte Quadratsumme	Wert der Testvariable	p-Wert
Zwischen den Gruppen	6.665	2	3.3325	0.58	0.5792
Innerhalb der Gruppen	51.6562	9	5.73958		
Gesamtstreuung	58.3212	11			

Das Ergebnis lässt sich auch mit einem Box-Plot grafisch plausibilisieren. Hierzu wird für jede Gruppe ein Box-Plot erzeugt und zum Vergleich nebeneinander dargestellt.

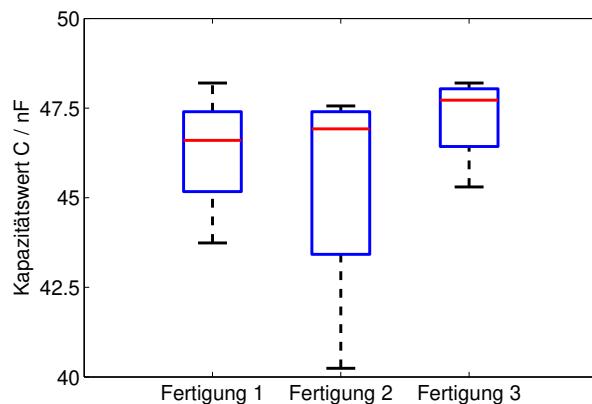


Bild 9.3: Box-Plot für das Beispiel der Kondensatorfertigung

Der Box-Plot in Bild 9.3 zeigt, dass sich die Verteilungen überlappen und deshalb hinsichtlich des Kapazitätswertes kein signifikanter Unterschied zwischen den Gruppen besteht.

Die Berechnung mit MATLAB ist in den folgenden Programmzeilen dargestellt.

```

1 % Messwerte einlesen
2 load Kondensatorfertigung.mat;
3
4 % Berechnung der ANOVA-Tabelle
5 [p, AnovaTab, TestStatistik] = anova1(data, 'on')
6
7 % Erzeugen des Box-Plot
8 boxplot([data(:,1) data(:,2) data(:,3)])

```

9.2 Mehrfaktorielle Varianzanalyse

Bei der einfaktoriellen Varianzanalyse wird gezeigt, wie der Einfluss eines Parameters auf eine Zielgröße geprüft wird. Im Beispiel in Abschnitt 9.1 war dies der Einfluss der verschiedenen Fertigungseinrichtungen auf den Wert des produzierten Kondensators. In vielen Fällen ist die Überprüfung eines Parameters nicht ausreichend, da eine Zielgröße oft von mehr als einer Einflussgröße abhängt. Zum Beispiel könnte der Kapazitätswert von dem Zulieferer des Basismaterials, der Fertigungseinrichtung und dem Bediener der Fertigungseinrichtung abhängen. Um die Signifikanz dieser unterschiedlichen Einflussgrößen zu untersuchen, kann eine mehrfaktorielle Varianzanalyse durchgeführt werden. Die Herleitung dazu wird wegen der vielen Indizes schnell unübersichtlich. Aus diesem Grund beschränkt sich die Herleitung in diesem Abschnitt auf die zweifaktorielle Varianzanalyse. Die dabei erlangten Erkenntnisse können auf beliebig viele Dimensionen erweitert werden.

9.2.1 Datenstruktur und Modellansatz der zweifaktoriellen Varianzanalyse

Bei der zweifaktoriellen Varianzanalyse wird die Bewertung gegenüber der einfaktoriellen Analyse um einen Einflussparameter β und eine Wechselwirkung $\alpha\beta$ erweitert. Dazu muss zunächst die Nomenklatur der Stichprobenbezeichnung geklärt werden. Die Stichprobenwerte können in Matrizenform dargestellt werden.

Tabelle 9.6: Nomenklatur der Stichprobenindizes für die zweifaktorielle Varianzanalyse

		Einflussgröße β				
		1	...	k	...	K
Einflussgröße α	1	$x_{111} \dots x_{11n} \dots x_{11N}$		$x_{1K1} \dots x_{JKn} \dots x_{JKN}$
	⋮					
	j	⋮		$x_{jk1} \dots x_{jkn} \dots x_{jkN}$		⋮
	⋮					
	J	$x_{J11} \dots x_{J1n} \dots x_{J1N}$...		$x_{JK1} \dots x_{JKn} \dots x_{JKN}$

Der Zeilenindex j läuft von 1 bis J, der Spaltenindex k von 1 bis K. In jeder Gruppe befinden sich N Stichprobenwerte, die mit dem Index n bezeichnet werden. Damit ergibt sich die Bezeichnung für den einzelnen Stichprobenwert x_{jkn} .

Der einzelne Stichprobenwert x_{jkn} kann als Summe eines konstanten Mittelwertes μ , den systematischen Abweichungen α_j und β_k , der Wechselwirkung der beiden Größen $\alpha\beta_{jk}$ sowie einer zufälligen, normalverteilten Abweichung ϵ_{jkn} dargestellt werden.

$$x_{jkn} = \mu + \alpha_j + \beta_k + \alpha\beta_{jk} + \epsilon_{jkn} \quad (9.25)$$

Die Analyse der Frage, ob bezüglich eines Faktors oder der Wechselwirkung ein signifikanter Einfluss vorliegt, entspricht wieder einem Hypothesentest. Zum Beispiel kann die Signifikanz des Faktors α_j untersucht werden. Besitzt der Faktor keine Signifikanz, ist für α_j ein Wert nahe null zu erwarten. Damit lautet die Nullhypothese:

$$\alpha_1 = \alpha_2 = \alpha_3 = \dots = \alpha_J = 0 \quad (9.26)$$

Die Gegenhypothese lautet, dass zumindest eine Stichprobe einen signifikanten Einfluss α_j besitzt, und es gilt:

$$\alpha_j \neq 0 \quad (9.27)$$

Diese Form des Hypothesentests kann auf gleiche Weise auch für den Einflussfaktor β und die Wechselwirkung $\alpha\beta$ angewendet werden. Wie bei der einfaktoriellen Varianzanalyse kann direkt aus dem

Modellansatz aus Gleichung (9.25) der Modellansatz der Varianzen aus den Stichprobenwerten angegeben werden.

$$s_x^2 = s_\alpha^2 + s_\beta^2 + s_{\alpha\beta}^2 + s_\varepsilon^2 \quad (9.28)$$

Die Bestimmung der Varianzen erfolgt analog zur einfaktoriellen Varianzanalyse über die Berechnung von Quadratsummen, die auf ihre Anzahl von Freiheitsgraden normiert werden.

9.2.2 Berechnung der einzelnen Varianzen über die standardisierten Quadratsummen

Wie bei der einfaktoriellen Varianzanalyse bildet auch im mehrdimensionalen Fall die Quadratsumme der Abweichungen der Stichprobenwerte vom Mittelwert die Grundlage der Berechnung. Die Quadratsumme

$$q_x = \sum_{j=1}^J \sum_{k=1}^K \sum_{n=1}^N (x_{jkn} - \bar{x})^2 \quad (9.29)$$

kann analog zur einfaktoriellen Varianzanalyse zerlegt werden in

$$q_x = \sum_{j=1}^J \sum_{k=1}^K \sum_{n=1}^N (\bar{x}_j - \bar{x})^2 + \sum_{j=1}^J \sum_{k=1}^K \sum_{n=1}^N (\bar{x}_k - \bar{x})^2 \quad (9.30)$$

mit dem Mittelwert der j-ten Zeile

$$\bar{x}_j = \frac{1}{K \cdot N} \cdot \sum_{k=1}^K \sum_{n=1}^N x_{jkn} \quad (9.31)$$

und dem Mittelwert der k-ten Spalte

$$\bar{x}_k = \frac{1}{J \cdot N} \cdot \sum_{j=1}^J \sum_{n=1}^N x_{jkn} \quad (9.32)$$

Der Mittelwert der Stichprobe in der j-ten Zeile und der k-ten Spalte wird durch den Ausdruck

$$\bar{x}_{jk} = \frac{1}{N} \cdot \sum_{n=1}^N x_{jkn} \quad (9.33)$$

repräsentiert. Der Gesamtmittelwert ergibt sich aus

$$\bar{x} = \frac{1}{J \cdot K \cdot N} \cdot \sum_{j=1}^J \sum_{k=1}^K \sum_{n=1}^N x_{jkn} \quad (9.34)$$

Mit den Größen aus Gleichung (9.31) bis Gleichung (9.34) kann die Quadratsumme q_x aus Gleichung (9.29) in die aus dem Modellansatz in Gleichung (9.28) bekannten vier Streuungseinflüsse zerlegt werden.

$$q_x = q_\alpha + q_\beta + q_\gamma + q_\varepsilon \quad (9.35)$$

Daraus ergibt sich die Varianz der Gesamtstreuung zu

$$s_x^2 = \frac{q_x}{v_x} = \frac{q_\alpha + q_\beta + q_\gamma + q_\varepsilon}{J \cdot K \cdot N - 1} \quad (9.36)$$

Die Gesamtvarianz lässt sich in die Varianz des Einflussfaktors α

$$s_\alpha^2 = \frac{q_\alpha}{v_\alpha} = \frac{K \cdot N \cdot \sum_{j=1}^J (\bar{x}_j - \bar{x})^2}{J - 1} \quad (9.37)$$

und die Varianz des Einflussfaktors β

$$s_{\beta}^2 = \frac{q_{\beta}}{\nu_{\beta}} = \frac{J \cdot N \cdot \sum_{k=1}^K (\bar{x}_k - \bar{x})^2}{K - 1} \quad (9.38)$$

aufteilen. Zusätzlich bewirkt die Wechselwirkung der beiden Einflussfaktoren einen Beitrag zu der Gesamtvarianz. Dieser berechnet sich zu

$$s_{\alpha\beta}^2 = \frac{q_{\alpha\beta}}{\nu_{\alpha\beta}} = \frac{N \cdot \sum_{j=1}^J \sum_{k=1}^K (\bar{x}_{jk} - \bar{x}_j - \bar{x}_k + \bar{x})^2}{(J - 1) \cdot (K - 1)} \quad (9.39)$$

Die Varianz der Reststreuung, die weder den beiden Einflussparametern α und β noch deren Wechselwirkung $\alpha\beta$ zugeordnet werden kann, ergibt sich aus

$$s_{\varepsilon}^2 = \frac{q_{\varepsilon}}{\nu_{\varepsilon}} = \frac{\sum_{j=1}^J \sum_{k=1}^K \sum_{n=1}^N (x_{jkn} - \bar{x})^2}{J \cdot K \cdot (N - 1)} \quad (9.40)$$

Durch die Gleichung (9.36) bis Gleichung (9.40) sind somit alle Varianzen des Modellansatzes bestimmt.

9.2.3 Signifikanzbewertung der einzelnen Einflüsse

Wie bei der einfaktoriellen Varianzanalyse wird nach der Standardisierung der Zufallsvariablen das Verhältnis von der zu untersuchender Varianz und der zufälligen Varianz der Reststreuung gebildet. Zur Bewertung des Einflusses der ersten Größe α ergibt sich der Quotient

$$\nu_{0\alpha} = \frac{s_{\alpha}^2}{s_{\varepsilon}^2} = \frac{\frac{q_{\alpha}}{J - 1}}{\frac{q_{\varepsilon}}{J \cdot K \cdot (N - 1)}} \quad (9.41)$$

für die zweite Einflussgröße β

$$\nu_{0\beta} = \frac{s_{\beta}^2}{s_{\varepsilon}^2} = \frac{\frac{q_{\beta}}{K - 1}}{\frac{q_{\varepsilon}}{J \cdot K \cdot (N - 1)}} \quad (9.42)$$

und für die Wechselwirkung $\alpha\beta$

$$\nu_{0\alpha\beta} = \frac{s_{\alpha\beta}^2}{s_{\varepsilon}^2} = \frac{\frac{q_{\alpha\beta}}{(J - 1) \cdot (K - 1)}}{\frac{q_{\varepsilon}}{J \cdot K \cdot (N - 1)}} \quad (9.43)$$

Die Quotienten aus Gleichung (9.41) bis Gleichung (9.43) weisen einer F-Verteilung mit den Freiheitsgraden auf, die sich aus den Freiheitsgraden von Zähler und Nenner ergeben. Ob ein einzelner Parameter einen signifikanten Einfluss auf den Zielwert besitzt, wird durch den p-Wert des F-Tests aus 6.6.4 signalisiert, der die Wahrscheinlichkeit dafür angibt, dass der wahre Wert v_0 über der Variable ν_0 liegt. Der p-Wert wird wiederum mit dem Signifikanzniveau α verglichen. Liegt dieser oberhalb des Signifikanzniveaus, kann die Nullhypothese nicht verworfen werden, der Einfluss ist somit nicht signifikant.

Auch für die zweifaktorielle Varianzanalyse kann das Ergebnis als ANOVA-Tabelle zusammengefasst werden. Dabei gelten grundsätzlich die gleichen Bezeichnungen wie bei der eindimensionalen Varianzanalyse.

Tabelle 9.7: Zusammenfassung der zweifaktoriellen Varianzanalyse als ANOVA-Tabelle

Streuungsquelle	Quadratsumme	Freiheitsgrade	Standardisierte Quadratsumme	Wert der Testvariable	p-Wert
Zwischen den Gruppen α	q_α	$J - 1$	$s_\alpha^2 = \frac{q_\alpha}{J - 1}$	$v_{0\alpha} = \frac{s_\alpha^2}{s_\epsilon^2}$	$P(v_\alpha > v_{0\alpha})$
Innerhalb den Gruppen β	q_β	$K - 1$	$s_\beta^2 = \frac{q_\beta}{K - 1}$	$v_{0\beta} = \frac{s_\beta^2}{s_\epsilon^2}$	$P(v_\beta > v_{0\beta})$
Wechselwirkung zwischen α und β	$q_{\alpha\beta}$	$(J - 1) \cdot (K - 1)$	$s_{\alpha\beta}^2 = \frac{q_{\alpha\beta}}{(J - 1) \cdot (K - 1)}$	$v_{0\alpha\beta} = \frac{s_{\alpha\beta}^2}{s_\epsilon^2}$	$P(v_{\alpha\beta} > v_{0\alpha\beta})$
Restvariation innerhalb der Gruppe	q_ϵ	$J \cdot K \cdot (N - 1)$	$s_\epsilon^2 = \frac{q_\epsilon}{J \cdot K \cdot (N - 1)}$		
Gesamtvarianz	q_x	$J \cdot K \cdot N - 1$			

Beispiel: Abgleicheinrichtung für Spannungsregler

Für Generatoren werden Spannungsregler gefertigt und abgeglichen. Der Abgleich der Spannungsregler erfolgt in unterschiedlichen Abgleichvorrichtungen. Im Rahmen einer Qualitätskontrolle werden die Abgleichdaten von drei Fertigungsschichten und drei Abgleicheinrichtungen kontrolliert. Es ergeben sich die in Tabelle 9.8 aufgelisteten Messwerte der Spannung.

Tabelle 9.8: Vermessung von Spannungsreglern aus unterschiedlichen Abgleichstationen und unterschiedlichen Fertigungsschichten

Fertigungsschicht	Fertigungseinrichtung		
	A	B	C
1	16.1736	16.4598	16.4500
	16.0336	16.5174	16.5278
	16.0971	16.4884	16.3452
2	16.1243	16.7064	16.5261
	15.9743	16.5755	16.4987
	16.0653	16.4482	16.4420
3	15.9059	16.7010	16.5136
	15.8825	16.7071	16.2742
	15.8979	16.7317	16.1590

In dem Beispiel soll der Einfluss der Fertigungseinrichtung und der Fertigungsschicht auf den Abgleichwert untersucht werden. Für das Beispiel ergeben sich die Werte der Quadratsumme der Gesamtstreuung zu

$$q_x = 1.90195 \quad (9.44)$$

Sie kann aufgeteilt werden in

$$q_\alpha = 0.01925 \quad (9.45)$$

$$q_\beta = 1.56408 \quad (9.46)$$

$$q_{\alpha\beta} = 0.17569 \quad (9.47)$$

und

$$q_\varepsilon = 0.14294 \quad (9.48)$$

Analog zu der einfaktoriellen Varianzanalyse werden diese Quadratsummen auf ihre Anzahl Freiheitsgrade normiert. Diese normierten Quadratsummen stellen eine Schätzung der Varianz des entsprechenden Einflusses dar und können entsprechend auf ihre Signifikanz hin überprüft werden. Nach der Normierung ergeben sich Werte von

$$s_\alpha^2 = \frac{q_\alpha}{v_\alpha} = \frac{0.01925}{3-1} = 0.00962 \quad (9.49)$$

$$s_\beta^2 = \frac{q_\beta}{v_\beta} = \frac{1.56408}{3-1} = 0.78204 \quad (9.50)$$

$$s_{\alpha\beta}^2 = \frac{q_{\alpha\beta}}{v_{\alpha\beta}} = \frac{0.17569}{(3-1) \cdot (3-1)} = 0.04392 \quad (9.51)$$

und

$$s_\varepsilon^2 = \frac{q_\varepsilon}{v_\varepsilon} = \frac{0.14294}{3 \cdot 3 \cdot (3-1)} = 0.00794 \quad (9.52)$$

Zur Überprüfung der Signifikanz müssen die Prüfgrößen v_0 bestimmt werden. Sie ergeben sich zu

$$v_{0\alpha} = \frac{s_\alpha^2}{s_\varepsilon^2} = 1.21 \quad (9.53)$$

$$v_{0\beta} = \frac{s_\beta^2}{s_\varepsilon^2} = 98.48 \quad (9.54)$$

und

$$v_{0\alpha\beta} = \frac{s_{\alpha\beta}^2}{s_\varepsilon^2} = 5.53 \quad (9.55)$$

Die Verteilung die der Berechnung der Wahrscheinlichkeit für die Signifikanzbewertung des Einflussfaktors α und β zugrunde liegt, ist eine F-Verteilung mit (2, 18) Freiheitsgraden, sodass sich bei einem Signifikanzniveau von 5 % eine Grenze von

$$c_\alpha = c_\beta = 3.5546 \quad (9.56)$$

ergibt. Die Prüfgröße v_0 ist kleiner als der Grenzwert c , damit ist die Fertigungsschicht nicht signifikant für das Abgleichergebnis. Da der Wert v_0 größer ist als die Grenze c , ist die Abweichung signifikant. Die Fertigungseinrichtung hat demnach einen Einfluss auf das Abgleichergebnis.

Die Berechnung der Wahrscheinlichkeit für die Signifikanzbewertung des Einflussfaktors $\alpha\beta$ zugrunde liegende Verteilung ist eine F-Verteilung mit (4, 18) Freiheitsgraden, sodass sich bei einem Signifikanzniveau von 5 % eine Grenze von

$$c_{\alpha\beta} = 2.9277 \quad (9.57)$$

ergibt. Da der v_0 größer ist als die Grenze c , ist die Abweichung signifikant. Die Kombination von Fertigungseinrichtung und Fertigungsschicht hat demnach ebenfalls einen Einfluss auf das Abgleichergebnis.

Die Bewertung der Signifikanz der Einflüsse kann auch mithilfe des p-Wertes erfolgen. Die Berechnung ergibt Werte von

$$p_\alpha = 32.08\% \quad (9.58)$$

$$c_\beta = 0\% \quad (9.59)$$

und

$$p_{\alpha\beta} = 0.44\% \quad (9.60)$$

Der Wert p liegt mit 32.08 % weit über dem Signifikanzniveau von 5 %, die einzelnen Fertigungsschichten haben somit keinen signifikanten Einfluss auf den Spannungswert. Die Aussage, die bereits über den Vergleich der Testvariablen v_0 mit der kritischen Grenze c getroffen wurde, wird damit bestätigt. Analog kann dies auch für die Einflussgröße β und die Wechselwirkung $\alpha\beta$ durchgeführt werden.

Das Ergebnis der Untersuchung ist in Tabelle 9.9 als ANOVA-Tabelle dargestellt.

Tabelle 9.9: ANOVA-Tabelle für die Vermessung von Spannungsreglern aus unterschiedlichen Abgleichstationen und unterschiedlichen Fertigungsschichten

Streuungsquelle	Quadratsumme	Freiheitsgrade	Standardisierte Quadratsumme	Wert der Testvariable	p-Wert
Zwischen den Gruppen α	0.01925	2	0.00962	1.21	0.3208
Innerhalb den Gruppen β	1.56408	2	0.78204	98.48	0
Wechselwirkung zwischen Einrichtung und Schicht	0.17569	4	0.04392	5.53	0.0044
Restvariation innerhalb der Gruppe	0.14294	18	0.00794		
Gesamt-Varianz	1.90195	26			

Die Bewertung der Signifikanz der beiden Einflussgrößen α und β kann grafisch mithilfe des Box-Plots plausibilisiert werden.

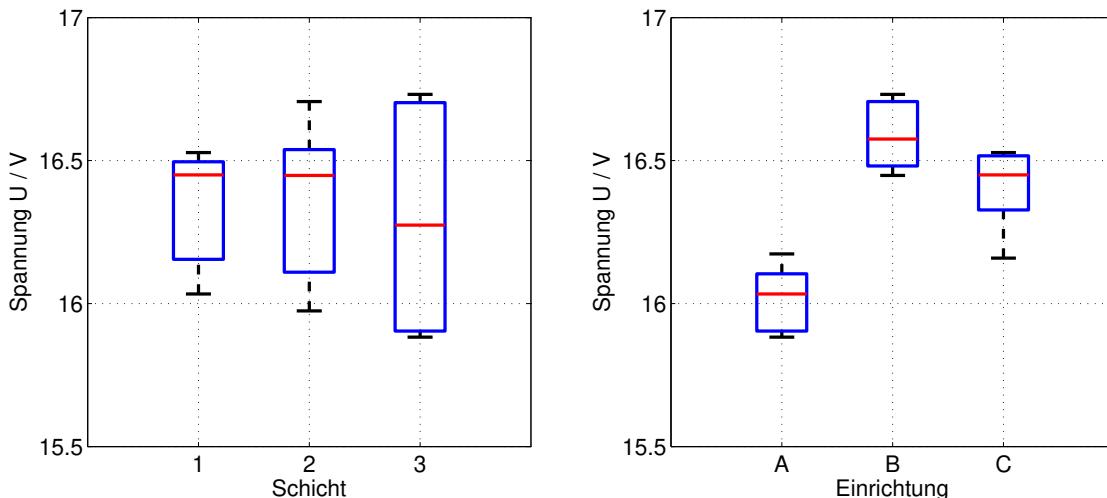


Bild 9.4: Boxplot für die Vermessung von Spannungsreglern aus unterschiedlichen Abgleichstationen und unterschiedlichen Fertigungsschichten

An der Überlappung der Interquartil-Bereiche im linken Diagramm lässt sich ablesen, dass die Schicht keinen signifikanten Einfluss auf das Abgleichergebnis hat. Der Box-Plot im rechten Diagramm bestätigt den signifikanten Einfluss der Einrichtung.

Die Berechnung mit MATLAB ist in den folgenden Programmzeilen dargestellt.

```

1 % Messwerte einlesen
2 load Abgleicheinrichtung.mat;
3 U = [U11 U12 U13 ; U21 U22 U23 ; U31 U32 U33];
4 U1 = [U11 U21 U31 ; U12 U22 U32 ; U13 U23 U33];
5
6 % Berechnung der ANOVA-Tabelle
7 [p, AnovaTab, TestStatistik] = anova2(U,3)
8
9 % Erzeugen der grafischen Plausibilisierung
10 f = figure(2);
11 % Box-Plot der Einflussgrößen \alpha
12 subplot(1,2,1)
13 boxplot([U1(:,1) U1(:,2) U1(:,3)]);
14
15 % Box-Plot der Einflussgrößen \beta
16 subplot(1,2,2)
17 boxplot([U(:,1) U(:,2) U(:,3)]);

```

9.2.4 Varianzanalysen mit mehr als zwei Einflussgrößen

Varianzanalysen mit mehr Einflussgrößen werden auf dieselbe Art durchgeführt, wie die ein- oder zweifaktorielle Varianzanalyse. Wegen des hohen numerischen Aufwands wird die Analyse typischerweise mit der Unterstützung eines Statistik-Programms durchgeführt, die jeweils unterschiedliche Syntax aufweisen. Gemeinsam ist die Darstellung des Ergebnisses als ANOVA-Tabelle, die die Zwischenergebnisse und das Ergebnis des Hypothesentests beinhaltet.

9.2.5 ANOVA-Tabellen in MATLAB und Python

Tabelle 9.10: Varianzanalyse mit MATLAB

MATLAB-Befehl	Funktionsbeschreibung
anova1(X)	Einfaktorielle Varianzanalyse
anova2(X)	Zweifaktorielle Varianzanalyse
anova3(X)	N-Dimensionale Varianzanalyse

In Python existieren unterschiedliche Methoden zur Berechnung einer Varianzanalyse. Eine universelle Lösung bietet die Bibliothek statsmodels.api, mit der auch Regressionsfunktionen berechnet werden können. Die Grundidee dabei ist, ein Modell kategorialer Variablen wie in Gleichung (9.25) zu definieren. Die konstante Größe ist keine Varianzursache und wird deshalb nicht aufgeführt. Der zufällige Fehler wird implizit berücksichtigt und nicht explizit aufgeführt.

$$\text{Modell} = \alpha_j + \beta_k + \alpha\beta_{jk} \quad (9.61)$$

Für dieses Modell wird eine Varianzanalyse durchgeführt. Der folgende Programmausschnitt zeigt das Vorgehen für das Beispiel der Spannungsregler aus Kapitel 9.2.3.

```

1 Bibliotheken importieren
2 import numpy as np
3 import pandas as pd
4 import matplotlib.pyplot as plt
5 import statsmodels.api as sm
6 from statsmodels.formula.api import ols
7
8 Data Frame Variable mit Daten erstellen
9 df = pd.DataFrame( 'Schicht' : np.tile(np.repeat([1, 2, 3], 3), 3),
10                     'Linie' : np.repeat(['A', 'B', 'C'], 9),
11                     'Spannung' : [16.1736, 16.0336, 16.0971,
12                           16.1243, 15.9743, 16.0653,
13                           15.9059, 15.8825, 15.8979,
14                           16.4598, 16.5174, 16.4884,
15                           16.7064, 16.5755, 16.4482,
16                           16.7010, 16.7071, 16.7317,
17                           16.4500, 16.5278, 16.3452,
18                           16.5261, 16.4987, 16.4420,
19                           16.5136, 16.2742, 16.1590])
20 print(df)
21
22 Modell aufbauen und ANOVA durchführen
23 model = ols('Spannung $\sim C(Linie) + C(Schicht) + C(Linie):C(
24 Schicht)', data=df).fit()
25 anova2 = sm.stats.anova_lm(model, typ=2)
26 print(anova2)
27
28 Boxplot erstellen
29 fig = plt.figure(2, figsize=(12, 4))
30 ax1, ax2 = fig.subplots(1, 2)
31 ax1 = df.boxplot('Spannung', by='Linie', ax=ax1)
32 ax2 = df.boxplot('Spannung', by='Schicht', ax=ax2)
33 plt.suptitle('')
```

Dabei sind die kategorialen Variablen zum Beispiel als $C(Linie)$ gekennzeichnet, das Produkt der kategorialen Größen wird als $C(Linie):C(Schicht)$ implementiert. Der ursprüngliche Datensatz sowie die entstehende ANOVA-Tabelle erwenden das Pandas Dataframe-Format. Das Dataframe-Format erlaubt außerdem eine besonders elegante Darstellung der Boxplots zur Validierung des Ergebnisses.

Tabelle 9.11: Varianzanalyse mit Python

Python-Befehl	Funktionsbeschreibung
statsmodels.formula.api.ols	Definition der Modellgleichung
statsmodels.api.stats.anova_lm	Durchführung der ANOVA

9.3 Anwendungsbeispiel: Homogenitätsprüfung eines Luftflusses

Ein wesentliches Ziel des Umweltschutzes ist es, schädliche Emissionen möglichst abzustellen oder so weit wie möglich zu reduzieren, um die Umwelt vor Luft-, Boden- oder Gewässerverschmutzung zu bewahren und Menschen vor Belastungen zu schützen. In dem Bundes-Immissionsschutzgesetz werden daher Grenzwerte für den Ausstoß von Schadstoffen aus großen Feuerungsanlagen wie Kohlewerken definiert.

Die Einhaltung dieser Grenzwerte muss in regelmäßigen Abständen durch zertifizierte Überwachungsstellen überprüft werden. Um den Aufwand zu minimieren, ist es vorteilhaft, nur an einer Stelle im Abluftkanal messen zu müssen. Voraussetzung dafür ist, dass die Emissionsverteilung über den Querschnitt des Abluftkanals ausreichend homogen ist. Der Homogenitätstest entspricht einer einfaktoriellen Varianzanalyse. Das Abgas wird als homogen über den Querschnitt angesehen, wenn sich der Messwert zwar zeitlich ändert, jedoch nicht von Messpunkt zu Messpunkt. Es wird deshalb überprüft, ob die Abweichungen der Messwerte zufällig sind oder ob durch die Inhomogenität des Luftflusses der Gehalt der Messgröße in der Abluft variiert.

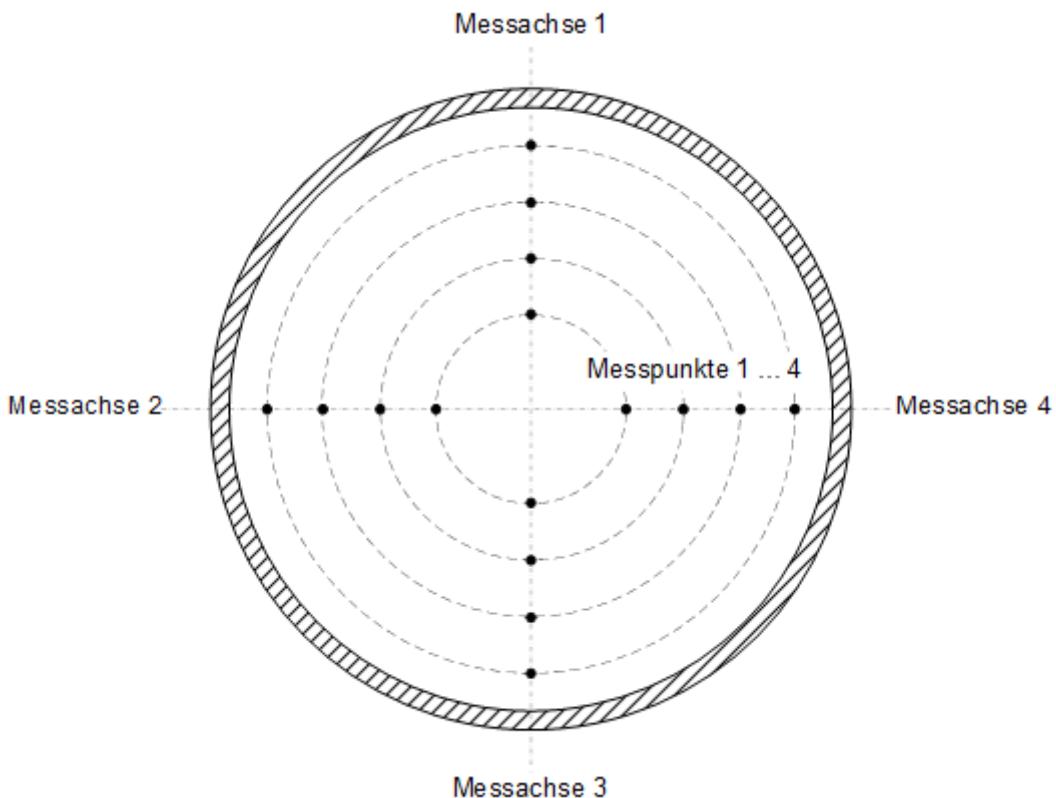


Bild 9.5: Querschnitt durch den Abluftkanal mit unterschiedlichen Messachsen und Messpunkten

Liegt für die Anlage keine gültige Homogenitätsprüfung vor, ist die Homogenität der Verteilung der Messgröße beziehungsweise eines Ersatzparameters im Messquerschnitt mithilfe von Netzmessungen und zusätzlicher Vergleichsmessungen mit einer unabhängigen Messeinrichtung an einem festen Punkt innerhalb der Messstrecke zu ermitteln. Dabei werden an mehreren Stellen des Abluftkanals Messsonden eingebracht und bei unterschiedlicher Eindringtiefe der Messwert erfasst. Die Homogenitätsbestimmung wird in diesem Beispiel für Stickoxide (NO_x) durchgeführt. Hierzu wurden die in Tabelle 9.12 aufgelisteten Messwerte aufgenommen. Parallel wird ein Referenzwert aufgenommen.

Tabelle 9.12: Messreihe zur Homogenitätsbestimmung für Stickoxide

Messachse	Messort		Netzmessung	Referenzmessung
	Messpunkt			
1	1		127	125
1	2		132	129
1	3		132	131
1	4		109	127
2	1		136	126
2	2		148	121
2	3		160	118
2	4		152	126
3	1		113	107
3	2		132	96
3	3		125	101
3	4		125	100
4	1		119	98
4	2		127	105
4	3		127	105
4	4		134	112

Netzmessung und Referenzmessung bilden eine Gruppe mit einem Stichprobenumfang von zwei Messungen. Die Differenz der Messwerte oder die Varianz innerhalb der Gruppe ist damit ein Maß für die Genauigkeit der Messung selbst.

Die Messungen werden an 16 Orten durchgeführt, sie bilden die Gruppen der Varianzanalyse. Die Varianz zwischen den Gruppen ist ein Maß für die Homogenität des Abluftstroms.

Bei einem homogenen Abluftstrom müsste die Varianz von Messort zu Messort kleiner sein als die Varianz zwischen Netzmessung und Referenzmessung. Diese Annahme kann mit einer Varianzanalyse mit $J = 16$ Stichproben mit je einem Umfang von $N = 2$ Messwerten überprüft werden. Die Auswertung ergibt die in Tabelle 9.13 dargestellte ANOVA-Tabelle.

Tabelle 9.13: Bewertung der Homogenität als ANOVA-Tabelle

Streuungsquelle	Quadratsumme	Freiheitsgrade	Standardisierte Quadratsumme	Wert der Testvariable	p-Wert
Zwischen den Gruppen (Homogenität)	3274.72	15	218.315	0.87	0.6043
Innerhalb der Gruppen (Genauigkeit)	4016.5	16	251.031		
Gesamtstreuung	7291.22	31			

Mit dem Signifikanzniveau $\alpha = 0.05$, den Freiheitsgraden $(J - 1) = 15$ beziehungsweise $J \cdot (N - 1) = 16$ und der inversen F-Verteilung ergibt sich für

$$F(c) = 1 - \alpha = 0.95 \quad (9.62)$$

die kritische Grenze $c = 2.3522$. Der Vergleich mit v_0 zeigt, dass $v_0 < c$ ist. Die Hypothese, dass alle Mittelwerte gleich sind, wird deshalb bestätigt. Aufgrund der vorliegenden Stichprobe kann also angenommen werden, dass der Luftfluss homogen ist. Die Messwerte schwanken nur zufällig um den tatsächlichen NOx-Gehalt.

Die Signifikanz des Messortes kann auch durch die Berechnung des p-Wertes geprüft werden.

$$p = 1 - F\left(\frac{s_\alpha^2}{s_\epsilon^2}\right) = 60.43\% \quad (9.63)$$

Da die Wahrscheinlichkeit p mit 60.43% über dem gewählten Signifikanzniveau von $\alpha = 5\%$ liegt, kann die Nullhypothese nicht verworfen werden. Dies stimmt mit der Einschätzung aus dem Vergleich der Größe v_0 mit der berechneten Grenze c überein. Durch die Auswertung der Messwerte ist die Homogenität des Luftflusses nachgewiesen. Der Messpunkt kann somit frei gewählt werden.

Die Auswertung der Messreihe wurde mit MATLAB durch die folgenden Programmzeilen durchgeführt.

```

1 % Messwerte einlesen
2 load Homogenitaetsbestimmung.mat;
3
4 % Berechnung der ANOVA-Tabelle
5 [p, AnovaTab, TestStatistik] = anova1([c\_Netz;c\_Punkt], 'on')
```

9.4 Literatur

- [Krey91] Kreyszig, Erwin: Statistische Methoden und ihre Anwendungen
4., unveränderter Nachdruck der 7. Auflage
Vandenhoeck & Ruprecht, Göttingen, 1991
- [Fahr96] Fahrmeir, Ludwig; Hamerle, Alfred; Tutz, Gerhard: Multivariate statistische Verfahren
2., überarbeitete Auflage
Walter de Gruyter & Co., Berlin
- [Ross06] Ross, M. Sheldon: Statistik für Ingenieure und Naturwissenschaftler
3. Auflage
Spektrum Akademischer Verlag, München, 2006
- [Hart07] Hartung, Joachim; Elpelt, Bärbel: Multivariate Statistik
7., unveränderte Auflage
R. Oldenbourg Verlag, München / Wien
- [Papu01] Papula, Lothar: Mathematik für Ingenieure und Naturwissenschaftler Band 3
4., verbesserte Auflage
Vieweg Teubner, Braunschweig / Wiesbaden, 2008

10 Korrelationsanalyse

In den Kapiteln 7 und 8 werden zwei- und mehrdimensionale Datensätze und Zufallsvariablen vorgestellt und beschrieben. Dabei wird die Kovarianz als Maß für den Zusammenhang zweier Zufallsgrößen diskutiert. Aufgrund einer fehlenden Normierung eignet sie sich jedoch nur bedingt zur Interpretation der Abhängigkeit. Eine geeignete Normierung liefert der Korrelationskoeffizient.

Ist der Korrelationskoeffizient ρ der Grundgesamtheit unbekannt, kann er auf Basis einer Stichprobe geschätzt werden. Die Bewertung dieser Schätzung erfolgt über einen Konfidenzbereich oder mithilfe eines Hypothesentests. Beide Verfahren werden in diesem Kapitel vorgestellt.

10.1 Korrelationskoeffizient einer Stichprobe

Der Korrelationskoeffizient r einer Stichprobe ist ein Maß dafür, wie ähnlich sich die zu untersuchenden Datensätze oder Zufallsvariablen sind. Er beschreibt den Grad der linearen Abhängigkeit. Die Daten oder Zufallsvariablen können dabei kontinuierlich oder diskret sein.

Zunächst wird der Korrelationskoeffizient zweidimensionaler Stichproben definiert. Diese Definition wird anschließend auf M-dimensionale Zufallsvektoren erweitert.

10.1.1 Korrelationskoeffizient r einer zweidimensionalen Stichprobe

Um die Skalierungsabhängigkeit der Kovarianz zu eliminieren, kann sie normiert werden. Eine geeignete Normierung bildet der Korrelationskoeffizient r .

$$r = \frac{s_{xy}}{s_x \cdot s_y} = \frac{\frac{1}{N-1} \cdot \sum_{n=1}^N ((x_n - \bar{x}) \cdot (y_n - \bar{y}))}{\sqrt{\frac{1}{N-1} \cdot \sum_{n=1}^N (x_n - \bar{x})^2} \cdot \sqrt{\frac{1}{N-1} \cdot \sum_{n=1}^N (y_n - \bar{y})^2}} \quad (10.1)$$

Durch die Normierung des Ausdrucks ist der Korrelationskoeffizient dimensionslos. Da die beiden Ausdrücke für die Standardabweichungen s_x und s_y im Nenner immer positiv sind, ist das Vorzeichen des Korrelationskoeffizienten dasselbe wie das der Kovarianz. Da die Reihenfolge der Faktoren in Zähler und Nenner beliebig ist, ist der Korrelationskoeffizient r unabhängig von der Reihenfolge der Stichprobengrößen.

$$r = \frac{s_{xy}}{s_x \cdot s_y} = \frac{s_{yx}}{s_y \cdot s_x} \quad (10.2)$$

Der Wert des Korrelationskoeffizienten r ist ein Maß dafür, wie stark die Zufallsvariablen linear voneinander abhängig sind. Bild 10.1 stellt Stichproben mit unterschiedlichen Werten für den Korrelationskoeffizienten r dar.

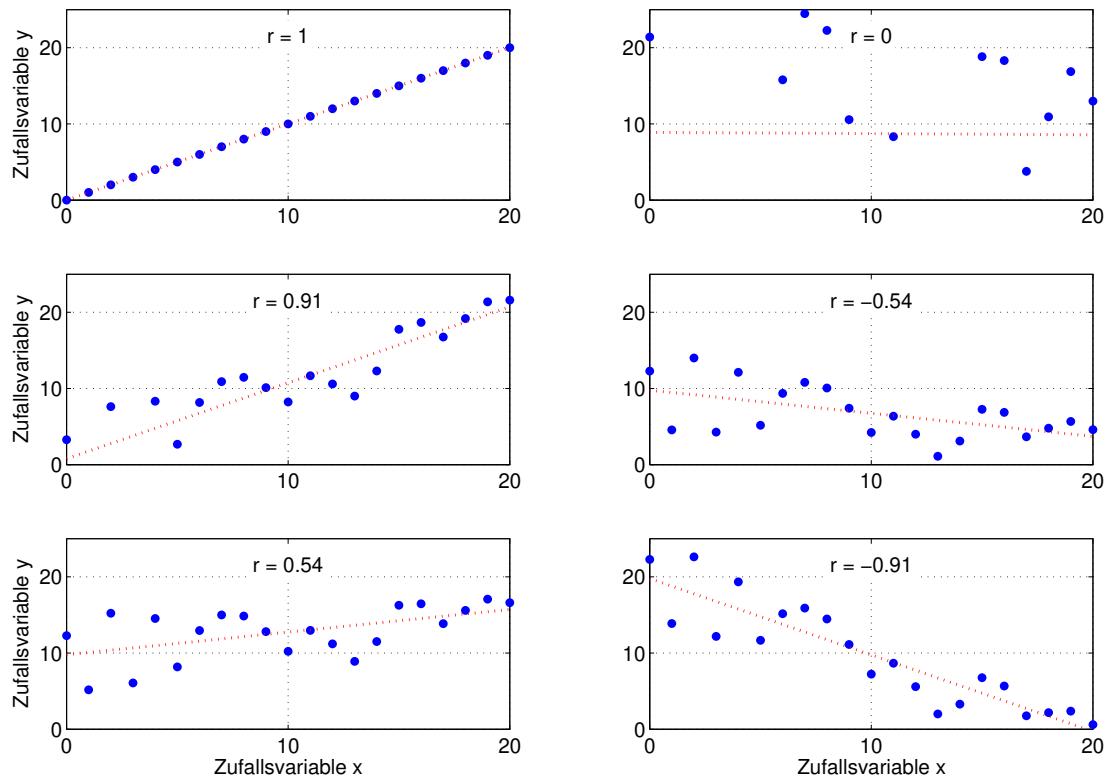


Bild 10.1: Stichproben mit unterschiedlichen Werten für den Korrelationskoeffizienten r und der entsprechenden linearen Approximation

In dem ersten Schaubild ist $r = 1$ und die Wertepaare liegen auf einer Geraden mit positiver Steigung. Im zweiten Diagramm ist der Korrelationskoeffizient $r = 0$, es existiert kein signifikanter Zusammenhang zwischen den Werten x_n und y_n der Stichprobe.

Die übrigen Diagramme zeigen den Zusammenhang zwischen Streudiagramm und Korrelationskoeffizient auf. Ein Betrag des Korrelationskoeffizienten r nahe 1 weist auf einen nahezu linearen Zusammenhang der beiden Größen x und y hin. Je linearer der Zusammenhang ist, desto größer ist der Betrag des Korrelationskoeffizienten r . Bei positivem Vorzeichen des Korrelationskoeffizienten steigen die Werte für y mit steigenden Werten für x an. Bei einem negativen Korrelationskoeffizienten fallen die Werte für y mit steigenden Werten für x . Tabelle 10.1 stellt die Eigenschaften des Korrelationskoeffizienten zusammen.

Tabelle 10.1: Eigenschaften des Korrelationskoeffizienten

Korrelationskoeffizient	Interpretation
$r = 0$	unkorrelierte Größen
$r > 0$	positive Korrelation gleichsinniger linearer Zusammenhang
$r < 0$	negative Korrelation gegensinniger linearer Zusammenhang

Nimmt der Korrelationskoeffizient r den Wert 1 oder -1 an, sind die Zufallsgrößen linear voneinander abhängig und damit stark korreliert. Für den Wert 0 des Korrelationskoeffizienten liegt keine lineare Abhängigkeit vor. Um die Korrelation auch zwischen diesen Eckpunkten einstufen zu können, wird

die Korrelation je nach Betrag des Korrelationskoeffizienten r in eine schwache, mittlere oder starke Korrelation eingeteilt. Die einzelnen Intervalle sind in Tabelle 10.2 aufgelistet.

Tabelle 10.2: Einstufungen des Korrelationskoeffizienten

Korrelationskoeffizient	Interpretation
$ r \leq 0.5$	schwache Korrelation
$ r < 0.5 \leq 0.8$	mittlere Korrelation
$0.8 \leq r $	starke Korrelation

Es sei noch einmal darauf hingewiesen, dass der Korrelationskoeffizient kein Maß für die Abhängigkeit schlechthin ist, sondern ein Maß für einen linearen Zusammenhang der beiden Zufallsvariablen.

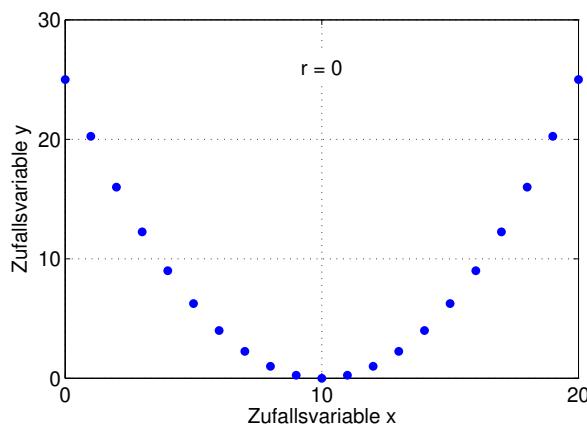


Bild 10.2: Stichprobe unkorrelierter Werte mit einem Korrelationskoeffizienten $r = 0$

Obwohl die beiden Zufallsvariablen einen quadratischen Zusammenhang aufweisen, ist ihre Korrelation $r = 0$. Es existiert kein linearer Zusammenhang zwischen den beiden Größen.

Beispiel: Untersuchung einer synthetischen Faser

Die gewonnenen Kenntnisse sollen an dem Beispiel einer zweidimensionalen Stichprobe verdeutlicht werden. Hierzu wird eine synthetische Faser mit einem festen Baumwollanteil untersucht. Es soll mittels Korrelation herausgefunden werden, ob zwischen der Zugfestigkeit dieser Faser und deren Trocknungszeit ein Zusammenhang besteht. Hierzu wurden 10 Faserproben analysiert, die in Tabelle 10.3 aufgelistet sind.

Tabelle 10.3: Daten zur Untersuchung einer synthetischen Faser

Nr.	Trocknungszeit t / h	Normierte Zugfestigkeit R	Nr.	Trocknungszeit t / h	Normierte Zugfestigkeit R
1	22.31	71.2467	6	25.48	72.9567
2	4.48	73.7833	7	34.79	79.7900
3	22.00	72.2867	8	41.71	81.3300
4	26.29	75.4200	9	41.20	77.7067
5	32.91	78.7000	10	43.34	80.3100

Um zunächst einen Überblick über die Daten zu erhalten, werden die in Tabelle 10.3 aufgelisteten Messwerte in einem Streudiagramm dargestellt.

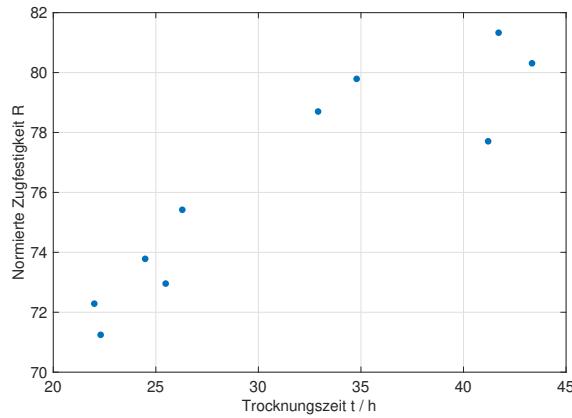


Bild 10.3: Streudiagramm der Messwerte zur Untersuchung einer synthetischen Faser

Die Kovarianz der Stichprobe berechnet sich zu

$$s_{tR} = \frac{1}{N-1} \cdot \sum_{n=1}^N ((t_n - \bar{t}) \cdot (R_n - \bar{R})) = 28.2574 \quad (10.3)$$

Durch die Normierung mit

$$s_t = \sqrt{\frac{1}{N-1} \cdot \sum_{n=1}^N (t_n - \bar{t})^2} = 8.4240 \quad (10.4)$$

und

$$s_R = \sqrt{\frac{1}{N-1} \cdot \sum_{n=1}^N (R_n - \bar{R})^2} = 3.6697 \quad (10.5)$$

ergibt sich der Korrelationskoeffizient der Stichprobe zu

$$r = \frac{s_{tR}}{s_t \cdot s_R} = \frac{28.2574}{8.4240 \cdot 3.6697} = 0.9141 \quad (10.6)$$

Mit MATLAB errechnet sich dieser Wert durch

```

1 % Messwerte einlesen
2 load ZugfestigkeitFasern.mat;
3
4 % Berechnung des Korrelationskoeffizienten
5 r = corr(x,y)

```

Entsprechend der Einstufung nach Tabelle 10.2 sind die Zugfestigkeit und die Trocknungszeit stark korreliert. Dieses Ergebnis entspricht der grafischen Bewertung von Bild 10.3, in dem zu erkennen ist, dass die Zugfestigkeit mit steigenden die Trocknungszeit zunimmt.

In Python wird für multivariate Aufgaben typischerweise die Pandas-Datenstruktur Dataframe verwendet. Für diesen Datentyp existiert die Methode corr, die die Korrelation zwischen Datenvektoren des Dataframes berechnet. Es entsteht eine Korrelationsmatrix, aus der das entsprechende Element ausgewählt wird.

```

1 Bibliotheken importieren
2 import pandas as pd
3 from scipy.io import loadmat
4
5 Laden der Daten
6 Formatierung als Dataframe und Berechnung der Korrelation
7 data = loadmat('ZugfestigkeitFasern')
8 df = pd.DataFrame('Trocknungszeit': data['values'][:, 1],
9 'Zugfestigkeit': data['values'][:, 2])
10 Corr = round((df.corr(method='pearson')).loc['Trocknungszeit', 'Zugfestigkeit'], 3)

```

10.1.2 Korrelationsmatrix R einer multivariaten Stichprobe

Die Darstellung multivariater Stichproben erfolgt mithilfe von Matrizen. Durch die Matrizen-Schreibweise ist es möglich, die multivariate Bewertung auf die Bewertung von jeweils zwei Zufallsvariablen zurückzuführen. Diese Vorgehensweise wird bei der Einführung der Kovarianzmatrix \mathbf{S} in Abschnitt 8.2.2 beschrieben. Ein vergleichbares Vorgehen führt zu einer Korrelationsmatrix \mathbf{R} , die im Folgenden diskutiert wird. Für jedes Paar x_j und x_k von Zufallsvariablen kann nach Gleichung (10.2) der zugehörige Korrelationskoeffizient r_{jk} angegeben werden zu

$$r_{jk} = \frac{s_{x_j x_k}}{s_{x_j} \cdot s_{x_k}} \quad (10.7)$$

Um die Wechselwirkung der Größen miteinander paarweise zu beschreiben, wird eine Matrix von Korrelationskoeffizienten aufgestellt. Bei einer Anzahl von M Zufallsgrößen ergibt sich eine Korrelationsmatrix \mathbf{R} der Dimension $M \times M$.

$$R = \begin{pmatrix} r_{11} & r_{12} & \dots & r_{1M} \\ r_{21} & r_{22} & \dots & r_{2M} \\ \dots & \dots & \dots & \dots \\ r_{M1} & r_{M2} & \dots & r_{MM} \end{pmatrix} = \begin{pmatrix} 1 & r_{12} & \dots & r_{1M} \\ r_{2,1} & 1 & \dots & r_{2M} \\ \dots & \dots & \dots & \dots \\ r_{M1} & r_{M2} & \dots & 1 \end{pmatrix} \quad (10.8)$$

Auf der Hauptdiagonale wird die Korrelation der Größen mit sich selbst berechnet. Da der Zusammenhang zwischen jeder Größe mit sich selbst streng linear ist, ist die Korrelation hier immer 1. Wegen der Symmetrie der Korrelation in Gleichung (10.2) ist die Korrelationsmatrix symmetrisch zur Hauptdiagonalen.

10.2 Definition des Korrelationskoeffizienten ρ der Grundgesamtheit

In Kapitel 5 und 6 werden Verfahren vorgestellt, mit denen Vertrauensbereiche und Hypothesentests aufgebaut werden. Grundlage für diese Tests ist die Analyse einer Zufallsvariable, die den Zusammenhang zwischen der Stichprobe und der Grundgesamtheit herstellt. Um für den Korrelationskoeffizienten ρ der Grundgesamtheit einen Konfidenzbereich angeben und einen Hypothesentest zur Signifikanzprüfung durchführen zu können, wird in diesem Abschnitt der Korrelationskoeffizient der Grundgesamtheit eingeführt, und es werden die grundlegenden Eigenschaften des Korrelationskoeffizienten erläutert.

10.2.1 Korrelationskoeffizient ρ der Grundgesamtheit von Wertepaaren

Die in der zweidimensionalen Grundgesamtheit vorkommenden Mittelwerte μ_x und μ_y der beiden Zufallsvariablen x und y sind durch den Erwartungswert-Operator definiert zu

$$\mu_x = E(x) \quad (10.9)$$

und

$$\mu_y = E(y) \quad (10.10)$$

Die zugehörigen Varianzen der Grundgesamtheit errechnen sich aus

$$\sigma_x^2 = E((x - \mu_x)^2) \quad (10.11)$$

und

$$\sigma_y^2 = E((y - \mu_y)^2) \quad (10.12)$$

Mit der Kovarianz σ_{xy} der Zufallsgrößen x und y

$$\sigma_{xy} = E((x - \mu_x) \cdot (y - \mu_y)) = E(x \cdot y) - E(x) \cdot E(y) \quad (10.13)$$

berechnet sich der Korrelationskoeffizient der Grundgesamtheit ρ in Anlehnung an Gleichung (10.2) zu

$$\rho = \frac{\sigma_{xy}}{\sigma_x \cdot \sigma_y} \quad (10.14)$$

10.2.2 Korrelation der Grundgesamtheit einer multivariaten Stichprobe

Analog zu der Korrelationsmatrix \mathbf{R} aus Abschnitt 10.1 kann für jedes Paar x_j und x_k von Zufallsvariablen nach Gleichung (10.7) die Korrelation der Grundgesamtheit angegeben werden zu

$$\rho_{jk} = \frac{\sigma_{x_j x_k}}{\sigma_{x_j} \cdot \sigma_{x_k}} \quad (10.15)$$

Um die Wechselwirkung der Größen miteinander paarweise zu beschreiben, wird eine Matrix von Korrelationskoeffizienten aufgestellt. Bei einer Anzahl von M Zufallsgrößen ergibt sich eine Korrelationsmatrix der Dimension $M \times M$.

$$P = \begin{pmatrix} \rho_{11} & \rho_{12} & \dots & \rho_{1M} \\ \rho_{21} & \rho_{22} & \dots & \rho_{2M} \\ \dots & \dots & \dots & \dots \\ \rho_{M1} & \rho_{M2} & \dots & \rho_{MM} \end{pmatrix} \quad (10.16)$$

Auf der Hauptdiagonale wird die Korrelation der Größen mit sich selbst berechnet. Da der Zusammenhang zwischen jeder Größen mit sich selbst streng linear ist, ist die Korrelation hier immer 1. Wegen der Symmetrie der Korrelation in Gleichung (10.15) ist die Korrelationsmatrix symmetrisch zur Hauptdiagonalen.

10.2.3 Wertebereich des Korrelationskoeffizienten

Der Korrelationskoeffizient ρ liegt zwischen -1 und 1.

$$-1 \leq \rho \leq 1 \quad (10.17)$$

Für einen Beweis dieser Ungleichung werden die beiden standardnormalverteilten Zufallsgrößen

$$z_x = \frac{x - \mu_x}{\sigma_x} \quad (10.18)$$

und

$$z_y = \frac{y - \mu_y}{\sigma_y} \quad (10.19)$$

zu der Zufallsgröße z

$$z = t \cdot z_x + z_y \quad (10.20)$$

verrechnet. Der Erwartungswert der Zufallsvariablen z ergibt sich wegen der Standardnormalverteilung der Zufallsvariablen z_x und z_y zu

$$E(z) = E(t \cdot z_x + z_y) = t \cdot E(z_x) + E(z_y) = 0 \quad (10.21)$$

Für die Varianz der Zufallsvariablen folgt

$$E(z^2) = E((t \cdot z_x + z_y)^2) = t^2 \cdot E(z_x^2) + 2 \cdot t \cdot E(z_x \cdot z_y) + E(z_y^2) = t^2 + 2 \cdot t \cdot E(z_x \cdot z_y) + 1 \quad (10.22)$$

Die Kovarianz der Zufallsgrößen z_x und z_y lässt sich umformen zu

$$E(z_x \cdot z_y) = E\left(\frac{x - \mu_x}{\sigma_x} \cdot \frac{y - \mu_y}{\sigma_y}\right) = \frac{1}{\sigma_x \cdot \sigma_y} \cdot E((x - \mu_x) \cdot (y - \mu_y)) = \frac{\sigma_{xy}}{\sigma_x \cdot \sigma_y} = \rho \quad (10.23)$$

Damit folgt Gleichung (10.22) zu

$$E(z^2) = t^2 + 2 \cdot t \cdot \rho + 1 \quad (10.24)$$

Um zu zeigen, dass der Korrelationskoeffizient ρ nie größer als ± 1 wird, muss Gleichung (10.24) umgeformt werden zu

$$E(z^2) = (t + \rho)^2 + 1 - \rho^2 \quad (10.25)$$

Die Varianz einer Zufallsvariable ist als die Quadratsumme der Abstände der Einzelwerte zum Mittelwert definiert und damit immer größer oder gleich null. Gleichung (10.25) genügt somit der Ungleichung

$$E(z^2) = (t + \rho)^2 + 1 - \rho^2 \geq 0 \quad (10.26)$$

Da der Teilausdruck $(t + \rho)^2$ stets größer oder gleich 0 ist, muss dies auch für den Teilausdruck $1 - \rho^2$ gelten, um die Ungleichung (10.26) nicht zu verletzen. Daraus ergibt sich, dass

$$\rho^2 \leq 1 \quad (10.27)$$

und damit

$$\rho \leq |1| \quad (10.28)$$

gilt.

Damit ist bewiesen, dass der Korrelationskoeffizient ρ zwischen -1 und 1 liegt.

$$-1 \leq \rho \leq 1 \quad (10.29)$$

Da der Korrelationskoeffizient der Stichprobe ein Schätzwert für den Korrelationskoeffizienten der Grundgesamtheit ist, gilt diese Ungleichung auch für den Korrelationskoeffizienten der Stichprobe.

$$-1 \leq r \leq 1 \quad (10.30)$$

10.2.4 Korrelation bei unabhängigen Zufallsvariablen

Handelt es sich bei den Zufallsvariablen x und y um unabhängige Zufallsvariablen, gilt für deren Kovarianz σ_{xy}

$$\sigma_{xy} = E((x - \mu_x) \cdot (y - \mu_y)) = 0 \quad (10.31)$$

Dieser Zusammenhang wird in Kapitel 8 bewiesen. Aus der Definition des Korrelationskoeffizienten

$$\rho = \frac{\sigma_{xy}}{\sigma_x \cdot \sigma_y} \quad (10.32)$$

folgt, dass dieser im Falle unabhängiger Zufallsvariablen ebenfalls zu null wird.

$$\rho = \frac{\sigma_{xy}}{\sigma_x \cdot \sigma_y} = \frac{0}{\sigma_x \cdot \sigma_y} = 0 \quad (10.33)$$

Bei einer zweidimensionalen Normalverteilung gilt auch die Umkehrung dieses Satzes. Im Allgemeinen kann aber nicht davon ausgegangen werden, dass bei verschwindendem Korrelationskoeffizienten die Variablen unabhängig sind. Der Korrelationskoeffizient ist kein allgemeines Maß für die Abhängigkeit, sondern ein Maß für die lineare Abhängigkeit der beiden Größen x und y .

10.2.5 Korrelation bei linearer Abhängigkeit der Zufallsvariablen

Im Folgenden wird gezeigt, welchen Wert der Korrelationskoeffizient annimmt, wenn die Zufallsvariablen linear voneinander abhängig sind. In diesem Fall gilt die Beziehung

$$y = b \cdot x + k \quad (10.34)$$

Der Korrelationskoeffizient der beiden Zufallsvariablen x und y kann durch die Gleichung

$$\rho = \frac{\sigma_{xy}}{\sigma_x \cdot \sigma_y} \quad (10.35)$$

berechnet werden. Die Kovarianz σ_{xy} wird durch den Erwartungswertoperator definiert

$$\sigma_{xy} = E((x - \mu_x)(y - \mu_y)) \quad (10.36)$$

Durch Einsetzen der Zufallsvariablen y aus Gleichung (10.34) in die Definitionsgleichung der Kovarianz folgt

$$\begin{aligned} E((x - \mu_x)(y - \mu_y)) &= E((x - E(x))(b \cdot x + k - E(b \cdot x + k))) \\ &= E((x - E(x))(b \cdot (x - E(x)))) = b \cdot E((x - E(x))^2) \end{aligned} \quad (10.37)$$

Die Standardabweichung σ_x und die Standardabweichung σ_y lassen sich ebenfalls mithilfe des Erwartungswertoperators ausdrücken zu

$$\sigma_x = \sqrt{E((x - \mu_x)^2)} = \sqrt{E((x - E(x))^2)} \quad (10.38)$$

und

$$\begin{aligned} \sigma_y &= \sqrt{E((y - \mu_y)^2)} = \sqrt{E((b \cdot x + k - E(b \cdot x + k))^2)} = \sqrt{E((b \cdot (x - E(x)))^2)} \\ &= |b| \cdot \sqrt{E((x - E(x))^2)} \end{aligned} \quad (10.39)$$

Mit Gleichung (10.37), Gleichung (10.38) und Gleichung (10.39) folgt aus der Definitionsgleichung des Korrelationskoeffizienten ρ

$$\rho = \frac{\sigma_{xy}}{\sigma_x \cdot \sigma_y} = \frac{b \cdot E((x - E(x))^2)}{\sqrt{E((x - E(x))^2)} \cdot |b| \cdot \sqrt{E((x - E(x))^2)}} = \frac{b}{|b|} \quad (10.40)$$

Der Korrelationskoeffizient ρ weist damit für positive Werte b einen Wert von $\rho = 1$ auf, für einen negativen Koeffizienten b ergibt sich ein Wert von $\rho = -1$. Der Korrelationskoeffizient ρ nimmt damit bei exakt linearer Abhängigkeit den Betrag 1 an. Für Abhängigkeiten, die sich approximativ als Gerade beschreiben lassen, ergibt sich ein Wert in den Grenzen von -1 und 1. Bei unabhängigen Zufallsvariablen besteht kein linearer Zusammenhang und der Korrelationskoeffizient nimmt den Wert 0 an, was in Abschnitt 10.2.3 gezeigt wird. Zusätzlich kann über das Vorzeichen des Korrelationskoeffizienten eine Aussage darüber getroffen werden, ob mit steigenden Werten für die Zufallsvariable x die Werte der Zufallsvariablen ab- oder zunehmen.

10.3 Bewertung des Korrelationskoeffizienten

Die Verallgemeinerung der Korrelation einer Stichprobe auf die Grundgesamtheit erfordert die Bestimmung des Konfidenzbereiches und einen Hypothesentest für die Prüfung der Korrelation auf Signifikanz.

10.3.1 Bewertung des Korrelationskoeffizienten

Aus der Literatur sind zwei unterschiedliche Zufallsvariablen zur Bewertung der Korrelationskoeffizienten bekannt [Fahr96].

Zufallsvariable für Korrelationskoeffizienten $\rho = 0$

Liegt für eine Stichprobe mit dem Stichprobenumfang N die Korrelation r nahe dem Wert $\rho = 0$, sind die Zufallsvariablen x und y voraussichtlich statistisch unabhängig. Für einen Test auf Unabhängigkeit wird die Hypothese $\rho = 0$ geprüft. Für diesen Test wird die Zufallsvariable

$$t = r \cdot \sqrt{\frac{N - 2}{1 - r^2}} \quad (10.41)$$

eingesetzt. Sie weist für $\rho = 0$ eine t-Verteilung mit $N - 2$ Freiheitsgraden auf.

Zufallsvariable für Korrelationskoeffizienten $\rho = \rho_0$

Die Zufallsgröße t aus Gleichung (10.41) geht von einer Unabhängigkeit der Zufallsgrößen x und y aus. Soll die Korrelation ρ auf einen Wert $\rho_0 \neq 0$ gepüft oder der Konfidenzbereich des Korrelationskoeffizienten angegeben werden, ist diese Voraussetzung nicht erfüllt. Aus diesem Grund wird in diesem Fall die Zufallsvariable

$$z = \frac{1}{2} \cdot \left(\ln\left(\frac{1+r}{1-r}\right) - \ln\left(\frac{1+\rho}{1-\rho}\right) \right) \cdot \sqrt{N-3} = (\tanh^{-1}(r) - \tanh^{-1}(\rho)) \cdot \sqrt{N-3} \quad (10.42)$$

für eine Bewertung des Korrelationskoeffizienten eingeführt. Sie ist für einen Stichprobenumfang $N > 25$ asymptotisch standardnormalverteilt.

Auf Basis dieser Zufallsvariablen und ihren Verteilungen kann der Konfidenzbereich des Korrelationskoeffizienten ρ bestimmt und mit einem Hypothesentest die Hypothese geprüft werden.

10.3.2 Konfidenzbereich des Korrelationskoeffizienten

Der Konfidenzbereich des Korrelationskoeffizienten wird über die standardnormalverteilte Zufallsvariable

$$z = \frac{1}{2} \cdot \left(\ln\left(\frac{1+r}{1-r}\right) - \ln\left(\frac{1+\rho}{1-\rho}\right) \right) \cdot \sqrt{N-3} = (\tanh^{-1}(r) - \tanh^{-1}(\rho)) \cdot \sqrt{N-3} \quad (10.43)$$

ausgewertet. Die Wahrscheinlichkeit γ , mit der die Variable z in dem Intervall $c_1 \dots c_2$ liegt, ist definiert als

$$P(c_1 < z \leq c_2) = F(c_2) - F(c_1) = \gamma \quad (10.44)$$

Bei Annahme eines symmetrischen Konfidenzbereiches ergeben sich die Konstanten c_1 und c_2 aus der inversen Standardnormalverteilung zu

$$c_1 = F^{-1}\left(\frac{1 - \gamma}{2}\right) \quad (10.45)$$

und

$$c_2 = F^{-1}\left(\frac{1 + \gamma}{2}\right) \quad (10.46)$$

Durch Umformungen ergibt sich ein Ausdruck für den Konfidenzbereich des Korrelationskoeffizienten ρ der Grundgesamtheit.

$$\begin{aligned}
 \gamma &= P\left(c_1 < \left(\tanh^{-1}(r) - \tanh^{-1}(\rho)\right) \cdot \sqrt{N-3} \leq c_2\right) \\
 &= P\left(\frac{c_1}{\sqrt{N-3}} < \tanh^{-1}(r) - \tanh^{-1}(\rho) \leq \frac{c_2}{\sqrt{N-3}}\right) \\
 &= P\left(\frac{c_1}{\sqrt{N-3}} - \tanh^{-1}(r) < -\tanh^{-1}(\rho) \leq \frac{c_2}{\sqrt{N-3}} - \tanh^{-1}(r)\right) \\
 &= P\left(\tanh^{-1}(r) - \frac{c_2}{\sqrt{N-3}} < \tanh^{-1}(\rho) \leq \tanh^{-1}(r) - \frac{c_1}{\sqrt{N-3}}\right) \\
 &= P\left(\tanh\left(\tanh^{-1}(r) - \frac{c_2}{\sqrt{N-3}}\right) < \rho \leq \tanh\left(\tanh^{-1}(r) - \frac{c_1}{\sqrt{N-3}}\right)\right)
 \end{aligned} \tag{10.47}$$

Mit der gewählten Wahrscheinlichkeit γ liegt der Korrelationskoeffizienten ρ der Grundgesamtheit in dem angegebenen Konfidenzintervall. Das Vorgehen zur Bestimmung des Konfidenzintervalls für den Korrelationskoeffizienten ρ wird in Tabelle 10.4 zusammengefasst.

Tabelle 10.4: Vorgehen zur Bestimmung des Konfidenzintervalls für den Korrelationskoeffizienten ρ

Nr.	Prozessschritt
1	Wahl eines Signifikanzniveaus γ
2	Bestimmung des zugehörigen Parameter c_1 und c_2 aus der inversen Standardnormalverteilung $c_1 = F^{-1}\left(\frac{1-\gamma}{2}\right)$ und $c_2 = F^{-1}\left(\frac{1+\gamma}{2}\right)$
3	Berechnung des Korrelationskoeffizienten der Stichprobe mit dem Stichprobenumfang N $r = \frac{s_{xy}}{s_x \cdot s_y} = \frac{\sum_{n=1}^N ((x_n - \bar{x}) \cdot (y_n - \bar{y}))}{\sqrt{\sum_{n=1}^N (x_n - \bar{x})^2} \cdot \sqrt{\sum_{n=1}^N (y_n - \bar{y})^2}}$
4	Bestimmung des Konfidenzintervalls $\tanh\left(\tanh^{-1}(r) - \frac{c_2}{\sqrt{N-3}}\right) < \rho \leq \tanh\left(\tanh^{-1}(r) - \frac{c_1}{\sqrt{N-3}}\right)$

10.3.3 Hypothesentest zum Korrelationskoeffizienten einer zweidimensionalen Stichprobe

Die Zufallsvariable zur Bewertung des Korrelationskoeffizienten ρ und ihre Verteilung sind von der Korrelation der Grundgesamtheit ρ_0 abhängig. Damit sind auch die Hypothesentests zum Korrelationskoeffizienten einer zweidimensionalen Stichprobe von der Korrelation der Grundgesamtheit ρ_0 abhängig.

Zufallsvariable für Korrelationskoeffizienten $\rho = 0$

Bei einer zweidimensionalen Normalverteilung ist der Korrelationskoeffizient r einer Stichprobe ein Schätzwert für den Korrelationskoeffizienten ρ der Grundgesamtheit. In diesem Fall kann die Hypothese $\rho = 0$ gegen eine Alternative, zum Beispiel $\rho \neq 0$ getestet werden. Es wird also geprüft, ob die Grundgesamtheit unkorreliert ist. Das Verfahren dazu basiert darauf, dass die Zufallsvariable

$$t = r \cdot \sqrt{\frac{N - 2}{1 - r^2}} \quad (10.48)$$

bei der Richtigkeit der Hypothese eine t-Verteilung mit $N - 2$ Freiheitsgraden besitzt. In diesem Fall sollte der Betrag von t klein sein. Liegt t außerhalb berechneter Grenzen, wird die Hypothese $\rho = 0$ verworfen. Mit der Kenntnis der Verteilung ist es möglich, einen Ausdruck für den Annahmebereich der Nullhypothese, dass die geschätzte Korrelation ρ mit einer spezifizierten Wahrscheinlichkeit γ einen Wert $\rho = 0$ aufweist, anzugeben.

$$\gamma = 1 - \alpha = P(c_1 < t \leq c_2) = P(c_1 < r \cdot \sqrt{\frac{N - 2}{1 - r^2}} \leq c_2) \quad (10.49)$$

Die Konstante c_1 und c_2 ergeben sich dabei aus der inversen t-Verteilung mit $N - 2$ Freiheitsgraden. Bei Annahme eines symmetrischen Konfidenzbereiches ergeben sich die Konstanten c_1 und c_2 aus

$$c_1 = F^{-1}\left(\frac{1 - \gamma}{2}\right) = F^{-1}\left(\frac{\alpha}{2}\right) \quad (10.50)$$

und

$$c_2 = F^{-1}\left(\frac{1 + \gamma}{2}\right) = F^{-1}\left(1 - \frac{\alpha}{2}\right) \quad (10.51)$$

Die Umrechnung des Annahmebereiches in Grenzen für die Korrelation r führt zu einer quadratischen Gleichung, die nicht eindeutig aufgelöst werden kann. Aus diesem Grund wird der Hypothesentest direkt für die Variable t durchgeführt. Liegt der Stichprobenmittelwert t in dem Annahmebereich

$$c_1 \leq t < c_2 \quad (10.52)$$

kann die Nullhypothese beibehalten werden, andernfalls gilt die Alternativhypothese. Alternativ kann auch der p-Value mit der Gleichung

$$p = F(t) \quad (10.53)$$

berechnet werden. Das Vorgehen mit den einzelnen Prozessschritten ist in Tabelle 10.5 dargestellt.

Tabelle 10.5: Test der Hypothese $\rho = 0$ gegen $\rho \neq 0$ für die Korrelation einer zweidimensionalen Stichprobe

Nr.	Prozessschritt	
1	Wahl eines Signifikanzniveaus α	
2	Bestimmung der zugehörigen Parameter c_1 und c_2 aus der inversen t-Verteilung mit $N - 2$ Freiheitsgraden $F(c_1) = \frac{\alpha}{2}$ und $F(c_2) = 1 - \frac{\alpha}{2}$	
3	Berechnung des Korrelationskoeffizienten der Stichprobe mit dem Stichprobenumfang N $r = \frac{s_{xy}}{s_x \cdot s_y} = \frac{\sum_{n=1}^N ((x_n - \bar{x}) \cdot (y_n - \bar{y}))}{\sqrt{\sum_{n=1}^N (x_n - \bar{x})^2} \cdot \sqrt{\sum_{n=1}^N (y_n - \bar{y})^2}}$	
4	Berechnung der Größe t aus dem Korrelationskoeffizienten r der Stichprobe und einem Stichprobenumfang N $t = r \cdot \sqrt{\frac{N-2}{1-r^2}}$	
5	Bestimmung des Annahmebereichs $c_1 \leq t < c_2$	Berechnung des p-Values mit der f-Verteilung $p = F(t)$
6	Für $c_1 \leq t < c_2$ wird die Hypothese angenommen, für $t \leq c_1$ oder $t > c_2$ wird die Hypothese verworfen	Für $\alpha/2 \leq p < 1 - \alpha/2$ wird die Hypothese angenommen, für $p < \alpha/2$ und $p \geq 1 - \alpha/2$ wird die Hypothese verworfen

Zufallsvariable für Korrelationskoeffizienten $\rho = \rho_0$

Ein Test der Hypothese $\rho = \rho_0$ soll gegen eine Alternative, zum Beispiel $\rho \neq \rho_0$ getestet werden. Dieser Test erfolgt auf Basis der standardnormalverteilten Zufallsvariable

$$z = \frac{1}{2} \cdot \left(\ln\left(\frac{1+r}{1-r}\right) - \ln\left(\frac{1+\rho_0}{1-\rho_0}\right) \right) \cdot \sqrt{N-3} = (\tanh^{-1}(r) - \tanh^{-1}(\rho_0)) \cdot \sqrt{N-3} \quad (10.54)$$

Ist die Hypothese $\rho = \rho_0$ erfüllt, sollte der Betrag von z klein sein. Liegt z außerhalb berechneter Grenzen, wird die Hypothese $\rho = \rho_0$ verworfen. Damit kann ein Ausdruck für den Annahmebereich der Nullhypothese, dass die geschätzte Korrelation ρ mit einer spezifizierten Wahrscheinlichkeit γ einen Wert $\rho = \rho_0$ aufweist, angegeben werden.

$$\gamma = P(c_1 < z \leq c_2) = P\left(c_1 < (\tanh^{-1}(r) - \tanh^{-1}(\rho_0)) \cdot \sqrt{N-3} \leq c_2\right) \quad (10.55)$$

Die Konstante c_1 und c_2 ergeben sich dabei aus der inversen Standardnormalverteilung. Bei Annahme eines symmetrischen Konfidenzbereiches ergeben sich die Konstanten c_1 und c_2 aus

$$c_1 = F^{-1}\left(\frac{1-\gamma}{2}\right) = F^{-1}\left(\frac{\alpha}{2}\right) \quad (10.56)$$

und

$$c_2 = F^{-1} \left(\frac{1 + \gamma}{2} \right) = F^{-1} \left(1 - \frac{\alpha}{2} \right) \quad (10.57)$$

Die Ungleichung (10.55) kann umgeformt werden zu

$$\tanh \left(\tanh^{-1}(\rho_0) + \frac{c_1}{\sqrt{N-3}} \right) = r_{c1} < r \leq r_{c2} = \tanh \left(\tanh^{-1}(\rho_0) + \frac{c_2}{\sqrt{N-3}} \right) \quad (10.58)$$

Alternativ kann auch der p-Value mit der Gleichung

$$p = F(z) \quad (10.59)$$

berechnet werden. Das Vorgehen mit den einzelnen Prozessschritten ist in Tabelle 10.6 dargestellt.

Tabelle 10.6: Test der Hypothese $\rho = \rho_0$ gegen $\rho \neq \rho_0$ für die Korrelation einer zweidimensionalen Stichprobe

Nr.	Prozessschritt	
1	Wahl eines Signifikanzniveaus α	
2	Bestimmung der zugehörigen Parameter c_1 und c_2 aus der inversen t-Verteilung mit $N - 2$ Freiheitsgraden $F(c_1) = \frac{\alpha}{2}$ und $F(c_2) = 1 - \frac{\alpha}{2}$	
3	Berechnung des Korrelationskoeffizienten der Stichprobe mit dem Stichprobenumfang N $r = \frac{s_{xy}}{s_x \cdot s_y} = \frac{\sum_{n=1}^N ((x_n - \bar{x}) \cdot (y_n - \bar{y}))}{\sqrt{\sum_{n=1}^N (x_n - \bar{x})^2} \cdot \sqrt{\sum_{n=1}^N (y_n - \bar{y})^2}}$	
4	Berechnung der Größe z aus dem Korrelationskoeffizienten r der Stichprobe und dem Stichprobenumfang N $z = (\tanh^{-1}(r) - \tanh^{-1}(\rho_0)) \cdot \sqrt{N - 3}$	
5	Bestimmung der Annahmegrenzen $r_{c1} = \tanh\left(\tanh^{-1}(\rho_0) + \frac{c_1}{\sqrt{N - 3}}\right)$ $r_{c2} = \tanh\left(\tanh^{-1}(\rho_0) + \frac{c_2}{\sqrt{N - 3}}\right)$ Bestimmung des Annahmebereichs $r_{c1} < r \leq r_{c2}$	Berechnung des p-Values mit Standardnormalverteilung $p = F(t)$
	Für $r_{c1} \leq r < r_{c2}$ wird die Hypothese angenommen, für $r \leq r_{c1}$ oder $r > r_{c2}$ wird die Hypothese verworfen	Für $\alpha/2 \leq p < 1 - \alpha/2$ wird die Hypothese angenommen, für $p < \alpha/2$ und $p \geq 1 - \alpha/2$ wird die Hypothese verworfen

Beispiel: Lagerspiel eines Motors

Zur Untersuchung von Alterungserscheinungen bei Motoren wurde das Wellenlagerspiel w / mm und das Pleuellagerspiel p / mm aufgenommen. Die gemessenen Werte sind in der folgenden Tabelle wiedergegeben.

Tabelle 10.7: Messreihe zur Bewertung des Lagerspiels eines Motors

Probe	1	2	3	4	5	6	7
Wellenlagerspiel w / mm	1.4300	1.0325	1.4241	1.6251	0.4493	0.1108	0.6877
Pleuellagerspiel p / mm	0.8286	0.1602	0.4373	0.6851	0.7219	0.8106	1.8924

Die gemessenen Werte des Wellenlagerspiels w und des Pleuellagerspiels p sollen auf Korrelation überprüft werden. Um sich einen ersten Eindruck über die Daten zu bekommen, werden diese zunächst grafisch in einem Streudiagramm dargestellt.

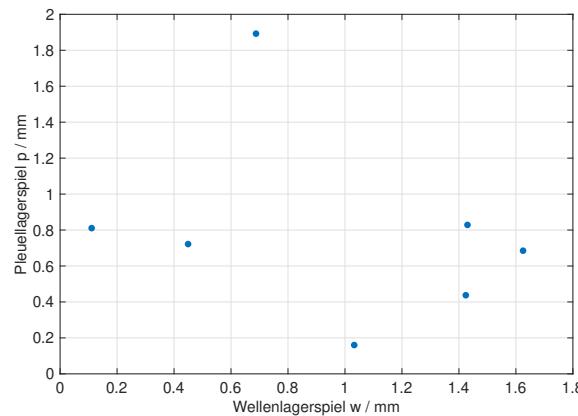


Bild 10.4: Streudiagramm des Lagerspiels eines Motors

In Bild 10.4 ist keine eindeutige Tendenz der Messwerte zu erkennen, sodass von einer geringen Korrelation der Messwerte ausgegangen werden kann. Eine nähere Aussage liefert die numerische Auswertung des Korrelationskoeffizienten r der Stichprobe. Dieser folgt zu

$$r = \frac{s_{pw}}{s_p \cdot s_w} = -0.2949 \quad (10.60)$$

Der Korrelationskoeffizient ist leicht negativ. Das heißt, dass mit steigendem Wert des Wellenlagerspiels der Wert des Pleuellagerspiels leicht abnimmt. Mit einem Betrag von $r < 0.5$ bestätigt sich die Annahme einer schwachen Korrelation.

Mithilfe eines Hypothesentests soll die Signifikanz der Korrelation untersucht werden. Hierzu wird die Hypothese $\rho = 0$ gegen die Alternative $\rho \neq 0$ bei einem Signifikanzniveau von $\alpha = 5\%$ geprüft. Die Grenzen ergeben sich aus der inversen t-Verteilung mit $N - 2$ Freiheitsgraden zu

$$c_1 = -2.5706 \quad (10.61)$$

und

$$c_2 = 2.5706 \quad (10.62)$$

Die Bewertungsgröße t berechnet sich zu

$$t = r \cdot \sqrt{\frac{N - 2}{1 - r^2}} = -0.6901 \quad (10.63)$$

Da der Wert t innerhalb der berechneten Grenzen c_1 und c_2 liegt, wird die Hypothese $\rho = 0$ angenommen. Entsprechendes zeigt auch die aus der t-Verteilung mit $N - 2$ Freiheitsgraden berechnete Wahrscheinlichkeit von

$$p = F(t) = 52.09\% \quad (10.64)$$

Sie sagt aus, mit welcher Wahrscheinlichkeit eine Korrelation von -0.2949 auftritt, wenn der Korrelationskoeffizient $\rho = 0$ ist. Der Wahrscheinlichkeitswert liegt zwischen den Werten $\alpha/2$ und $1 - \alpha/2$. Die Hypothese $\rho = 0$ wird deshalb nicht verworfen. Die Auswertung mithilfe des Hypothesentestes entspricht der auf Basis des Streudiagrammes vorgenommenen Bewertung.

Da auch der Konfidenzbereich des Korrelationskoeffizienten zur Bewertung dessen Signifikanz herangezogen werden kann, soll dieser zusätzlich berechnet werden. Es gilt

$$\tanh\left(\tanh^{-1}(r) - \frac{c_2}{\sqrt{N-3}}\right) < \rho \leq \tanh\left(\tanh^{-1}(r) - \frac{c_1}{\sqrt{N-3}}\right) \quad (10.65)$$

Mit einer Konfidenzzahl $\gamma = 95\%$ errechnen sich die benötigten Grenzen c_1 und c_2 aus der inversen Standardnormalverteilung zu $c_1 = -1.96$ und $c_2 = 1.96$. Mit dem bereits berechneten Korrelationskoeffizienten der Stichprobe $r = -0.2949$ und dem Stichprobenumfang von $N = 7$ berechnet sich der Konfidenzbereich mit

$$\rho_1 = \tanh\left(\tanh^{-1}(r) - \frac{c_2}{\sqrt{N-3}}\right) = -0.8575 \quad (10.66)$$

und

$$\rho_2 = \tanh\left(\tanh^{-1}(r) - \frac{c_1}{\sqrt{N-3}}\right) = 0.5890 \quad (10.67)$$

zu

$$-0.8575 \leq \rho \leq 0.5890 \quad (10.68)$$

Da der Konfidenzbereich des Korrelationskoeffizienten den Wert $\rho = 0$ mit einschließt, bestätigt der Konfidenzbereich die Aussage des Hypothesentestes.

In MATLAB zeigt sich die Berechnung wie folgt:

```

1 % Messwerte einlesen
2 load Lagerspiel.mat;
3
4 % Berechnung der Kenngrößen
5 [R,P,RLO,RUP] = corrcoef([p w])

```

Dabei steht in dem Rückgabeparameter R der Funktion die Korrelationsmatrix \mathbf{R} für die Größen p und w . Der Wahrscheinlichkeitswert P kann dazu verwendet werden, die Signifikanz des Korrelationskoeffizienten r zu überprüfen. Liegt der Wert P zwischen $\alpha/2$ und $1 - \alpha/2$, kann die Hypothese $\rho = 0$ auf Basis der vorliegenden Stichprobe nicht verworfen werden. In den beiden Zielgrößen RLO und RUP sind die obere beziehungsweise die untere Grenze des Konfidenzbereiches des Korrelationskoeffizienten enthalten.

Unterschiede bei der Berechnung des Wertes P sind darauf zurückzuführen, dass MATLAB im Widerspruch zu bekannten Literaturstellen mit der t-Verteilung statt der Standardnormalverteilung rechnet. Für Stichprobenumfänge mit $N > 25$ ist der Unterschied zu vernachlässigen.

In Python existiert eine Funktion zur Berechnung des Korrelationskoeffizienten und des Hypothesentests für $\rho = 0$, aber es existiert keine Funktion zur Berechnung des Konfidenzbereichs. Die Funktion kann aber mit den Angaben in Tabelle 10.6 selbst implementiert werden.

```

1 Bibliotheken importieren
2 import numpy as np
3 import pandas as pd
4 import scipy.stats as stats
5 from scipy.io import loadmat
6 from scipy.stats import norm
7
8 Stichprobenwerte aus Aufgabe übernehmen und als Dataframe speichern
9 data = loadmat('Lagerspiel')
10 df = pd.DataFrame({'w': data['values'][1, :],
11                     'p': data['values'][2, :]})
12
13 Berechnung Korrelation und des Hypothesentests roh = 0
14 Corr, p = stats.pearsonr(df['w'], df['p'])
15
16 Berechnung des Konfidenzbereichs
17 N = np.size(df['w'])
18 gamma = 0.95
19 c1 = norm.ppf((1-gamma)/2)
20 c2 = norm.ppf((1+gamma)/2)
21 roh1 = np.tanh(np.arctanh(Corr)-c2/np.sqrt(N-3))
22 roh2 = np.tanh(np.arctanh(Corr)-c1/np.sqrt(N-3))

```

10.4 Bewertung des Korrelationskoeffizienten mehrdimensionaler Stichproben

Äquivalent zu dem in Abschnitt 10.3.2 eingeführten Konfidenzbereich für den Korrelationskoeffizienten kann auch für Korrelationsmatrizen ein Konfidenzbereich angegeben werden.

$$\rho_{1jk} \leq \rho_{jk} \leq \rho_{2jk} \quad (10.69)$$

Dabei berechnen sich die Werte ρ_{1jk} und ρ_{2jk} wie bei zweidimensionalen Stichproben zu

$$\rho_{1jk} = \tanh\left(\tanh^{-1}(r_{jk}) - \frac{c_2}{\sqrt{N-3}}\right) \quad (10.70)$$

beziehungsweise

$$\rho_{2jk} = \tanh\left(\tanh^{-1}(r_{jk}) - \frac{c_1}{\sqrt{N-3}}\right) \quad (10.71)$$

Der Konfidenzbereich der Korrelationsmatrix kann damit beschrieben werden durch

$$\begin{pmatrix} 1 & \rho_{112} & \dots & \rho_{11M} \\ \rho_{121} & 1 & \dots & \rho_{12M} \\ \dots & \dots & \dots & \dots \\ \rho_{1M1} & \rho_{1M2} & \dots & 1 \end{pmatrix} \leq P \leq \begin{pmatrix} 1 & \rho_{212} & \dots & \rho_{21M} \\ \rho_{221} & 1 & \dots & \rho_{22M} \\ \dots & \dots & \dots & \dots \\ \rho_{2M1} & \rho_{2M2} & \dots & 1 \end{pmatrix} \quad (10.72)$$

Liegt der Wert $\rho_{jk} = 0$ innerhalb des Konfidenzbereiches sind die entsprechenden Datensätze nicht signifikant korreliert.

Da die Größen paarweise miteinander verglichen werden, kann auch der Hypothesentest für die Korrelationskoeffizienten durchgeführt werden. Dabei gelten dieselben Annahmen und Herleitungen wie für den Hypothesentest zweidimensionaler Stichproben.

Beispiel: Emissionsuntersuchung

Für eine Umweltstudie soll die Abhängigkeit der Fahrzeugemissionen E / g/km vom Hubraum H / cm³ des PKWs, der Leistung P / kW und der Masse M / kg untersucht werden. Hierzu wurden die in Tabelle 10.8 dargestellten Werte aufgenommenen.

Tabelle 10.8: Messreihe zur Untersuchung der Emission von Fahrzeugen

Hubraum H / cm ³	Leistung P / kW	Masse M / kg	Emissionen E /g/km
1197	63	1205	113
1197	81	1210	114
1197	81	1229	112
1395	81	1382	94
1395	81	1382	124
1395	81	1395	92
1395	81	1395	119
1395	192	1225	120
1395	92	1249	116
1395	110	1268	119
1395	110	1293	121
1395	110	1288	116
1395	110	1270	109
1395	110	1296	112
1395	110	1290	110
1984	162	1351	139
1984	162	1370	149
1984	169	1382	165
1984	169	1402	159
1984	221	1476	99
1984	221	1495	102
1598	81	1299	119
1598	81	1317	85
1598	81	1432	106
1598	81	1265	117
1968	110	1354	119
1968	110	1375	122
1968	135	1394	109
1968	135	1449	119
1968	135	1377	122

Die Daten werden zunächst grafisch dargestellt.

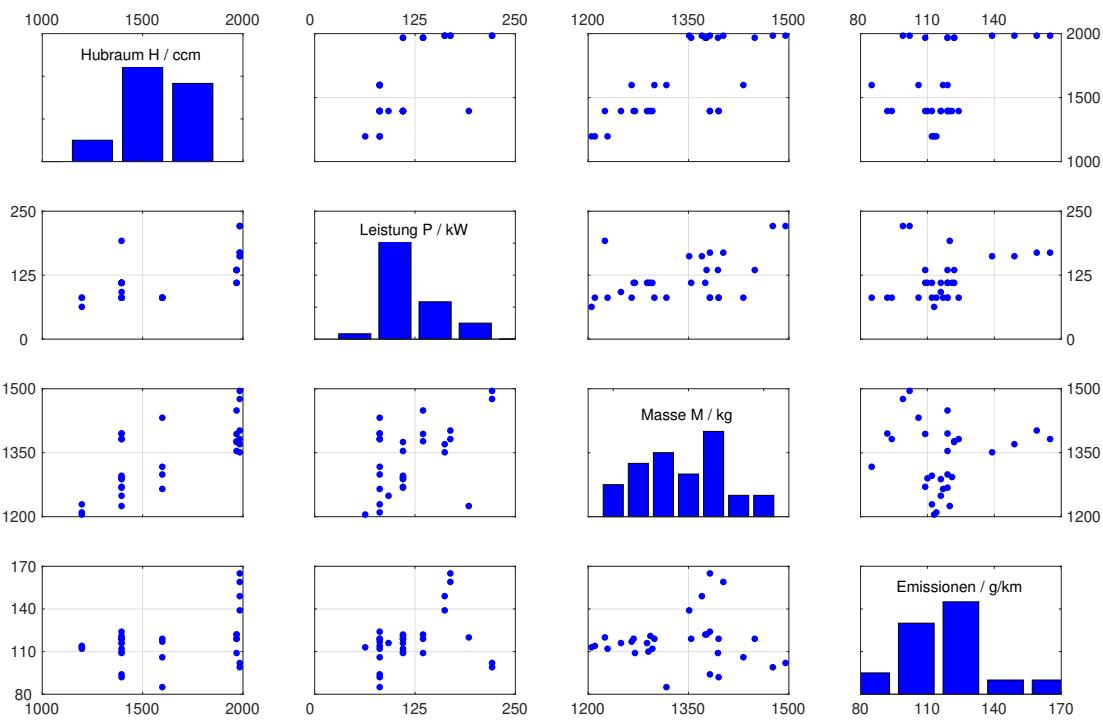


Bild 10.5: Streudiagramm-Matrix der Messwerte

Es ergibt sich die Korrelationsmatrix von

$$R = \begin{pmatrix} 1 & 0.6641 & 0.7122 & 0.4019 \\ 0.6641 & 1 & 0.4789 & 0.3345 \\ 0.7122 & 0.4789 & 1 & 0.0261 \\ 0.4019 & 0.3345 & 0.0261 & 1 \end{pmatrix} \quad (10.73)$$

Zwischen den untersuchten Größen existiert nur eine schwache bis mittlere Korrelation. Die höchste Korrelation besteht zwischen Hubraum und Masse sowie Hubraum und Leistung. Motoren mit größerem Hubraum haben ein höheres Gewicht und eine höhere Leistung.

Für das Beispiel aus Tabelle 10.8 ergibt sich die Wahrscheinlichkeit p dafür, dass die Stichproben nicht korreliert sind, gegen die Alternative, dass sie eine von null verschiedene Korrelation haben, von

$$P(R = 0) = \begin{pmatrix} 0 & 0.0001 & 0.0000 & 0.0277 \\ 0.0001 & 0 & 0.0074 & 0.0708 \\ 0.0000 & 0.0074 & 0 & 0.8910 \\ 0.0277 & 0.0708 & 0.8910 & 0 \end{pmatrix} \quad (10.74)$$

In Kombination mit dem Signifikanzniveau von $\alpha = 0.05$ ergibt sich, dass alle Korrelationen signifikant sind bis auf die Korrelation zwischen Masse und Emissionen sowie Leistung und Emissionen. Zu derselben Aussage kommt die Bewertung des Konfidenzbereiches des Korrelationskoeffizienten.

$$\rho_{1jk} \leq \rho_{jk} \leq \rho_{2jk} \quad (10.75)$$

Für das Beispiel aus Tabelle 10.8 errechnet sich der Konfidenzbereich zu

$$\begin{pmatrix} 1 & 0.3994 & 0.4733 & 0.0487 \\ 0.3994 & 1 & 0.1433 & -0.0293 \\ 0.4733 & 0.1433 & 1 & -0.3373 \\ 0.0487 & 0.0293 & -0.3373 & 1 \end{pmatrix} \leq P \leq \begin{pmatrix} 1 & 0.8266 & 0.8535 & 0.6658 \\ 0.8266 & 1 & 0.7157 & 0.6201 \\ 0.8535 & 0.7157 & 1 & 0.3828 \\ 0.6658 & 0.6201 & 0.3828 & 1 \end{pmatrix} \quad (10.76)$$

Dabei wird folgende MATLAB-Befehlssequenz verwendet.

```
1 % Messwerte einlesen
2 load Fahrzeugemissionen.mat;
3
4 % Berechnung der Kenngrößen
5 [R,P,RLO,RUP] = corrcoef(values)
```

Der Wert $\rho = 0$ liegt für die Korrelationen von Masse und Emissionen sowie Leistung und Emissionen innerhalb des Konfidenzbereiches, diese Korrelationskoeffizienten sind somit nicht signifikant. Ein Vergleich mit den Ergebnissen des Hypothesentest zeigt, dass sich beide Aussagen entsprechen.

Nach diesem Datensatz hängen die Fahrzeugemissionen im Wesentlichen von dem Hubraum des Motors ab.

In Python ist die Berechnung der Korrelationsmatrix ebenfalls möglich, der Hypothesentest für $\rho = 0$ kann jedoch nur für zwei Vektoren durchgeführt werden. Die Berechnung des Konfidenzbereichsmuss mit den Angaben in Tabelle 10.6 selbst implementiert werden.

```
1 Bibliotheken importieren
2 import pandas as pd
3 import matplotlib.pyplot as plt
4 from scipy.io import loadmat
5
6 Stichprobenwerte aus Aufgabe übernehmen und als Dataframe speichern
7 data = loadmat('Fahrzeugemissionen')
8 df = pd.DataFrame({'H': data['values'][:, 0],
9                     'P': data['values'][:, 1],
10                    'M': data['values'][:, 2],
11                    'E': data['values'][:, 3]})
12
13 Grafische Darstellung als Streudiagramm-Matrix
14 fig = plt.figure(1, figsize=(12,8))
15 fig.suptitle('')
16 ax1 = fig.subplots(1,1)
17 ax1= pd.plotting.scatter_matrix(df, alpha=1, Color='b', hist_kwds=dict(
18                               Color='b'))
19 Berechnung Korrelation und des Hypothesentests roh = 0
20 Corr1 = df.corr(method='pearson')
```

10.5 Korrelation und Kausalzusammenhang

Der Korrelationskoeffizient stellt ein Maß für den numerischen Zusammenhang zweier Größen dar. Häufig wird die Korrelation zweier Größen dazu benutzt, einen Hinweis darauf zu bekommen, ob zwei statistische Größen ursächlich oder kausal miteinander zusammenhängen.

Ein kleiner Betrag des Korrelationskoeffizienten nahe null weist darauf hin, dass die Größen nicht oder nur gering korreliert sind. Damit ist es unwahrscheinlich, dass die beiden untersuchten Größen der Stichprobe einen Kausalzusammenhang aufweisen.

Aus einem großen Betrag des Korrelationskoeffizienten kann auf eine starke numerische Korrelation geschlossen werden. Damit muss aber nicht zwangsläufig ein starker Kausalzusammenhang im Sinne von Ursache und Wirkung zwischen den Größen bestehen. Ein Kausalzusammenhang kann nur durch eine geeignete Systemanalyse begründet werden. Der Korrelationskoeffizient von $r = 1$ bestätigt, dass die beiden Größen linear voneinander abhängig sind, sagt aber nichts über den Proportionalitätsfaktor zwischen den Größen aus.

Kann der Kausalzusammenhang nicht bestätigt werden, wird die Korrelation als Scheinkorrelation zwischen zwei Größen bezeichnet. Solche Scheinzusammenhänge können dadurch entstehen, dass eine mit beiden Größen korrelierte dritte Größe existiert, die indirekt einen Kausalzusammenhang zwischen den beiden zunächst analysierten Größen vortäuscht.

Ein bekanntes Beispiel dazu ist die Korrelation zwischen dem Rückgang der Störche im Burgenland und einem Rückgang der Anzahl Neugeborener. Diese Ereignisse haben nichts miteinander zu tun - weder bringen Störche Kinder noch umgekehrt. Ein großer Korrelationskoeffizient bedeutet, sie sind kausal allenfalls über eine dritte Größe miteinander verbunden, etwa über die Verstädterung, die sowohl Nistplätze vernichtet als auch Kleinfamilien ohne Kinder fördert.

Um einen Kausalzusammenhang wirklich herstellen und um Proportionalitätsfaktoren bestimmen zu können, wären Experimente nötig, bei denen ein Faktor experimentell festgelegt wird und die andere Größe gemessen wird. In Bereichen, wo solche Experimente oftmals aus Gründen der zu langen Dauer oder zu hoher Kosten nicht möglich sind, kommt die Korrelation zum Einsatz.

Ist es möglich Experimente durchzuführen, würden die Ergebnisse mithilfe der Regressionsanalyse bewertet werden. Im Gegensatz zur Korrelation beschreibt die in Kapitel 11 behandelte Regressionsanalyse einen funktionalen Zusammenhang. Die Korrelation beschreibt dagegen lediglich einen statistischen Zusammenhang zwischen den untersuchten Größen.

10.6 Literatur

- [Krey91] Kreyszig, Erwin: Statistische Methoden und ihre Anwendungen
4., unveränderter Nachdruck der 7. Auflage
Vandenhoeck & Ruprecht, Göttingen, 1991
- [Fahr96] Fahrmeir, Ludwig; Hamerle, Alfred; Tutz, Gerhard: Multivariate statistische Verfahren
2., überarbeitete Auflage
Walter de Gruyter & Co., Berlin
- [Ross06] Ross, M. Sheldon: Statistik für Ingenieure und Naturwissenschaftler
3. Auflage
Spektrum Akademischer Verlag, München, 2006
- [Hart07] Hartung, Joachim; Elpelt, Bärbel: Multivariate Statistik
7., unveränderte Auflage
R. Oldenbourg Verlag, München / Wien
- [Papu01] Papula, Lothar: Mathematik für Ingenieure und Naturwissenschaftler Band 3
4., verbesserte Auflage
Vieweg Teubner, Braunschweig / Wiesbaden, 2008

11 Kapitel11

12 Regression zweidimensionaler Datensätze

Eines der zentralen Ziele von Design For Six Sigma besteht darin, Ursachen-Wirkungs-Beziehungen zu ermitteln, um Produkte und Prozesse bestmöglich zu verstehen. Zur Beschreibung des Zusammenhangs von Eingangs-, Stör- und Zielgrößen werden unterschiedliche Methoden angewendet:

- Analytische Berechnung Auf Basis eines Systemmodells werden die Zielgrößen als Funktion der Eingangsgrößen berechnet. Notwendig ist ein physikalisches Verhaltensmodell, das das Systemverhalten mit einer ausreichenden Präzision beschreibt.
- Simulation Eine Simulation wird durchgeführt, wenn das Modell zwar grundsätzlich bekannt ist, die analytische Berechnung aber aufgrund der komplexen Geometrie oder den Randbedingungen unübersichtlich und aufwendig wird.
- Experiment Mit Experimenten werden die Simulationen und Berechnungen bestätigt. Mit dem Hintergrundwissen zum Modellverhalten, das aus der analytischen Rechnung und der Simulation kommt, kann der experimentelle Aufwand klein gehalten werden.

Der Zusammenhang zwischen den unterschiedlichen Methoden zur Modellierung eines Systems ist in Bild 12.1 zusammenfassend dargestellt.

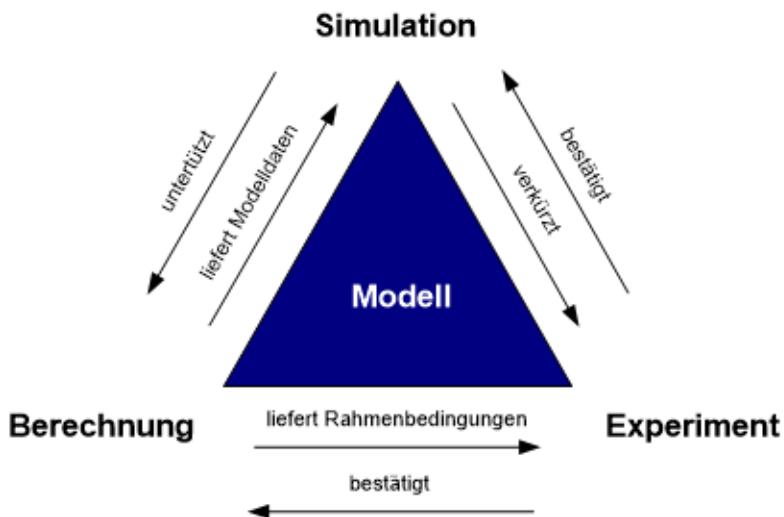
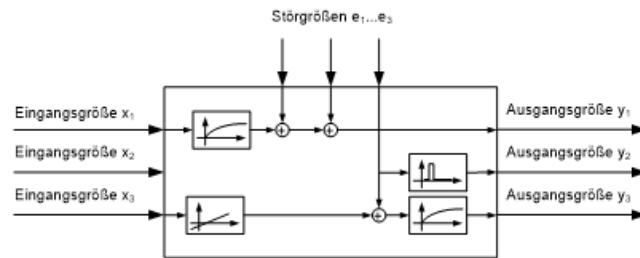


Bild 12.1: Gegenüberstellung von analytischer Berechnung, numerischer Simulation und Experiment

Insbesondere die Simulation, aber auch die analytische Berechnung erfordern ein Modell des abzubildenden Systems oder Prozesses. Es werden mathematische und physikalische Modelle unterschieden, beide werden in Bild 12.1 gegenübergestellt.

a) Physikalisches Systemmodell



b) Mathematisches Systemmodell

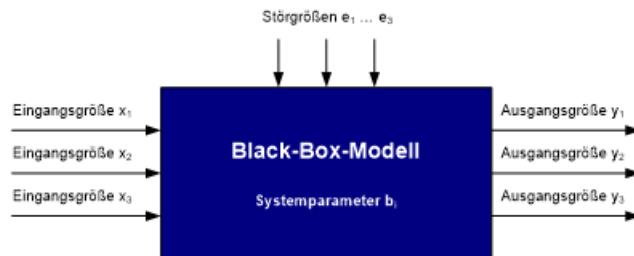


Bild 12.2: Vergleich von physikalischen und mathematischen Systemmodellen

Während in der klassischen Modellbildung das physikalische Modell analytisch hergeleitet wird, wird im Rahmen der mathematischen Modellbildung das zu untersuchende System als Black-Box betrachtet. Dieser Ansatz führt zu einem Modell, bei dem das Zusammenwirken der Ein- und Zielgrößen mathematisch über sogenannte Regressionsfunktionen beschrieben wird.

In diesem Kapitel werden zunächst Regressionsfunktionen für zweidimensionale Datensätze berechnet und ihr Konfidenzbereich ermittelt. Dabei werden lineare und nichtlineare Regressionen betrachtet. In Kapitel 12 wird das Vorgehen auf M-dimensionale Datensätze verallgemeinert.

12.1 Lineare Regression

Anhand der linearen Regression zweidimensionaler Datensätze werden die Grundprinzipien der Regression und die wesentlichen statistischen Verfahren zur Bewertung der Regressionsfunktion vorgestellt. Liegt eine Beobachtung mit einer Stichprobe

$$(x_1, y_1), (x_2, y_2), \dots, (x_N, y_N) \quad (12.1)$$

aus einer zweidimensionalen Grundgesamtheit vor, können diese Punkte oftmals in guter Näherung durch eine Geradengleichung der Form

$$y(x) = b_0 + b_1 \cdot x \quad (12.2)$$

beschrieben werden kann. Der Parameter b_0 beschreibt den Schnittpunkt mit der y-Achse, während der Parameter b_1 die Steigung der Geraden beschreibt. Der durch die Funktion geschätzte Wert wird mit $\hat{y}(x)$ bezeichnet.

Beispiel: Temperatursensor

Die Ausgangssituation wird an einem Beispiel verdeutlicht, bei dem der Zusammenhang zwischen der Temperatur eines Motoröls und der Ausgangsspannung eines Temperatursensors beschrieben werden soll. Die aufgenommenen Stichprobenwerte sind in Tabelle 12.1 aufgelistet.

Tabelle 12.1: Zusammenhang zwischen Öltemperatur und Ausgangsspannung eines Temperatursensors

Temperatur T / °C	0	10	20	30	40	50
Spannung U / V	2.766	2.862	3.005	3.120	3.173	3.411
Temperatur T / °C	60	70	80	90	100	
Spannung U / V	3.676	3.803	3.944	4.188	4.165	

In Bild 12.3 wird die Stichprobe als Streudiagramm visualisiert. Der funktionale Zusammenhang zwischen der Öltemperatur und der Ausgangsspannung des Temperatursensors kann in diesem Fall über eine Gerade approximiert werden. Hierzu ist in Bild 12.3 zusätzlich die entsprechende Regressionsgerade eingezeichnet.

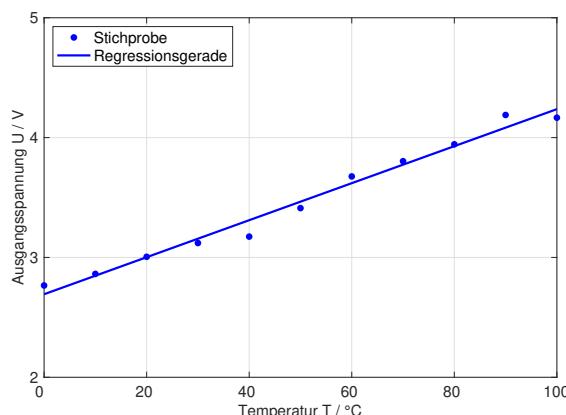


Bild 12.3: Grafische Darstellung der Stichprobe und der entsprechenden Regressionsgerade für den Zusammenhang zwischen der Öltemperatur und der Ausgangsspannung des Temperatursensors

Die Punkte in Bild 12.3 liegen nicht präzise auf einer Linie, sodass das Einzeichnen einer Geraden zunächst nicht eindeutig ist. Um insbesondere bei großen Datenmengen eine Funktion eindeutig berechnen zu können, wurde von Gauß das Prinzip der kleinsten Quadrate entwickelt. Nach diesem Prinzip ist die Gerade so zu legen, dass die Summe der Quadrate aller Abstände von den Stichprobenwerten zu der Geraden möglichst klein wird. Die durch dieses Verfahren bestimmte Funktion wird als Regressionsfunktion bezeichnet. Der Abstand der Stichprobenwerte von der Regressionsgeraden wird als Residuum bezeichnet. 12.4 verdeutlicht den Begriff des Residuums.

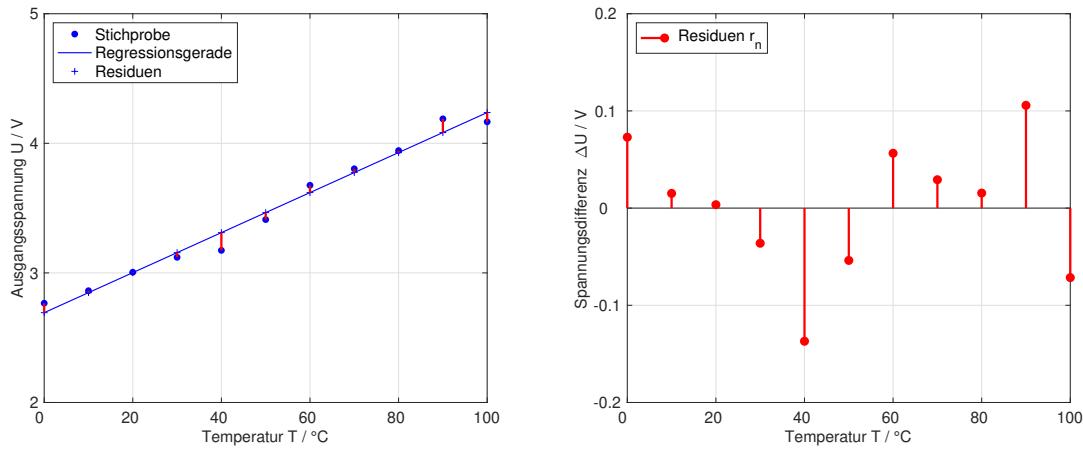


Bild 12.4: Grafische Darstellung des Abstandes der Stichprobenwerte von der Regressionsgeraden (Residuen r_n)

Wird die Regressionsgerade über die Gleichung

$$y(x) = b_0 + b_1 \cdot x \quad (12.3)$$

beschrieben, ergibt sich für das Wertepaar (x_n, y_n) sich das Residuum r_n zu

$$r_n = y_n - y(x_n) = y_n - (b_0 + b_1 \cdot x_n) \quad (12.4)$$

Die Parameter b_0 und b_1 sollen so bestimmt werden, dass die Summe der Abstandsquadrate a minimal wird. Die Summe der Abstandsquadrate a errechnet sich zu

$$a = \sum_{n=1}^N r_n^2 = \sum_{n=1}^N (y_n - b_0 - b_1 \cdot x_n)^2 \quad (12.5)$$

Damit diese Funktion ein Minimum aufweist, müssen die partiellen Ableitungen von a nach den Parametern b_m der Regressionsfunktion verschwinden. Damit ergeben sich die notwendigen Bedingungen

$$\frac{\partial a}{\partial b_1} = 0 \quad (12.6)$$

und

$$\frac{\partial a}{\partial b_0} = 0 \quad (12.7)$$

Die Ableitungen aus Gleichung (12.6) und Gleichung (12.7) berechnen sich zu

$$\frac{\partial a}{\partial b_1} = -2 \cdot \sum_{n=1}^N (x_n \cdot (y_n - b_0 - b_1 \cdot x_n)) \quad (12.8)$$

und

$$\frac{\partial a}{\partial b_0} = -2 \cdot \sum_{n=1}^N (y_n - b_0 - b_1 \cdot x_n) \quad (12.9)$$

Beide Ausdrücke werden zu null gesetzt, und es ergeben sich die Gleichungen

$$0 = \sum_{n=1}^N x_n \cdot (y_n - b_0 - b_1 \cdot x_n) = \sum_{n=1}^N x_n \cdot y_n - b_0 \cdot \sum_{n=1}^N x_n - b_1 \cdot \sum_{n=1}^N x_n^2 \quad (12.10)$$

und

$$0 = \sum_{n=1}^N (y_n - b_0 - b_1 \cdot x_n) = \sum_{n=1}^N y_n - \sum_{n=1}^N b_0 - b_1 \cdot \sum_{n=1}^N x_n = \sum_{n=1}^N y_n - N \cdot b_0 - b_1 \cdot \sum_{n=1}^N x_n \quad (12.11)$$

Unter Berücksichtigung der Definition für den Mittelwert

$$\bar{x} = \frac{1}{N} \cdot (x_1 + x_2 + \dots + x_N) = \frac{1}{N} \cdot \sum_{n=1}^N x_n \quad (12.12)$$

beziehungsweise

$$\bar{y} = \frac{1}{N} \cdot (y_1 + y_2 + \dots + y_N) = \frac{1}{N} \cdot \sum_{n=1}^N y_n \quad (12.13)$$

ergibt sich aus Gleichung (12.10)

$$b_0 \cdot N \cdot \bar{x} + b_1 \cdot \sum_{n=1}^N x_n^2 = \sum_{n=1}^N x_n \cdot y_n \quad (12.14)$$

und aus Gleichung (12.12) folgt

$$b_0 + b_1 \cdot \bar{x} = \bar{y} \quad (12.15)$$

Dieses Gleichungssystem kann nach b_1 und b_0 aufgelöst werden. Mit den Ausdrücken für die Varianz

$$s_x^2 = \frac{1}{N-1} \cdot \sum_{n=1}^N (x_n - \bar{x})^2 = \frac{1}{N-1} \cdot \left(\sum_{n=1}^N x_n^2 - N \cdot \bar{x}^2 \right) \quad (12.16)$$

und die Kovarianz

$$s_{xy} = \frac{1}{N-1} \cdot \sum_{n=1}^N ((x_n - \bar{x}) \cdot (y_n - \bar{y})) = \frac{1}{N-1} \cdot \left(\sum_{n=1}^N x_n \cdot y_n - N \cdot \bar{x} \cdot \bar{y} \right) \quad (12.17)$$

können sie umgeformt werden zu

$$b_1 = \frac{\sum_{n=1}^N x_n \cdot y_n - N \cdot \bar{x} \cdot \bar{y}}{\sum_{n=1}^N x_n^2 - N \cdot \bar{x}^2} = \frac{s_{xy}}{s_x^2} \quad (12.18)$$

und

$$b_0 = \bar{y} - b_1 \cdot \bar{x} \quad (12.19)$$

Damit ergibt sich die Geradengleichung

$$y(x) - \bar{y} = b_1 \cdot (x - \bar{x}) \quad (12.20)$$

und die Summe aller Residuen errechnet sich zu

$$\begin{aligned} \sum_{n=1}^N r_n &= \sum_{n=1}^N (y_n - y(x_n)) = \sum_{n=1}^N (y_n - b_1 \cdot (x_n - \bar{x}) - \bar{y}) = \sum_{n=1}^N (y_n - b_1 \cdot x_n + b_1 \cdot \bar{x} - \bar{y}) \\ &= \sum_{n=1}^N y_n - b_1 \cdot \sum_{n=1}^N x_n + b_1 \cdot N \cdot \bar{x} - N \cdot \bar{y} = N \cdot \bar{y} - b_1 \cdot N \cdot \bar{x} + b_1 \cdot N \cdot \bar{x} - N \cdot \bar{y} = 0 \end{aligned} \quad (12.21)$$

Für das Beispiel des Öltemperatursensors aus Tabelle 11.1 ergeben sich die Koeffizienten b_1 und b_0 zu

$$b_1 = \frac{\sum_{n=1}^N x_n \cdot y_n - N \cdot \bar{x} \cdot \bar{y}}{\sum_{n=1}^N x_n^2 - N \cdot \bar{x}^2} = \frac{s_{xy}}{s_x^2} = 0.0154 \quad (12.22)$$

und

$$b_0 = \bar{y} - b_1 \cdot \bar{x} = 2.693 \quad (12.23)$$

Die entsprechende Gerade ist bereits in Bild 12.3 und Bild 12.4 eingezeichnet.

Das Verfahren zur Berechnung einer Regressionsgeraden auf Basis einer Stichprobe ist in Tabelle 12.2 zusammengefasst.

Tabelle 12.2: Vorgehen zur Bestimmung des Parameters b_1 und b_0 einer Regressionsgeraden

Nr.	Prozessschritt
1	Berechnen der Mittelwerte aus der Stichprobe $\bar{x} = \frac{1}{N} \cdot \sum_{n=1}^N x_n$ und $\bar{y} = \frac{1}{N} \cdot \sum_{n=1}^N y_n$
2	Berechnen der Varianz und der Kovarianz aus der Stichprobe $s_x^2 = \frac{1}{N-1} \cdot \sum_{n=1}^N (x_n - \bar{x})^2$ und $s_{xy} = \frac{1}{N-1} \cdot \sum_{n=1}^N (x_n - \bar{x}) \cdot (y_n - \bar{y})$
3	Berechnen der Parameter b_0 und b_1 der Regressionsgeraden $b_1 = \frac{s_{xy}}{s_x^2}$ und $b_0 = \bar{y} - b_1 \cdot \bar{x}$
4	Beschreiben der Stichprobe mit der Regressionsfunktion $y(x) = b_1 \cdot x + b_0 = b_1 \cdot (x - \bar{x}) + \bar{y}$

Damit ist das Vorgehen zur Festlegung der Regressionsgeraden einer Stichprobe festgelegt. Für diese Stichprobe ist die berechnete Gerade im Sinne der Summe der Fehlerquadrate die bestmögliche Lösung. Wie bei der Bestimmung der Konfidenzintervalle bleibt aber zunächst die Frage offen, inwieweit die bestimmte Geradengleichung die Grundgesamtheit wiedergibt.

12.1.1 Modell zur statistischen Bewertung der Regression

Zur statistischen Bewertung der Regression wird die Grundgesamtheit der Zielgröße beschrieben als

$$y = \beta_0 + \beta_1 \cdot x + e \quad (12.24)$$

Dabei ist die Größe x die Eingangsgröße, für die die Zielgröße y bestimmt werden soll. Sie wird nicht durch den Prozess festgelegt und besitzt damit keine Streuung. Die Messwerte unterliegen einem Messfehler. Für den Messfehler wird davon ausgegangen, dass er normalverteilt ist. Er weist einen Erwartungswert von

$$E(e) = \mu_e = 0 \quad (12.25)$$

und eine Varianz

$$\left((e - \mu_e)^2 \right) = \sigma^2 \quad (12.26)$$

auf. Bild 12.5 stellt die Annahmen zur statistischen Bewertung der Regression grafisch dar.

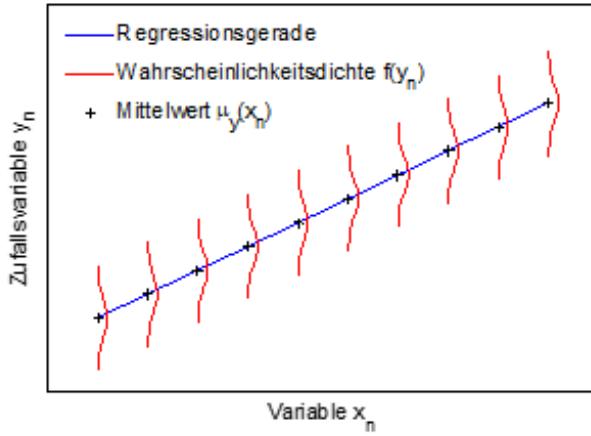


Bild 12.5: Grafische Darstellung der Annahmen zur Bewertung der Regressionskoeffizienten

Damit hat die Zielgröße y den Erwartungswert

$$E(y) = E(\beta_0 + \beta_1 \cdot x + e) = \beta_0 + \beta_1 \cdot x + E(e) = \beta_0 + \beta_1 \cdot x \quad (12.27)$$

Da ausschließlich der Messfehler e streut, hat die Zielgröße y eine Varianz von

$$\sigma_y^2 = \sigma^2 \quad (12.28)$$

Der Regressionskoeffizient b_1 wird mit Gleichung (12.18) berechnet zu

$$b_1 = \frac{\sum_{n=1}^N ((x_n - \bar{x}) \cdot (y_n - \bar{y}))}{\sum_{n=1}^N (x_n - \bar{x})^2} \quad (12.29)$$

Der Ausdruck kann mit

$$(y_n - \bar{y}) = \beta_1 \cdot (x_n - \bar{x}) + (e_n - \bar{e}) \quad (12.30)$$

umgeformt werden zu

$$\begin{aligned} b_1 &= \frac{\sum_{n=1}^N ((x_n - \bar{x}) \cdot (\beta_1 \cdot (x_n - \bar{x}) + (e_n - \bar{e})))}{\sum_{n=1}^N (x_n - \bar{x})^2} = \frac{\sum_{n=1}^N \beta_1 \cdot (x_n - \bar{x})^2 + (x_n - \bar{x}) \cdot (e_n - \bar{e})}{\sum_{n=1}^N (x_n - \bar{x})^2} \\ &= \frac{\beta_1 \cdot \sum_{n=1}^N (x_n - \bar{x})^2 + \sum_{n=1}^N (x_n \cdot e_n - x_n \cdot \bar{e} - \bar{x} \cdot e_n + \bar{x} \cdot \bar{e})}{\sum_{n=1}^N (x_n - \bar{x})^2} = \beta_1 + \frac{\sum_{n=1}^N (x_n - \bar{x}) \cdot e_n}{\sum_{n=1}^N (x_n - \bar{x})^2} \end{aligned} \quad (12.31)$$

Der Erwartungswert des Messfehlers ist null. Damit ist die Schätzung b_1 des Regressionskoeffizienten β_1 erwartungstreu. Die Varianz der Schätzung berechnet sich

$$\sigma_{b_1}^2 = \frac{\sum_{n=1}^N (x_n - \bar{x})^2 \cdot \sigma^2}{\left(\sum_{n=1}^N (x_n - \bar{x})^2 \right)^2} = \frac{\sigma^2}{\sum_{n=1}^N (x_n - \bar{x})^2} = \frac{\sigma^2}{(N-1) \cdot s_x^2} \quad (12.32)$$

Der Regressionskoeffizient b_0 errechnet sich nach Gleichung (12.19) zu

$$b_0 = \bar{y} - b_1 \cdot \bar{x} = \beta_0 + \beta_1 \cdot \bar{x} + \bar{e} - b_1 \cdot \bar{x} = \beta_0 + (\beta_1 - b_1) \cdot \bar{x} + \bar{e} \quad (12.33)$$

Da die Schätzung des Regressionskoeffizienten β_1 erwartungstreu ist und der Messfehler e mittelwertsfrei ist, ist die Schätzung b_0 des Regressionskoeffizienten β_0 erwartungstreu.

$$E(b_0) = E(\beta_0 + (\beta_1 - b_1) \cdot \bar{x} + \bar{e}) = E(\beta_0) + E((\beta_1 - b_1) \cdot \bar{x}) + E(\bar{e}) = \beta_0 \quad (12.34)$$

Aus der Gleichung für den Regressionskoeffizienten

$$b_0 = \bar{y} - b_1 \cdot \bar{x} = \frac{1}{N} \cdot \sum_{n=1}^N y_n - \frac{\sum_{n=1}^N (x_n - \bar{x}) \cdot y_n}{\sum_{n=1}^N x_n^2 - N \cdot \bar{x}^2} \cdot \bar{x} \quad (12.35)$$

ergibt sich mit den Rechenregeln für Funktionen mehrerer Zufallsvariablen eine Varianz von

$$\sigma_{b_0}^2 = \frac{1}{N^2} \cdot N \cdot \sigma_y^2 + \frac{\sigma_y^2}{(N-1) \cdot s_x^2} \cdot \bar{x}^2 = \frac{\sum_{n=1}^N x_n^2 - N \cdot \bar{x}^2 + N \cdot \bar{x}^2}{N \cdot (N-1) \cdot s_x^2} \cdot \sigma_y^2 = \frac{\sum_{n=1}^N x_n^2}{N \cdot (N-1) \cdot s_x^2} \cdot \sigma^2 \quad (12.36)$$

Zur Berechnung der unterschiedlichen Varianzen wird die unbekannte Varianz der Messung σ^2 benötigt. Zur Abschätzung der Varianz werden die Residuen verwendet. Ihre Stichprobenvarianz ergibt sich aus

$$\sigma^2 \approx s^2 = \frac{1}{N-2} \cdot \sum_{n=1}^N r_n^2 = \frac{1}{N-2} \cdot \sum_{n=1}^N (y_n - b_0 - b_1 \cdot x_n)^2 = \frac{a}{N-2} \quad (12.37)$$

Die Summe wird durch den Faktor $N - 2$ geteilt, da bei der Berechnung von b_0 und b_1 zwei Freiheitsgrade verloren gehen.

12.1.2 Bewertung des Regressionskoeffizienten β_1

Zur Bewertung des Regressionskoeffizienten β_1 wird die Zufallsvariable

$$z = \frac{b_1 - \beta_1}{\sigma_{b_1}} = \frac{b_1 - \beta_1}{\sigma} \cdot \sqrt{N - 1} \cdot s_x \quad (12.38)$$

aufgestellt. Mit den Vorüberlegungen in Abschnitt 11.1.1 weist sie eine Standard-Normalverteilung auf. Die Varianz σ^2 wird mit Gleichung (12.37) geschätzt. Die Schätzung weist $N - 2$ Freiheitsgrade auf. Damit ist die Größe

$$t = \frac{b_1 - \beta_1}{\sqrt{a}} \cdot \sqrt{N - 1} \cdot \sqrt{N - 2} \cdot s_x = \frac{b_1 - \beta_1}{s_{b1}} \quad (12.39)$$

eine t-Verteilung mit $N - 2$ Freiheitsgraden auf, und die Standardabweichung s_{b1} errechnet sich zu

$$s_{b1} = \frac{\sqrt{a}}{\sqrt{N - 1} \cdot \sqrt{N - 2} \cdot s_x} \quad (12.40)$$

Konfidenzintervall für den Regressionskoeffizienten β_1

Zur Berechnung des Konfidenzintervalls des Regressionskoeffizienten β_1 wird die Zufallsvariable t verwendet. Nach den Ausführungen in Kapitel 5 berechnet sich der Konfidenzbereich aus der Wahrscheinlichkeit

$$P(c_1 < t \leq c_2) = F(c_2) - F(c_1) = \gamma \quad (12.41)$$

Durch die Symmetrie des Konfidenzbereichs ergeben sich die Konstanten c_1 und c_2 zu

$$c_1 = F^{-1}\left(\frac{1 - \gamma}{2}\right) \quad (12.42)$$

und

$$c_2 = F^{-1}\left(\frac{1 + \gamma}{2}\right) \quad (12.43)$$

Durch Umformungen von Gleichung (12.41) ergibt sich ein Ausdruck für den Konfidenzbereich des Regressionskoeffizienten β_1 von

$$\gamma = P(c_1 < t \leq c_2) = P\left(c_1 < \frac{b_1 - \beta_1}{s_{b1}} \leq c_2\right) = P(b_1 - c_2 \cdot s_{b1} \leq \beta_1 < b_1 - c_1 \cdot s_{b1}) \quad (12.44)$$

Die Berechnung wird in Tabelle 12.3 zusammengefasst.

Tabelle 12.3: Vorgehen zur Bestimmung des Konfidenzbereichs für den Regressionskoeffizienten β_1

Nr.	Prozessschritt
1	Wahl einer Konfidenzzahl γ
2	Bestimmung der zugehörigen Parameter c_1 und c_2 aus der inversen t-Verteilung mit $N - 2$ Freiheitsgraden $c_1 = F^{-1}\left(\frac{1-\gamma}{2}\right) \quad \text{und} \quad c_2 = F^{-1}\left(\frac{1-\gamma}{2}\right)$
3	Berechnung der Mittelwerte der Stichprobe $\bar{x} = \frac{1}{N} \cdot \sum_{n=1}^N x_n \quad \text{und} \quad \bar{y} = \frac{1}{N} \cdot \sum_{n=1}^N y_n$
4	Bestimmung der Standardabweichung und der Kovarianz der Stichprobe $s_x = \sqrt{\frac{1}{N-1} \cdot \sum_{n=1}^N (x_n - \bar{x})^2} \quad \text{und} \quad s_{xy} = \frac{1}{N-1} \cdot \left(\sum_{n=1}^N x_n \cdot y_n - N \cdot \bar{x} \cdot \bar{y} \right)$
5	Bestimmung der Parameter b_0 und b_1 der Regressionsgeraden $b_1 = \frac{\sum_{n=1}^N x_n \cdot y_n - N \cdot \bar{x} \cdot \bar{y}}{\sum_{n=1}^N x_n^2 - N \cdot \bar{x}^2} = \frac{s_{xy}}{s_x^2} \quad \text{und} \quad b_0 = \bar{y} - b_1 \cdot \bar{x}$
6	Berechnung der Fehlerquadratsumme $a = \sum_{n=1}^N r_n^2 = \sum_{n=1}^N (y_n - b_1 \cdot x_n - b_0)^2$
7	Bestimmung des Konfidenzintervalls $b_1 - c_2 \cdot s_{b1} \leq \beta_1 < b_1 + c_1 \cdot s_{b1}$

Beispiel: Temperatursensor

Die Durchführung der Rechnung wird wieder an dem Beispiel des Temperatursensors aus Tabelle 11.1 verdeutlicht. Es liegen $N = 11$ Stichprobenwerte vor, sodass zur Bestimmung der Konstante c_1 und c_2 eine t-Verteilung mit 9 Freiheitsgraden herangezogen wird. Für eine Konfidenzzahl von $\gamma = 0.95$ ergeben sich die Grenze c_1 und c_2 zu

$$c_1 = F^{-1}(1 - 0.975) = -2.2622 \quad (12.45)$$

und

$$c_2 = F^{-1}(1 - 0.025) = 2.2622 \quad (12.46)$$

Mit den Angaben errechnet sich der Konfidenzbereich des Regressionskoeffizienten β_1 mit dem geschätzten Regressionsparameter $b_1 = 0.0154$ und der Standardabweichung $s_x = 33.1662$ zu

$$0.0138 \leq \beta_1 < 0.0170 \quad (12.47)$$

In das Temperatur-Spannungs-Diagramm in Bild 12.6 sind neben den Messwerten und der Regressionsgeraden mit dem geschätzten Regressionskoeffizienten b_1 noch zwei weitere Gerade eingezeichnet. Diese ergeben sich durch die Extremwerte des 95% - Konfidenzintervalls.

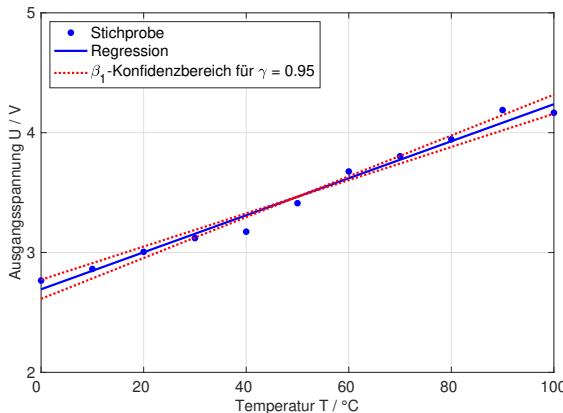


Bild 12.6: Grafische Darstellung der Stichprobe für den Zusammenhang zwischen der Öltemperatur und der Ausgangsspannung eines Temperatursensors mit unterschiedlichen Regressionsgeraden

Test des Regressionskoeffizienten β_1 auf Signifikanz

Zur Reduzierung der Komplexität von Regressionsfunktionen muss die Frage beantwortet werden, ob die berechneten Koeffizienten einen signifikanten Einfluss auf die Zielgröße besitzen. Es existieren verschiedene Verfahren zur Bewertung der Signifikanz von Regressionskoeffizienten, von denen zwei dargestellt werden sollen. Dies sind der t-Test und die Analyse des Konfidenzbereichs.

Zur Bewertung der Signifikanz des Regressionskoeffizienten β_1 wird die Nullhypothese aufgestellt, dass der Korrelationskoeffizient β_1 einem Wert $\beta_1 = 0$ entspricht. Trifft diese Hypothese zu, ist die Zielgröße y nicht von der Eingangsgröße x abhängig. Wird die Nullhypothese auf Basis der vorliegenden Stichprobe abgelehnt, kann davon ausgegangen werden, dass die Zielgröße y signifikant von dem Term $\beta_1 \cdot c$ abhängt.

Damit ein über die Stichprobe geschätzter Regressionskoeffizient b_1 mit einer spezifizierten Wahrscheinlichkeit zu der t-Verteilung aus Gleichung (12.39) gehört, muss dieser in dem Intervall $\beta_{1C1} < b_1 \leq \beta_{1C2}$ liegen. Wird die Wahrscheinlichkeit dafür mit γ bezeichnet, gilt die Gleichung

$$P(\beta_{1C1} < b_1 \leq \beta_{1C2}) = \gamma = 1 - \alpha \quad (12.48)$$

Mit der Verteilung aus Gleichung (12.39) wird die Wahrscheinlichkeit γ , mit der die Variable t innerhalb des Intervalls $c_1 \dots c_2$ liegt, definiert als

$$\gamma = P(c_1 < t \leq c_2) = F(c_2) - F(c_1) \quad (12.49)$$

Bei Annahme eines symmetrischen Tests ergeben sich die Konstanten c_1 und c_2 aus den Bedingungen

$$F(c_1) = \frac{1 - \gamma}{2} = \frac{\alpha}{2} \quad (12.50)$$

und

$$F(c_2) = 1 - \frac{1 - \gamma}{2} = 1 - \frac{\alpha}{2} \quad (12.51)$$

Auflösen nach c_1 und c_2 führt zu

$$c_1 = F^{-1}\left(\frac{\alpha}{2}\right) \quad (12.52)$$

und

$$c_2 = F^{-1} \left(1 - \frac{\alpha}{2} \right) \quad (12.53)$$

Durch Umformungen von Gleichung (12.39) und (12.49) ergibt sich ein Ausdruck für den Annahmebereich der Nullhypothese, nämlich dass der geschätzte Regressionskoeffizient b_1 mit einer spezifizierten Wahrscheinlichkeit γ zu der angenommenen t-Verteilung gehört.

$$\gamma = P(c_1 < t \leq c_2) = P(c_1 \cdot s_{b1} < b_1 \leq c_2 \cdot s_{b1}) \quad (12.54)$$

Alternativ kann, wie in Kapitel 5 gezeigt wird, eine Unterschreitungswahrscheinlichkeit p der Prüfgröße b_1 bestimmt werden und mit dem Signifikanzniveau α verglichen werden. Bei Hypothesentests mit beidseitigem Verwerfungsbereich $\beta_1 \neq 0$ müssen für die Annahme der Nullhypothese die Bedingungen

$$p = P(t_1) > \frac{\alpha}{2} \quad (12.55)$$

und

$$p = P(t_1) \leq 1 - \frac{\alpha}{2} \quad (12.56)$$

erfüllt werden. Damit lässt sich der Test mit der Hypothese $\beta_1 = 0$ und der Alternative $\beta_1 \neq 0$ in folgenden Prozessschritten zusammenfassen.

Tabelle 12.4: Test der Hypothese $\beta_1 = 0$ gegen $\beta_1 \neq 0$ für den Regressionskoeffizienten β_1 einer linearen Regression

Nr.	Prozessschritt	
1	Wahl eines Signifikanzniveaus α	
2	Bestimmung der zugehörigen Parameter c_1 und c_2 aus der inversen t-Verteilung mit N - 2 Freiheitsgraden $F(c_1) = \frac{\alpha}{2}$ und $F(c_2) = 1 - \frac{\alpha}{2}$	
3	Berechnung der Mittelwerte der Stichprobe $\bar{x} = \frac{1}{N} \cdot \sum_{n=1}^N x_n$ und $\bar{y} = \frac{1}{N} \cdot \sum_{n=1}^N y_n$	
4	Bestimmung der Standardabweichung und der Kovarianz der Stichprobe $s_x = \sqrt{\frac{1}{N-1} \cdot \sum_{n=1}^N (x_n - \bar{x})^2}$ und $s_{xy} = \frac{1}{N-1} \cdot \left(\sum_{n=1}^N x_n \cdot y_n - N \cdot \bar{x} \cdot \bar{y} \right)$	
5	Bestimmung der Parameter b_0 und b_1 der Regressionsgeraden $b_1 = \frac{\sum_{n=1}^N x_n \cdot y_n - N \cdot \bar{x} \cdot \bar{y}}{\sum_{n=1}^N x_n^2 - N \cdot \bar{x}^2} = \frac{s_{xy}}{s_x^2}$ und $b_0 = \bar{y} - b_1 \cdot \bar{x}$	
6	Berechnung der Fehlerquadratsumme $a = \sum_{n=1}^N r_n^2 = \sum_{n=1}^N (y_n - b_1 \cdot x_n - b_0)^2$	
7	Bestimmung der Standardabweichung s_{b1} $s_{b1} = \frac{\sqrt{a}}{\sqrt{N-1} \cdot \sqrt{N-2} \cdot s_x}$	
8	Berechnung der Grenzen des Annahmebereiches $\beta_{1C1} = c_1 \cdot s_{b1}$ und $\beta_{1C2} = c_2 \cdot s_{b1}$	
9	Bestimmung des Annahmebereichs $\beta_{1C1} < b_1 \leq \beta_{1C2}$	Berechnung des p-Wertes mit der t-Verteilung mit N - 2 Freiheitsgraden $p = F\left(\frac{b_1}{s_{b1}}\right)$
10	Für $\beta_{1C1} \leq b_1 < \beta_{1C2}$ wird die Hypothese angenommen, für $b_1 \leq \beta_{1C1}$ oder $b_1 > \beta_{1C2}$ wird die Hypothese verworfen	Für $\alpha/2 \leq p < 1 - \alpha/2$ wird die Hypothese angenommen, für $p < \alpha/2$ und $p \geq 1 - \alpha/2$ wird die Hypothese verworfen

Beispiel: Temperatursensor

Für das Beispiel des Temperatursensors aus Tabelle 12.1 wird die Nullhypothese $\beta_1 = 0$ gegen die Alternativhypothese $\beta_1 \neq 0$ getestet. Es liegen $N = 11$ Stichprobenwerte vor, sodass zur Bestimmung der Konstanten c_1 und c_2 eine t-Verteilung mit 9 Freiheitsgraden herangezogen wird. Für ein Signifikanzniveau $\alpha = 0.05$ ergeben sich die Grenzen c_1 und c_2 zu

$$c_1 = F^{-1}(1 - 0.975) = -2.2622 \quad (12.57)$$

und

$$c_2 = F^{-1}(1 - 0.025) = 2.2622 \quad (12.58)$$

Mit den obigen Angaben errechnet sich der Annahmebereich der Nullhypothese zu

$$-0.0016 < b_1 \leq 0.0016 \quad (12.59)$$

Der geschätzte Wert des Regressionskoeffizienten liegt mit $b_1 = 0.0154$ außerhalb dem in Gleichung (12.59) berechneten Annahmebereich. Die Nullhypothese wird somit auf Grundlage der vorliegenden Stichprobe verworfen, der Regressionskoeffizient β_1 ist ungleich 0 und damit signifikant.

Alternativ kann zur Bewertung des Hypothesentests auch der p-Wert herangezogen werden. Dieser berechnet sich für die vorliegenden Hypothesen durch

$$p = F\left(\frac{b_1}{s_{b1}}\right) \quad (12.60)$$

Mit einem Wert von $p = 1$ liegt dieser deutlich über der Grenze von $1 - \alpha/2 = 0.975$. Die Nullhypothese muss daher verworfen werden, womit die Signifikanz des Regressionskoeffizienten β_1 bestätigt wird.

Eine weitere Möglichkeit zur Überprüfung der Signifikanz eines Regressionskoeffizienten ist dessen Konfidenzbereich. In Kapitel 6 wird gezeigt, dass wegen der linearen Abhängigkeit ein Hypothesentest in die Betrachtung eines Konfidenzbereichs überführt werden kann. Dadurch wird eine einfache Interpretation des Konfidenzbereichs zur Bewertung der Signifikanz ermöglicht: Schließt das Konfidenzintervall des Regressionskoeffizienten der Grundgesamtheit den entsprechenden Stichprobenwert ein, ist der untersuchte Koeffizient nicht signifikant. Andernfalls wird der Regressionskoeffizient als signifikant angenommen.

Für das Beispiel aus Tabelle 11.1 wurde der Konfidenzbereich des Regressionskoeffizienten β_1 bestimmt zu

$$0.0138 < \beta_1 \leq 0.0170 \quad (12.61)$$

Dieses Konfidenzintervall schließt die Zahl 0 nicht mit ein, sodass der Regressionskoeffizient auch nach diesem Kriterium als signifikant angenommen werden muss.

12.1.3 Bewertung des Regressionskoeffizienten β_0

Die Vorgehensweise zur Berechnung der Verteilungsfunktion des Regressionskoeffizienten β_0 ist vergleichbar zu der des Regressionskoeffizienten β_1 . Es wird die Zufallsvariable

$$z = \frac{b_0 - \beta_0}{\sigma_{b_0}} = \frac{b_0 - \beta_0}{\sqrt{\frac{\sum_{n=1}^N x_n^2}{N \cdot (N-1)} \cdot \frac{\sigma}{s_x}}} = \frac{b_0 - \beta_0}{\sigma} \cdot \frac{\sqrt{N} \cdot \sqrt{N-1} \cdot s_x}{\sqrt{\sum_{n=1}^N x_n^2}} \quad (12.62)$$

aufgestellt. Mit den Vorüberlegungen in Abschnitt 11.1.1 weist sie eine Standard-Normalverteilung auf. Die Varianz σ^2 wird mit Gleichung (12.37) geschätzt. Die Schätzung weist $N - 2$ Freiheitsgrade auf. Damit ist die Größe

$$t = \frac{b_0 - \beta_0}{\sqrt{a} \cdot \sqrt{\sum_{n=1}^N x_n^2}} \cdot \sqrt{N} \cdot \sqrt{N-1} \cdot \sqrt{N-2} \cdot s_x = \frac{b_0 - \beta_0}{s_{b_0}} \quad (12.63)$$

eine t-Verteilung mit $N - 2$ Freiheitsgraden auf, und die Standardabweichung s_{b_0} errechnet sich zu

$$s_{b_0} = \frac{\sqrt{a} \cdot \sqrt{\sum_{n=1}^N x_n^2}}{\sqrt{N} \cdot \sqrt{N-1} \cdot \sqrt{N-2} \cdot s_x} \quad (12.64)$$

Konfidenzintervall für den Regressionskoeffizienten β_0

Zur Berechnung des Konfidenzintervalls des Regressionskoeffizienten β_0 wird analog zum Regressionskoeffizienten β_1 die Zufallsvariable t verwendet. Der Konfidenzbereich berechnet sich dabei aus der Wahrscheinlichkeit

$$P(c_1 < t \leq c_2) = F(c_2) - F(c_1) = \gamma \quad (12.65)$$

Durch die Symmetrie des Konfidenzbereichs ergeben sich die Konstanten c_1 und c_2 zu

$$c_1 = F^{-1}\left(\frac{1-\gamma}{2}\right) \quad (12.66)$$

und

$$c_2 = F^{-1}\left(\frac{1+\gamma}{2}\right) \quad (12.67)$$

Umformungen führen zu einem Ausdruck für den Konfidenzbereich des Regressionskoeffizienten β_0 .

$$\gamma = P(c_1 < t \leq c_2) = P\left(c_1 < \frac{b_0 - \beta_0}{s_{b_0}} \leq c_2\right) = P(b_0 - c_2 \cdot s_{b_0} \leq \beta_0 < b_0 - c_1 \cdot s_{b_0}) \quad (12.68)$$

Die Prozessschritte für das hier vorgestellte Vorgehen sind in Tabelle 11.3 zusammengefasst.

Tabelle 12.5: Vorgehen zur Bestimmung des Konfidenzbereichs für den Regressionskoeffizienten β_0

Nr.	Prozessschritt
1	Wahl einer Konfidenzzahl γ
2	Bestimmung der zugehörigen Parameter c_1 und c_2 aus der inversen t-Verteilung mit $N - 2$ Freiheitsgraden $c_1 = F^{-1}\left(\frac{1-\gamma}{2}\right) \quad \text{und} \quad c_2 = F^{-1}\left(\frac{1-\gamma}{2}\right)$
3	Berechnung der Mittelwerte der Stichprobe $\bar{x} = \frac{1}{N} \cdot \sum_{n=1}^N x_n \quad \text{und} \quad \bar{y} = \frac{1}{N} \cdot \sum_{n=1}^N y_n$
4	Bestimmung der Standardabweichung und der Kovarianz der Stichprobe $s_x = \sqrt{\frac{1}{N-1} \cdot \sum_{n=1}^N (x_n - \bar{x})^2} \quad \text{und} \quad s_{xy} = \frac{1}{N-1} \cdot \left(\sum_{n=1}^N x_n \cdot y_n - N \cdot \bar{x} \cdot \bar{y} \right)$
5	Bestimmung der Parameter b_0 und b_1 der Regressionsgeraden $b_1 = \frac{\sum_{n=1}^N x_n \cdot y_n - N \cdot \bar{x} \cdot \bar{y}}{\sum_{n=1}^N x_n^2 - N \cdot \bar{x}^2} = \frac{s_{xy}}{s_x^2} \quad \text{und} \quad b_0 = \bar{y} - b_1 \cdot \bar{x}$
6	Berechnung der Fehlerquadratsumme $a = \sum_{n=1}^N r_n^2 = \sum_{n=1}^N (y_n - b_1 \cdot x_n - b_0)^2$
7	Bestimmung der Standardabweichung $s_{b0} = \frac{\sqrt{a} \cdot \sqrt{\sum_{n=1}^N x_n^2}}{\sqrt{N} \cdot \sqrt{N-1} \cdot \sqrt{N-2} \cdot s_x}$
8	Bestimmung des Konfidenzintervalls $b_0 - c_2 \cdot s_{b0} \leq \beta_1 < b_0 + c_1 \cdot s_{b0}$

Beispiel: Temperatursensor

Die Vorgehensweise zur Berechnung des Konfidenzbereichs des Regressionskoeffizienten β_0 wird wieder an dem Beispiel des Temperatursensors aus Tabelle 12.1 verdeutlicht. Es liegen wiederum $N = 11$ Stichprobenwerte vor, sodass zur Bestimmung der Konstante c_1 und c_2 eine t-Verteilung mit 9 Freiheitsgraden herangezogen wird. Für eine Konfidenzzahl von $\gamma = 0.95$ ergeben sich die Grenze c_1 und c_2 zu

$$c_1 = F^{-1}(1 - 0.975) = -2.2622 \quad (12.69)$$

und

$$c_2 = F^{-1}(1 - 0.025) = 2.2622 \quad (12.70)$$

Mit den Angaben errechnet sich der Konfidenzbereich des Regressionskoeffizienten β_0 mit dem geschätzten Regressionsparameter $b_0 = 2.6930$ und der Standardabweichung $s_x = 33.1662$ zu

$$2.5988 \leq \beta_0 < 2.7873 \quad (12.71)$$

In das Temperatur-Spannungs-Diagramm in Bild 11.6 wurde neben den Messwerten und der geschätzten Regressionsgeraden mit dem Regressionskoeffizienten b_0 noch zwei weitere Gerade eingezeichnet. Diese ergeben sich durch die Extremwerte des 95% - Konfidenzintervalls.

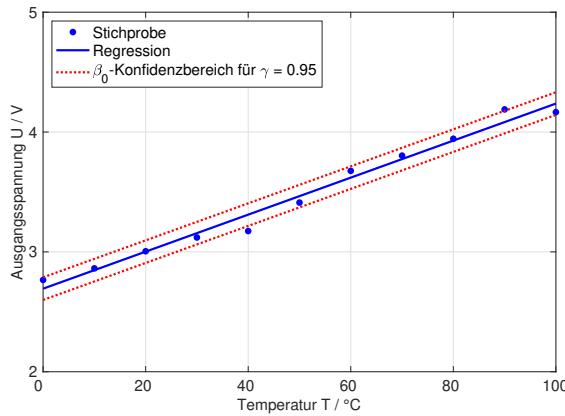


Bild 12.7: Grafische Darstellung der Stichprobe für den Zusammenhang zwischen Öltemperatur und Ausgangsspannung eines Temperatursensors mit unterschiedlichen Regressionsgeraden

Test des Regressionskoeffizienten β_0 auf Signifikanz

Zur Bewertung der Signifikanz des Regressionskoeffizienten β_0 wird die Nullhypothese geprüft, dass der Korrelationskoeffizient β_0 einem Wert $\beta_0 = 0$ entspricht. Trifft diese Hypothese zu, liegt bei dem untersuchten System kein Gleichanteil vor. Wird die Nullhypothese auf Basis der vorliegenden Stichprobe abgelehnt, kann davon ausgegangen werden, dass das untersuchte System einen Gleichanteil besitzt.

Damit ein über die Stichprobe geschätzter Regressionskoeffizient b_0 mit einer spezifizierten Wahrscheinlichkeit zu der t-Verteilung aus Gleichung (12.63) gehört, muss dieser in dem Intervall $\beta_{0C1} < b_0 \leq \beta_{0C2}$ liegen. Wird die Wahrscheinlichkeit dafür mit γ bezeichnet, gilt die Gleichung

$$P(\beta_{0C1} < b_0 \leq \beta_{0C2}) = \gamma = 1 - \alpha \quad (12.72)$$

Mit der Verteilung aus Gleichung (12.63) wird nach Gleichung (12.72) die Wahrscheinlichkeit γ , mit der die Variable t innerhalb des Intervalls $c_1 \dots c_2$ liegt, definiert als

$$\gamma = P(c_1 < t \leq c_2) = F(c_2) - F(c_1) \quad (12.73)$$

Bei Annahme eines symmetrischen Tests ergeben sich die Konstanten c_1 und c_2 aus den Bedingungen

$$c_1 = F^{-1}\left(\frac{\alpha}{2}\right) \quad (12.74)$$

und

$$c_2 = F^{-1}\left(1 - \frac{\alpha}{2}\right) \quad (12.75)$$

Durch Umformungen von Gleichung (12.63) und (12.73) ergibt sich ein Ausdruck für den Annahmebereich der Nullhypothese, nämlich dass der geschätzte Regressionskoeffizient b_0 mit einer spezifizierten Wahrscheinlichkeit γ zu der angenommenen t-Verteilung gehört.

$$\gamma = P(c_1 < t \leq c_2) = P(c_1 \cdot s_{b0} < b_0 \leq c_2 \cdot s_{b0}) \quad (12.76)$$

Alternativ kann, wie in Kapitel 6 gezeigt wird, eine Unterschreitungswahrscheinlichkeit p der Prüfgröße b_0 bestimmt werden und mit dem Signifikanzniveau α verglichen werden. Bei Hypothesentests mit beidseitigem Verwerfungsbereich $\beta_0 \neq 0$ müssen für die Annahme der Nullhypothese die Bedingungen

$$p = F(t) > \frac{\alpha}{2} \quad (12.77)$$

und

$$p = F(t) > \frac{\alpha}{2} \quad (12.78)$$

erfüllt werden. Damit lässt sich der Test mit der Hypothese $\beta_0 = 0$ und der Alternative $\beta_0 \neq 0$ in folgenden Prozessschritten zusammenfassen.

Tabelle 12.6: Test der Hypothese $\beta_0 = 0$ gegen $\beta_0 \neq 0$ für den Regressionskoeffizienten β_0 einer linearen Regression

Nr.	Prozessschritt	
1	Wahl eines Signifikanzniveaus α	
2	Bestimmung der zugehörigen Parameter c_1 und c_2 aus der inversen t-Verteilung mit $N - 2$ Freiheitsgraden $c_1 = F^{-1}\left(\frac{\alpha}{2}\right) \quad \text{und} \quad c_2 = F^{-1}(1 - \frac{\alpha}{2})$	
3	Berechnung der Mittelwerte der Stichprobe $\bar{x} = \frac{1}{N} \cdot \sum_{n=1}^N x_n \quad \text{und} \quad \bar{y} = \frac{1}{N} \cdot \sum_{n=1}^N y_n$	
4	Bestimmung der Standardabweichung und der Kovarianz der Stichprobe $s_x = \sqrt{\frac{1}{N-1} \cdot \sum_{n=1}^N (x_n - \bar{x})^2} \quad \text{und} \quad s_{xy} = \frac{1}{N-1} \cdot \left(\sum_{n=1}^N x_n \cdot y_n - N \cdot \bar{x} \cdot \bar{y} \right)$	
5	Bestimmung der Parameter b_0 und b_1 der Regressionsgeraden $b_1 = \frac{\sum_{n=1}^N x_n \cdot y_n - N \cdot \bar{x} \cdot \bar{y}}{\sum_{n=1}^N x_n^2 - N \cdot \bar{x}^2} = \frac{s_{xy}}{s_x^2} \quad \text{und} \quad b_0 = \bar{y} - b_1 \cdot \bar{x}$	
6	Berechnung der Fehlerquadratsumme $a = \sum_{n=1}^N r_n^2 = \sum_{n=1}^N (y_n - b_1 \cdot x_n - b_0)^2$	
7	Bestimmung der Standardabweichung s_{b1} $s_{b0} = \frac{\sqrt{a} \cdot \sqrt{\sum_{n=1}^N x_n^2}}{\sqrt{N} \cdot \sqrt{N-1} \cdot \sqrt{N-2} \cdot s_x}$	
8	Berechnung der Grenzen des Annahmebereiches $\beta_{0C1} = c_1 \cdot s_{b0} \quad \text{und} \quad \beta_{0C2} = c_2 \cdot s_{b0}$	
9	Bestimmung des Annahmebereichs $\beta_{0C1} < b_1 \leq \beta_{0C2}$	Berechnung des p-Wertes mit der t-Verteilung mit $N - 2$ Freiheitsgraden $p = F\left(\frac{b_0}{s_{b0}}\right)$
10	Für $\beta_{0C1} \leq b_0 < \beta_{0C2}$ wird die Hypothese angenommen, für $b_0 \leq \beta_{0C1}$ oder $b_0 > \beta_{0C2}$ wird die Hypothese verworfen	Für $\alpha/2 \leq p < 1 - \alpha/2$ wird die Hypothese angenommen, für $p < \alpha/2$ und $p \geq 1 - \alpha/2$ wird die Hypothese verworfen

Beispiel: Temperatursensor

Für das Beispiel des Temperatursensors aus Tabelle 11.1 wird die Nullhypothese $\beta_0 = 0$ gegen die Alternativhypothese $\beta_0 \neq 0$ getestet. Es liegen $N = 11$ Stichprobenwerte vor, sodass zur Bestimmung der Konstanten c_1 und c_2 eine t-Verteilung mit 9 Freiheitsgraden herangezogen wird. Für ein Signifikanzniveau $\alpha = 0.05$ ergeben sich die Grenzen c_1 und c_2 zu

$$c_1 = F^{-1}(1 - 0.975) = -2.2622 \quad (12.79)$$

und

$$c_2 = F^{-1}(1 - 0.025) = 2.2622 \quad (12.80)$$

Mit den Angaben errechnet sich der Annahmebereich der Nullhypothese zu

$$-0.0943 < b_0 \leq 0.0943 \quad (12.81)$$

Der geschätzte Wert des Regressionskoeffizienten liegt mit $b_0 = 2.6930$ außerhalb des in Gleichung (12.59) berechneten Annahmebereich. Die Nullhypothese wird somit auf Grundlage der vorliegenden Stichprobe verworfen, der Regressionskoeffizient β_0 ist ungleich 0 und damit signifikant.

Alternativ kann zur Bewertung des Hypothesentests auch der p-Wert herangezogen werden. Dieser berechnet sich für die vorliegenden Hypothesen durch

$$p = F\left(\frac{b_0}{s_{b0}}\right) \quad (12.82)$$

Der Wert $p = 1$ liegt deutlich über der Grenze von $1 - \alpha/2 = 0.975$. Die Nullhypothese muss daher verworfen werden, womit die Signifikanz des Regressionskoeffizienten β_0 aufgezeigt wird.

Eine weitere Möglichkeit zur Überprüfung der Signifikanz eines Regressionskoeffizienten ist dessen Konfidenzbereich. Für das Beispiel aus Tabelle 11.1 wurde der Konfidenzbereich des Regressionskoeffizienten β_0 bestimmt zu

$$2.5988 \leq \beta_0 < 2.7873 \quad (12.83)$$

Dieses Konfidenzintervall schließt die Zahl 0 nicht mit ein, sodass der Regressionskoeffizient auch nach diesem Kriterium als signifikant angenommen werden muss.

12.1.4 Konfidenzintervall für den Erwartungswert

Mit der Regressionsgleichung wird die Zielgröße y_0 als Funktion der festen Eingangsgrößen x_0 geschätzt.

$$y(x_0) = b_0 + b_1 \cdot x_0 \quad (12.84)$$

Dabei werden b_0 und b_1 auf Basis des vorliegenden Datensatzes geschätzt, sie sind damit selber Zufallsvariablen. Für den Erwartungswert gilt:

$$\mu_y(x_0) = E(y_0) = E(b_0 + b_1 \cdot x_0) = E(b_0) + E(b_1 \cdot x_0) = \beta_0 + \beta_1 \cdot x_0 \quad (12.85)$$

Die Regression ist erwartungstreu, weil die Regressionskoeffizienten β_0 und β_1 erwartungstreu geschätzt werden. Bei der Berechnung des Konfidenzintervalls für den Erwartungswert $\mu_y(x_0)$ wird eine Variable benötigt, deren Verteilung bekannt ist und in der der Erwartungswert sowie der geschätzte Wert $y(x_0)$ vorkommt. Der Schätzwert berechnet sich zu

$$\bar{y}(x_0) = b_1 \cdot x_0 + b_0 = b_1 \cdot (x_0 - \bar{x}) + \bar{y} \quad (12.86)$$

Damit ist die Größe

$$z = \frac{y(x_0) - \mu_y(x_0)}{\sigma_{y0}} = \frac{b_1 \cdot (x_0 - \bar{x}) + \bar{y} - \mu_y(x_0)}{\sqrt{\sigma_{b_1}^2 \cdot (x_0 - \bar{x})^2 + \sigma_{\bar{y}}^2}} = \frac{b_1 \cdot (x_0 - \bar{x}) + \bar{y} - \mu_y(x_0)}{\sigma \cdot \sqrt{\frac{(x_0 - \bar{x})^2}{(N-1) \cdot s_x^2} + \frac{1}{N}}} \quad (12.87)$$

standard-normalverteilt. Die Varianz σ^2 wird mit Gleichung (12.37) geschätzt. Die Schätzung weist $N - 2$ Freiheitsgrade auf. Damit besitzt die Größe

$$t = \frac{b_1 \cdot (x_0 - \bar{x}) + \bar{y} - \mu_y(x_0)}{\sqrt{\frac{(x_0 - \bar{x})^2}{(N-1) \cdot s_x^2} + \frac{1}{N}}} = \frac{y(x_0) - \mu_y(x_0)}{s_{y0}} \quad (12.88)$$

eine t-Verteilung mit $N - 2$ Freiheitsgraden auf, und die Standardabweichung s_{y0} errechnet sich zu

$$s_{y0} = \sqrt{\frac{(x_0 - \bar{x})^2}{(N-1) \cdot s_x^2} + \frac{1}{N}} \cdot \sqrt{\frac{a}{N-2}} \quad (12.89)$$

Die Zufallsvariable t wird zur Bestimmung des Konfidenzbereichs für den Erwartungswert $\mu_y(x_0)$ verwendet. Er berechnet sich aus der Wahrscheinlichkeit

$$P(c_1 < t \leq c_2) = F(c_2) - F(c_1) = \gamma \quad (12.90)$$

Durch die Symmetrie des Konfidenzbereichs ergeben sich die Konstanten c_1 und c_2 zu

$$c_1 = F^{-1}\left(\frac{1-\gamma}{2}\right) \quad (12.91)$$

und

$$c_2 = F^{-1}\left(\frac{1-\gamma}{2}\right) \quad (12.92)$$

Durch Umformungen ergibt sich ein Ausdruck für den Konfidenzbereich des Mittelwertes.

$$\gamma = P(c_1 < t \leq c_2) = P(y(x_0) - c_2 \cdot s_{y0} \leq \mu_y(x_0) < y(x_0) - c_1 \cdot s_{y0}) \quad (12.93)$$

Das Verfahren wird in Tabelle 12.7 beschrieben.

Tabelle 12.7: Vorgehen zur Bestimmung des Konfidenzbereichs für den Mittelwert einer Regressionsgerade

Nr.	Prozessschritt
1	Wahl einer Konfidenzzahl γ
2	Bestimmung der zugehörigen Parameter c_1 und c_2 aus der inversen t-Verteilung mit $N - 2$ Freiheitsgraden $c_1 = F^{-1}\left(\frac{1-\gamma}{2}\right) \quad \text{und} \quad c_2 = F^{-1}\left(\frac{1-\gamma}{2}\right)$
3	Berechnung der Mittelwerte der Stichprobe $\bar{x} = \frac{1}{N} \cdot \sum_{n=1}^N x_n \quad \text{und} \quad \bar{y} = \frac{1}{N} \cdot \sum_{n=1}^N y_n$
4	Bestimmung der Standardabweichung und der Kovarianz der Stichprobe $s_x = \sqrt{\frac{1}{N-1} \cdot \sum_{n=1}^N (x_n - \bar{x})^2} \quad \text{und} \quad s_{xy} = \frac{1}{N-1} \cdot \left(\sum_{n=1}^N x_n \cdot y_n - N \cdot \bar{x} \cdot \bar{y} \right)$
5	Bestimmung der Parameter b_0 und b_1 der Regressionsgeraden $b_1 = \frac{\sum_{n=1}^N x_n \cdot y_n - N \cdot \bar{x} \cdot \bar{y}}{\sum_{n=1}^N x_n^2 - N \cdot \bar{x}^2} = \frac{s_{xy}}{s_x^2} \quad \text{und} \quad b_0 = \bar{y} - b_1 \cdot \bar{x}$
6	Berechnung der Fehlerquadratsumme $a = \sum_{n=1}^N r_n^2 = \sum_{n=1}^N (y_n - b_1 \cdot x_n - b_0)^2$
7	Bestimmung der Standardabweichung s_{y0} $s_{y0} = \sqrt{\frac{(x_0 - \bar{x})^2}{(N-1) \cdot s_x^2} + \frac{1}{N} \cdot \sqrt{\frac{a}{N-2}}}$
8	Bestimmung des Konfidenzintervalls $y(x_0) - c_2 \cdot s_{y0} \leq \mu_y(x_0) < y(x_0) + c_1 \cdot s_{y0}$

Der Konfidenzbereich des Mittelwertes ist von der Zufallsvariablen x abhängig und erreicht seine minimale Länge an der Stelle $x = \bar{x}$. Mit steigendem Abstand x von dem Mittelwert steigt die Länge des Konvergenzintervalls an.

Beispiel: Temperatursensor

Bild 12.8 zeigt den Konvergenzbereich für das Beispiel des Öltemperatursensors aus Tabelle 12.1.

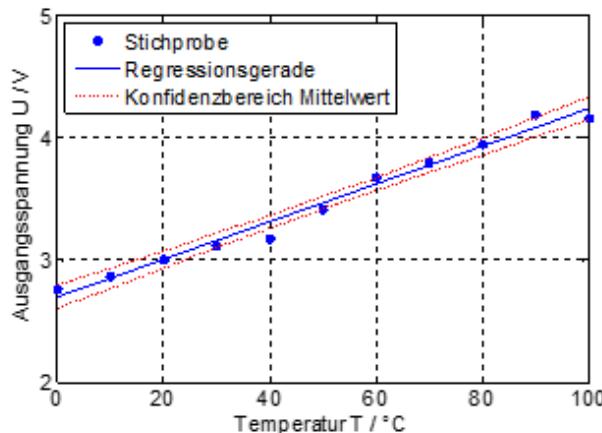


Bild 12.8: Grafische Darstellung des Konvergenzintervalls für den Mittelwert für das Beispiel aus Tabelle 12.1

12.1.5 Prognoseintervall für zukünftige Werte

Um zukünftige Werte einer Stichprobe vorhersagen zu können, ist es erforderlich, das Prognoseintervall für die Reaktion y_{N+1} an einer Stelle x_0 abschätzen zu können. Die Lage eines zukünftigen Wertes ergibt sich aus dem Mittelwert und dem überlagerten Messfehler.

$$y_{N+1}(x_0) = b_1 \cdot x_0 + b_0 + e = b_1 \cdot (x_0 - \bar{x}) + \bar{y} + e \quad (12.94)$$

Der Messfehler e ist mittelwertsfrei, sodass sich der Erwartungswert von

$$E(y_{N+1}(x_0)) = E(b_1 \cdot x_0 + b_0 + e) = E(b_0) + E(b_1 \cdot x_0) + E(e) = \beta_0 + \beta_1 \cdot x_0 \quad (12.95)$$

ergibt. Die Varianz der Größe berechnet sich zu

$$\sigma_{y_{N+1}}^2 = \sigma_{b_1}^2 \cdot (x_0 - \bar{x})^2 + \sigma_y^2 + \sigma^2 = \left(\frac{(x_0 - \bar{x})^2}{(N-1) \cdot s_x^2} + \frac{1}{N} + 1 \right) \cdot \sigma^2 \quad (12.96)$$

Damit besitzt die Zufallsvariable

$$z = \frac{y_{N+1}(x_0) - (b_1 \cdot (x_0 - \bar{x}) + \bar{y})}{\sigma \cdot \sqrt{1 + \frac{(x_0 - \bar{x})^2}{(N-1) \cdot s_x^2} + \frac{1}{N}}} \quad (12.97)$$

eine Standardnormalverteilung. Die Varianz σ^2 wird mit Gleichung (10.37) geschätzt. Die Schätzung weist $N - 2$ Freiheitsgrade auf. Damit besitzt die Größe

$$t = \frac{y_{N+1}(x_0) - y_0}{\sqrt{\frac{(x_0 - \bar{x})^2}{(N-1) \cdot s_x^2} + \frac{N+1}{N}} \cdot \sqrt{\frac{a}{N-2}}} \quad (12.98)$$

eine t-Verteilung mit $N - 2$ Freiheitsgraden auf, und die Standardabweichung $s_{y_{N+1}}$ errechnet sich zu

$$s_{y_{N+1}} = \sqrt{\frac{(x_0 - \bar{x})^2}{(N-1) \cdot s_x^2} + \frac{N+1}{N}} \cdot \sqrt{\frac{a}{N-2}} \quad (12.99)$$

Der Zähler der Variable stellt den Abstand des Schätzwertes $y_{N+1}(x_0)$ an der Stelle x_0 zu der Regressionsgeraden dar, sodass diese Zufallsvariable zur Bestimmung des Prognoseintervalls verwendet werden kann.

$$P(c_1 < t \leq c_2) = F(c_2) - F(c_1) = \gamma \quad (12.100)$$

Durch die Symmetrie des Prognosebereichs ergeben sich die Konstanten c_1 und c_2 zu

$$c_1 = F^{-1}\left(\frac{1-\gamma}{2}\right) \quad (12.101)$$

und

$$c_2 = F^{-1}\left(\frac{1+\gamma}{2}\right) \quad (12.102)$$

Durch Umformungen ergibt sich ein Ausdruck für den Prognosebereich zukünftiger Stichprobenwerte

$$\gamma = P(c_1 < t \leq c_2) = P(y(x_0) + c_1 \cdot s_{y_{N+1}} < y_{N+1}(x_0) \leq y(x_0) + c_2 \cdot s_{y_{N+1}}) \quad (12.103)$$

mit

$$y(x_0) = (b_1 \cdot (x_0 - \bar{x}) + \bar{y}) \quad (12.104)$$

Der Prognosebereich ist von der Zufallsvariablen x abhängig und erreicht seine minimale Länge an der Stelle $x = \bar{x}$. Mit steigendem Abstand x von dem Mittelwert steigt die Länge des Prognoseintervalls an.

Ein Vergleich von Konfidenz- und Prognoseintervall zeigt, dass sie sich formell nur leicht unterscheiden. Für den Konfidenzbereich des Mittelwertes existiert ein Summand $1/N$ und bei der Berechnung des Prognoseintervalls für zukünftige Werte ein Summand $(N+1)/N$. Da $N > 1$ ist, ist der Konfidenzbereich des Mittelwertes erwartungsgemäß kleiner als der Prognosebereich für zukünftige Werte.

Das Verfahren ist in Tabelle 12.8 zusammengefasst.

Tabelle 12.8: Vorgehen zur Bestimmung des Prognosebereiches für zukünftige Werte einer Regressionsgerade

Nr.	Prozessschritt
1	Wahl einer Konfidenzzahl γ
2	Bestimmung der zugehörigen Parameter c_1 und c_2 aus der inversen t-Verteilung mit $N - 2$ Freiheitsgraden $c_1 = F^{-1}\left(\frac{1-\gamma}{2}\right) \quad \text{und} \quad c_2 = F^{-1}\left(\frac{1-\gamma}{2}\right)$
3	Berechnung der Mittelwerte der Stichprobe $\bar{x} = \frac{1}{N} \cdot \sum_{n=1}^N x_n \quad \text{und} \quad \bar{y} = \frac{1}{N} \cdot \sum_{n=1}^N y_n$
4	Bestimmung der Standardabweichung und der Kovarianz der Stichprobe $s_x = \sqrt{\frac{1}{N-1} \cdot \sum_{n=1}^N (x_n - \bar{x})^2} \quad \text{und} \quad s_{xy} = \frac{1}{N-1} \cdot \left(\sum_{n=1}^N x_n \cdot y_n - N \cdot \bar{x} \cdot \bar{y} \right)$
5	Bestimmung der Parameter b_0 und b_1 der Regressionsgeraden $b_1 = \frac{\sum_{n=1}^N x_n \cdot y_n - N \cdot \bar{x} \cdot \bar{y}}{\sum_{n=1}^N x_n^2 - N \cdot \bar{x}^2} = \frac{s_{xy}}{s_x^2} \quad \text{und} \quad b_0 = \bar{y} - b_1 \cdot \bar{x}$
6	Berechnung der Fehlerquadratsumme $a = \sum_{n=1}^N r_n^2 = \sum_{n=1}^N (y_n - b_1 \cdot x_n - b_0)^2$
7	Bestimmung der Standardabweichung $s_{y_{N+1}} = \sqrt{\frac{(x_0 - \bar{x})^2}{(N-1) \cdot s_x^2} + \frac{N+1}{N}} \cdot \sqrt{\frac{a}{N-2}}$
8	Bestimmung des Prognoseintervalls $y(x_0) - c_1 \cdot s_{y_{N+1}} \leq \mu_{y_{N+1}}(x_0) \leq y(x_0) + c_2 \cdot s_{y_{N+1}}$

Beispiel: Temperatursensor

Bild 12.9 zeigt die beiden Intervalle für das Beispiel des Öltemperatursensors aus Tabelle 12.1.

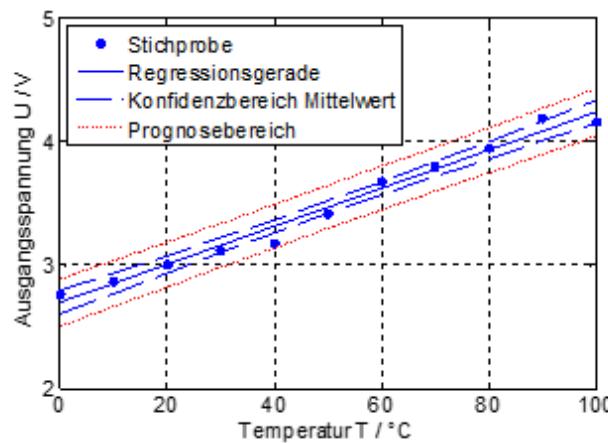


Bild 12.9: Grafischer Vergleich des Konfidenzbereichs für den Mittelwert und des Prognosebereiches für zukünftige Werte bei einer Konfidenzzahl $\gamma = 0.95$

12.1.6 Lineare Regression zweidimensionaler Datensätze mit MATLAB

MATLAB unterstützt die Regression zwei- und mehrdimensionaler Datensätze mit einem umfangreichen Befehlssatz. Die wichtigsten Befehle sind in Tabelle 12.9 zusammengefasst.

Tabelle 12.9: MATLAB-Befehle zur Berechnung und Bewertung von Regressionsfunktionen zweidimensionaler Datensätze

Befehl	Beschreibung
$[B,s] = \text{polyfit}(X, Y, M)$	Polyfit berechnet zu einem Eingangsvektor X und einem Ausgangsvektor Y ein Regressionspolynom M-ter Ordnung, die Koeffizienten werden in dem Vektor B dargestellt. Außerdem wird eine Struktur s erzeugt, die zur Berechnung von Konfidenzbereich des Mittelwertes und des Prognosebereiches zukünftiger Stichprobenwerte erforderlich ist
$[Y, DEL] = \text{polyval}(B, X, s)$	Polyval berechnet die Regressionswerte des Polynoms mit den Koeffizienten P an den Stelle X, außerdem die Standardabweichung des Konfidenzintervalls des Mittelwertes, Basis sind die über polyfit berechneten Werte B und s
$[Y, DEL] = \text{polyconf}(B, X, s)$	Polyconf hat eine ähnliche Funktion wie polyval, erlaubt aber durch die Angabe der Schlüsselwörter observation und curve auch die Berechnung zukünftiger Prognosewerte. Mit dem Schlüsselwort observation lassen sich die Prognoseintervalle und mit dem Schlüsselwort curve die Konfidenzintervalle berechnen. Als default Einstellung in MATLAB ist observation mit $\alpha = 5\%$ definiert.
$stats = \text{regstats}(Y, X, model)$	Regstats liefert eine Datenstruktur zur Bewertung der Regression. Mit den Werten kann eine Signifikanzbewertung der Terme durchgeführt werden, und es können Konfidenzbereiche der Residuen angegeben werden. Außerdem bietet sich die Möglichkeit der Bewertung über Bestimmtheitsmaße
$[B, Bint] = \text{regress}(Y, X)$	Regress berechnet die Regressionskoeffizienten und deren Konfidenzintervalle

Für das Beispiel des Temperatursensors ergibt sich die folgende Befehlssequenz:

```
1 % Definition der Stichprobe: Spannungsvektor gekürzt
2 T = 0:10:100;
3 U = [2.7660 2.8626 ... 4.1887 4.1659];
4
5 % Regressionsfunktion der Ordnung 1 mit Konfidenz- und Prognosebereich
6 [P,s] = polyfit(T,U,1);
7 [Ureg, delmean] = polyconf(P,T,s, 'alpha', 0.05, 'predopt', 'curve');
8 [Ureg, delprog] = polyconf(P,T,s, 'alpha', 0.05, 'predopt', 'observation');
9
10 % MATLAB-Funktionen zur Bewertung der Signifikanz
11 stats = regstats(U,T, 'linear')
12 [B,Bint] = regress(U',[ones(length(T),1) T'])
```

Zu beachten sind die unterschiedlichen Zusammensetzungen der Matrix von Eingangsgrößen bei den Befehlen `regstats` und `regress`, dazu sind in der MATLAB-Hilfe detailliertere Information zu finden.

Die Ergebnisse für die Regression entsprechen den analytisch berechneten Werten. Einige statistische Informationen des `regstats`-Befehls werden in den folgenden Abschnitten noch aufgegriffen.

12.2 Regression mit Polynomen

In Abschnitt 11.1 wird eine zweidimensionale Stichprobe über eine Gerade approximiert. Mithilfe statistischer Verfahren werden für die Regressionsparameter und den funktionalen Zusammenhang selbst Konfidenzintervalle bestimmt.

Bei technischen Aufgabenstellungen ist es insbesondere bei Fehlerabschätzungen erforderlich, funktionale Zusammenhänge höherer Ordnung über Polynome beschreiben zu können. Dieser Ansatz führt zur Regression mit Polynomen.

Aus Gründen der Übersichtlichkeit werden für die Regression mit Polynomen in diesem Abschnitt keine statistischen Bewertungen hergeleitet. Zur Bewertung werden verfügbare Verfahren aus Software-Paketen verwendet. Eine ausführliche Herleitung ist in Kapitel 12 zu finden.

12.3 Berechnung und Bewertung der Regressionskoeffizienten

Analog zum Vorgehen zur Bestimmung einer Regressionsgeraden nach dem Prinzip des kleinsten Fehlerquadrates kann auch ein Polynom höherer Ordnung als Regressionsfunktion verwendet werden. Im allgemeinen Fall ergibt sich ein Polynom M-ter Ordnung der Form

$$y(x_0) = b_0 + b_1 \cdot x + \dots + b_M \cdot x^M \quad (12.105)$$

Zur Bestimmung der Koeffizienten b_m wird wieder gefordert, dass die Summe der quadratischen Fehler ein Minimum aufweist. Diese Forderung führt zu den $M + 1$ Gleichungen

$$\frac{\partial a}{\partial b_0} = 0, \frac{\partial a}{\partial b_1} = 0, \dots, \frac{\partial a}{\partial b_M} = 0 \quad (12.106)$$

Zur Auswertung dieser Forderung stehen Software-Pakete zur Verfügung, die neben der Bestimmung der Koeffizienten b_m des Regressionspolynoms eine Berechnung des Konfidenzintervalls erlauben. Bild 12.10 zeigt die Regressionsfunktion für das Beispiel aus Tabelle 12.1 mit der Ordnung $M = 1, M = 2, M = 3$, und $M = 6$.

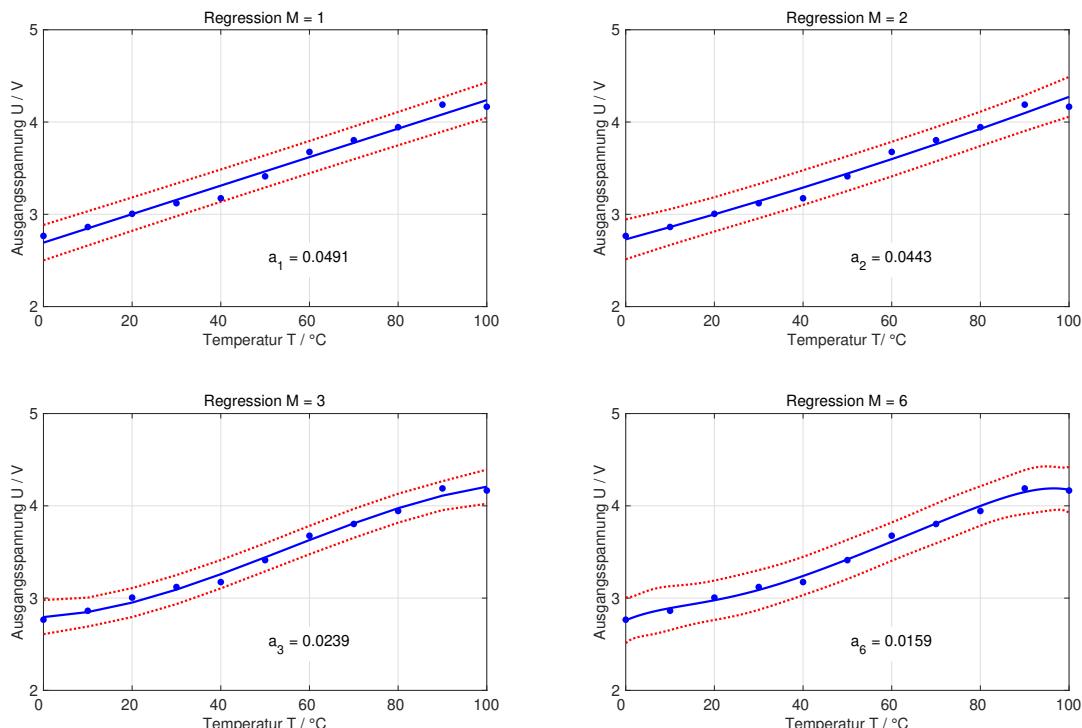


Bild 12.10: Grafische Darstellung des Konvergenzintervalls für das Beispiel aus Tabelle 11.1 und Regressionspolynome der Ordnung $M = 1, M = 2, M = 3$ und $M = 6$, Vorhersagebereich mit $\gamma = 95\%$

In dem Beispiel sinkt die Summe der Fehlerquadrate mit steigender Ordnung M des Regressionspolynoms von $a_1 = 0.0491$ auf $a_6 = 0.0159$ ab. Generell wird die Summe der Fehlerquadrate mit steigender Ordnung des Polynoms immer sinken. Trotzdem muss der Ansatz eines Regressionspolynoms höherer Ordnung nicht unbedingt zielführend sein. Ergibt sich aus dem physikalischen Hintergrund ein linearer Zusammenhang, ist eine lineare Regressionsfunktion, die den physikalischen Zusammenhang beschreibt, sinnvoller als ein Regressionspolynom höherer Ordnung, das lediglich die Messfehler gut approximiert. Über die Güte einer Regression entscheidet deshalb neben der Summe der Fehlerquadrate das Ziel, das mit der Regression verfolgt wird.

Sollen die Daten über den bekannten Datenbereich extrapoliert werden, ergibt sich ein weiterer Grund für eine Regression mit einer geringen Ordnung. Dazu zeigt Bild 12.12 für Regressionsfunktionen der Ordnung $M = 1$ und $M = 3$ den Konfidenzbereich bei Extrapolation der Daten.

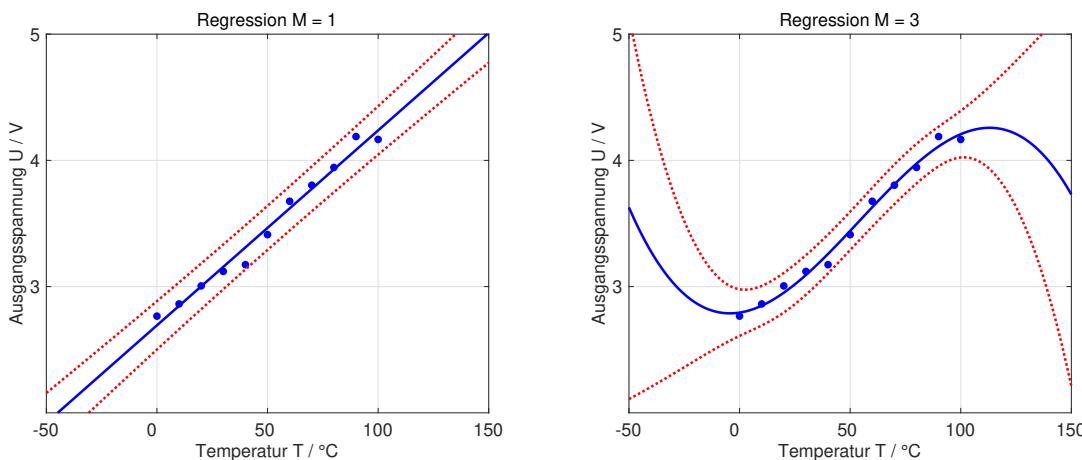


Bild 12.11: Grafische Darstellung des Vorhersagebereichs mit $\gamma = 95\%$ für das Beispiel aus Tabelle 11.1 und Regressionspolynome der Ordnung $M = 1$ und $M = 3$ bei Extrapolation

Es ist deutlich zu erkennen, dass das Konfidenzintervall bei Regressionsfunktionen höherer Ordnung viel stärker wächst als bei Regressionen niedriger Ordnung. Daraus lässt sich die Regel ableiten, die Regression mit einer möglich geringen Ordnung zu realisieren und auf eine Extrapolation der Daten soweit wie möglich zu vermeiden.

Die Berechnung in MATLAB entspricht dem in Abschnitt 12.1.6 beschriebenen Vorgehen, es wird lediglich die Ordnung M des Polynoms verändert.

12.3.1 Signifikanz der einzelnen Regressionsterme bei Regressionsfunktionen höherer Ordnung

In dem Beispiel des Temperatursensors bleibt die Quadratsumme der Residuen bei einer Regression mit einem Polynom zweiter Ordnung ($a_2 = 0.0443$) fast genauso groß wie bei einer Regression mit einer Geraden ($a_1 = 0.0491$). Wird als Regressionsfunktion ein Polynom dritter Ordnung verwendet, wird die Quadratsumme der Residuen ($a_3 = 0.0239$) praktisch halbiert. Offensichtlich scheint der quadratische Term weniger wichtig oder weniger signifikant zu sein als der kubische Term.

Die Vermutung kann anhand der einzelnen Regressionskoeffizienten verifiziert werden. Die einzelnen Regressionskoeffizienten b_m werden durch Minimierung des quadratischen Fehlers für die vorliegende Stichprobe gewonnen. Wegen des endlichen Stichprobenumfangs weisen die Regressionsparameter b_m eine Unsicherheit auf, die durch die ihre Standardabweichung s_{bm} ausgedrückt wird. Ist die Streuung s_{bm} gegenüber dem Regressionskoeffizienten b_m klein, ist der Regressionskoeffizient b_m mit hoher Wahrscheinlichkeit von null verschieden.

Diese Entscheidung kann als Hypothesentest formuliert werden. Es wird die Nullhypothese getestet, dass der Regressionskoeffizient b_m den Wert null aufweist. In Tabelle 12.10 wird dieser Test analog zu den Darstellungen in den Abschnitten 12.1.2 und 12.1.3 auf eine t-Verteilung mit $N - 2$ Freiheitsgraden und dem Signifikanzniveau α zurückgeführt. Ist die Wahrscheinlichkeit P kleiner als das Signifikanzniveau α , wird davon ausgegangen, dass der Regressionskoeffizient signifikant ist. Andernfalls ist der Koeffizient für die Regression nicht von Bedeutung, er kann zu null gesetzt werden. Die Frage der Signifikanz eines Regressionsparameters wird bei vielen Software-Paketen als Koeffiziententabelle zusammengefasst. Tabelle 12.10 stellt die Koeffiziententabelle für das Beispiel des Öltemperatursensors dar.

Tabelle 12.10: Bewertung der Signifikanz von Regressionskoeffizienten für das Beispiel aus Tabelle 12.1 mit einem Stichprobenumfang von $n = 11$ über den t-Test

Name	Regressionskoeffizient b_m	Standardabweichung s_{bm}	t-Wert	p-Wert	Signifikanz
b_0	2.7940	$5.1 \cdot 10^{-2}$	53.7917	$2.0 \cdot 10^{-10}$	ja
b_1	0.0027	$4.7 \cdot 10^{-3}$	0.5673	0.5882	nein
b_2	$2.96 \cdot 10^{-4}$	$1.13 \cdot 10^{-4}$	2.6147	0.0347	ja
b_3	$-1.8 \cdot 10^{-6}$	$7.44 \cdot 10^{-7}$	-2.4437	0.0445	ja

Die Analyse zeigt, dass alle Regressionskoeffizienten außer dem Koeffizienten b_1 unterhalb der Grenze von 5 % liegen und damit signifikant sind. Alternativ kann der Konfidenzbereich der Regressionskoeffizienten untersucht werden. Liegt die Zahl null innerhalb des Konfidenzintervalls, ist der betreffende Regressionskoeffizient für die zu untersuchende Aufgabenstellung nicht signifikant.

Die für eine Bewertung der Signifikanz erforderlichen Daten ergeben sich aus den MATLAB-Befehlen regstats beziehungsweise regress. Die Befehle sind in Abschnitt 11.1.6 beschrieben. Zur Bewertung der Signifikanz wird die strukturierte Variable stats verwendet.

```

1 % MATLAB-Funktionen zur Bewertung der Signifikanz
2 stats = regstats(U,T, 'quadratic')
3 stats.tstat

```

Tabelle 12.11: Bewertung der Signifikanz von Regressionskoeffizienten für das Beispiel aus Tabelle 11.1 mit einem Stichprobenumfang von $N = 11$ über den Konfidenzbereich der Regressionskoeffizienten

Name	Konfidenzintervall			Signifikanz
	Minimum	Mitte	Maximum	
b_0	2.6711	2.7940	2.9168	ja
b_1	-0.0085	0.0027	0.0139	nein
b_2	$0.28 \cdot 10^{-4}$	$2.96 \cdot 10^{-4}$	$5.64 \cdot 10^{-4}$	ja
b_3	$-3.6 \cdot 10^{-6}$	$-1.8 \cdot 10^{-6}$	$-0.1 \cdot 10^{-6}$	ja

Die Analyse der Konfidenzintervalle bestätigt den Hypothesentest. Zur Reduzierung der Komplexität kann das Regressionsmodell der Ausgangsspannung des Öltemperatursensors deshalb vereinfacht werden, indem die linearen Terme in T aus dem Modell eliminiert werden. Aus der ursprünglichen Regressionsfunktion dritter Ordnung

$$U(T) = b_0 + b_1 \cdot T + b_2 \cdot T^2 + b_3 \cdot T^3 \quad (12.107)$$

wird damit die reduzierte Regressionsfunktion mit

$$U(T) = b_0 + b_2 \cdot T^2 + b_3 \cdot T^3 \quad (12.108)$$

Die neuen Regressionskoeffizienten und ihre Konfidenzintervalle sind in Tabelle 12.12 zusammengefasst.

Tabelle 12.12: Bewertung der Signifikanz von Regressionskoeffizienten für das Beispiel aus Tabelle 12.1 mit einem Stichprobenumfang von $N = 11$ über den Konfidenzbereich der Regressionskoeffizienten

Name	Konfidenzintervall			Signifikanz
	Minimum	Mitte	Maximum	
b_0	2.74	2.81	2.88	ja
b_1	$2.87 \cdot 10^{-4}$	$3.58 \cdot 10^{-4}$	$4.28 \cdot 10^{-4}$	ja
b_2	$-1.91 \cdot 10^{-6}$	$-2.19 \cdot 10^{-6}$	$-1.48 \cdot 10^{-6}$	ja

Die Koeffizienten ändern sich bei Weglassen des linearen Regressionsterms. Die Analyse zeigt, dass nach dem Streichen des linearen Terms alle übrigen Regressionsterme signifikant sind, was durch einen Hypothesentest bestätigt werden könnte. Die Analysen für die Signifikanz von Regressionskoeffizienten werden wie auch die Bestimmung der Regressionskoeffizienten selbst mit entsprechenden Software-Paketen durchgeführt. Die Bewertung basiert auf dem in Abschnitt 11.1 dargestellten Verfahren. Das Verfahren zur Reduktion der Regressionsfunktion auf signifikante Terme ist in Bild 12.12 dargestellt.

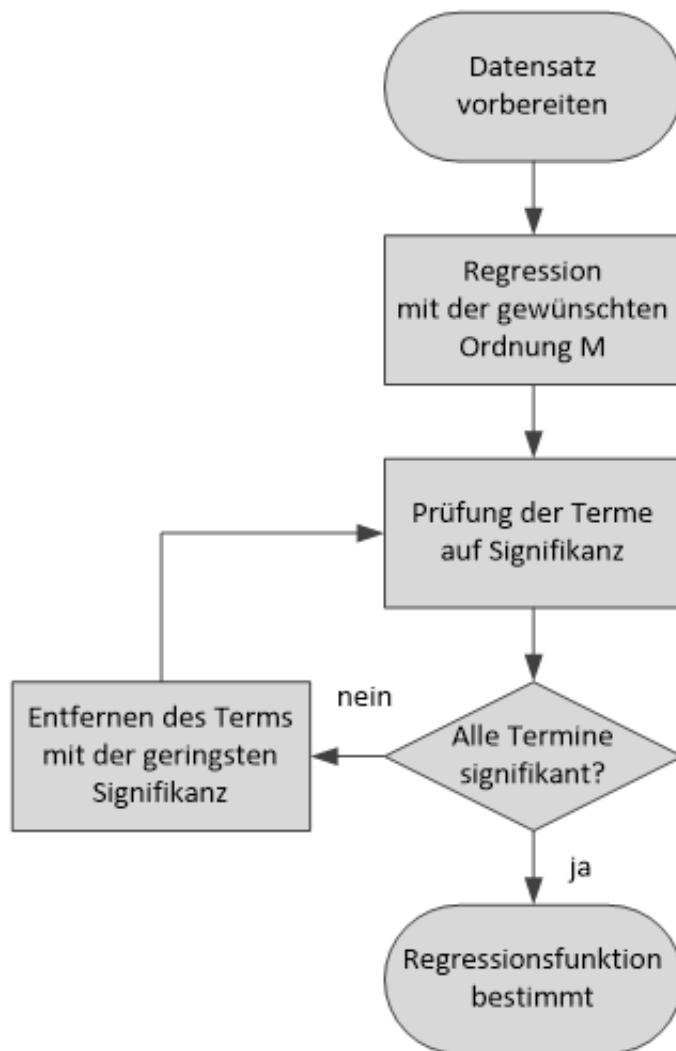


Bild 12.12: Ablaufdiagramm zur Reduktion der Regressionsfunktion auf signifikante Terme

Zunächst wird eine Regressionsfunktion mit der gewünschten Ordnung M erstellt. Mithilfe des beschriebenen t-Tests werden die einzelnen Terme hinsichtlich ihrer Signifikanz bewertet. Der Term mit dem höchsten p-Wert ist am wenigsten signifikant, er wird eliminiert. Nach der Eliminierung dieses einzelnen

Terms werden alle verbleibenden Terme erneut auf Signifikanz geprüft. Das Verfahren wird fortgesetzt, bis alle Terme signifikant sind, also einen p-Wert < 5% aufweisen.

12.4 Bewertung von Regressionen zweidimensionaler Datensätze

Nach Anpassung der Parameter muss das Ergebnis überprüft und bewertet werden. Dazu werden verschiedene Fragestellungen betrachtet:

- Wie gut ist die berechnete Regression?
- Wie verhält sich die durch die Regression nicht erklärte Reststreuung? Welche Verteilung hat sie, existieren Ausreißer? Welche Besonderheiten weist sie auf?
- Entsprechen die Ergebnisse den physikalischen Vorstellungen, die über den Prozess bekannt sind?

Diese Fragestellungen werden in den folgenden Abschnitten diskutiert. Dabei wird zunächst auf die Frage der Kenngrößen für die Güte einer Regression eingegangen.

12.4.1 Bewertung der Regressionsgüte über den Root-Mean-Square-Error

Die Frage nach der Güte einer Regression kann über die Summe der quadratischen Fehler beantwortet werden. In dem Beispiel des Öltemperatursensors zeigt sich, dass mit zunehmender Ordnung M die Summe der quadratischen Fehler abnimmt. Die Summe der quadratischen Fehler ist demnach ein Maß für die Güte der Regression. Um die Einheit der Zielgröße y zu erhalten, wird aus der Summe der quadratischen Fehler die Wurzel gezogen.

Mit steigender Anzahl N von Messpunkten steigt die Summe der quadratischen Fehler stetig, obwohl die Regression aufgrund der größeren Information immer besser wird. Um diesen Effekt zu kompensieren, muss die Wurzel der Summe von quadratischen Fehlern normiert werden. Durch diese Normierung ergibt sich ein Schätzwert für die Standardabweichung s_R der Residuen.

Liegt eine Regression der Ordnung M vor, müssen $M + 1$ Koeffizienten bestimmt werden. Weist die dazu vorhandene Stichprobe einen Umfang N auf, so muss zur eindeutigen Bestimmung der Koeffizienten die Bedingung

$$N \geq M + 1 \quad (12.109)$$

erfüllt sein. Die Differenz

$$\nu = N - (M + 1) = N - M - 1 \quad (12.110)$$

beschreibt die Anzahl von Stichproben, die zur Absicherung des Regressionsergebnisses verwendet werden. Für eine erwartungstreue Schätzung der Standardabweichung s_R muss eine Normierung mit der Anzahl von Freiheitsgraden ν erfolgen. Damit ergibt sich die Standardabweichung der Residuen zu

$$s_R = \sqrt{\frac{1}{N - M - 1} \cdot \sum_{n=1}^N (y_n - \bar{y}(x_n))^2} \quad (12.111)$$

Sie wird auch als Root-Mean-Square-Error (RMS-Error) bezeichnet. Die Standardabweichung s_R ist ein absolutes Maß in den Einheiten der Zielgröße und damit ein absolutes Maß für die Genauigkeit einer Regression. Um universelle Gütekriterien für Regressionen zu erhalten, die unabhängig von der zugrundeliegenden physikalischen Größe sind, müssen deshalb weitere Kenngrößen eingeführt werden.

12.4.2 Bewertung der Regressionsgüte über das Bestimmtheitsmaß

Ein bewährtes Maß der Statistik für die Bewertung von Streuungen ist die Varianz einer Zufallsvariable. Aus diesem Grund wird die Varianz der Stichprobenwerte y_n analysiert.

$$s_y^2 = \frac{1}{N - 1} \cdot \sum_{n=1}^N (y_n - \bar{y})^2 \quad (12.112)$$

Ähnlich wie bei der Varianzanalyse kann die Varianz der Stichprobenwerte in unterschiedliche Summanden zerlegt werden.

$$\begin{aligned}
 (N-1) \cdot s_y^2 &= \sum_{n=1}^N (y_n - \bar{y})^2 = \sum_{n=1}^N (y_n - y(x_n) + y(x_n) - \bar{y})^2 \\
 &= \sum_{n=1}^N ((y_n - y(x_n))^2 + 2 \cdot (y_n - y(x_n)) \cdot (y(x_n) - \bar{y}) + (y(x_n) - \bar{y})^2) \\
 &= \sum_{n=1}^N (y_n - y(x_n))^2 + 2 \cdot \sum_{n=1}^N ((y_n - y(x_n)) \cdot (y(x_n) - \bar{y})) + \sum_{n=1}^N (y(x_n) - \bar{y})^2
 \end{aligned} \tag{12.113}$$

Der mittlere Summand kann umgeformt werden zu

$$2 \cdot \sum_{n=1}^N ((y_n - y(x_n)) \cdot (y(x_n) - \bar{y})) = 2 \cdot \sum_{n=1}^N (r_n \cdot (y(x_n) - \bar{y})) = 2 \cdot \sum_{n=1}^N (r_n \cdot y(x_n)) - 2 \cdot \bar{y} \cdot \sum_{n=1}^N r_n \tag{12.114}$$

Da die Summe der Residuen null ist, vereinfacht sich der Ausdruck zu

$$2 \cdot \sum_{n=1}^N ((y_n - y(x_n)) \cdot (y(x_n) - \bar{y})) = 2 \cdot \sum_{n=1}^N (r_n \cdot y(x_n)) - 0 \tag{12.115}$$

Mit Gleichung (12.21) kann gezeigt werden, dass damit der gesamte Ausdruck zu null wird.

$$\begin{aligned}
 2 \cdot \sum_{n=1}^N (r_n \cdot y(x_n)) &= 2 \cdot \sum_{n=1}^N (r_n \cdot (b_1 \cdot x_n + b_0)) = 2 \cdot b_1 \cdot \sum_{n=1}^N (r_n \cdot x_n) + 2 \cdot b_0 \cdot \sum_{n=1}^N r_n \\
 &= 2 \cdot b_1 \cdot \sum_{n=1}^N (r_n \cdot x_n) + 0 = 2 \cdot b_1 \cdot \sum_{n=1}^N ((y_n - y(x_n)) \cdot x_n) \\
 &= 2 \cdot b_1 \cdot ((y_n - b_1 \cdot x_n - b_0) \cdot x_n) = 0
 \end{aligned} \tag{12.116}$$

Die Varianz der Stichprobenwerte y_n reduziert sich damit auf zwei Teilsummen

$$(N-1) \cdot s_y^2 = \sum_{n=1}^N (y_n - \bar{y})^2 = \sum_{n=1}^N (y_n - y(x_n))^2 + \sum_{n=1}^N (y(x_n) - \bar{y})^2 = \sum_{n=1}^N r_n^2 + \sum_{n=1}^N (y(x_n) - \bar{y})^2 \tag{12.117}$$

Dabei ist der erste Summand die Quadratsumme der Residuen, der zweite Summand beschreibt die Quadratsumme der Schätzwerte. Division durch die linke Seite ergibt

$$1 = \frac{\sum_{n=1}^N r_n^2}{\sum_{n=1}^N (y_n - \bar{y})^2} + \frac{\sum_{n=1}^N (y(x_n) - \bar{y})^2}{\sum_{n=1}^N (y_n - \bar{y})^2} \tag{12.118}$$

Der erste Summand ist ein Maß für die nicht erklärte Reststreuung. Der zweite Summand wird als Bestimmtheitsmaß R^2 bezeichnet.

$$R^2 = \frac{N-1}{N-1} \cdot \frac{\sum_{n=1}^N (y(x_n) - \bar{y})^2}{\sum_{n=1}^N (y_n - \bar{y})^2} = \frac{\frac{1}{N-1} \cdot \sum_{n=1}^N (y(x_n) - \bar{y})^2}{\frac{1}{N-1} \cdot \sum_{n=1}^N (y_n - \bar{y})^2} = \frac{\frac{1}{N-1} \cdot \sum_{n=1}^N (y(x_n) - \bar{y})^2}{\frac{1}{N-1} \cdot \sum_{n=1}^N (y_n - \bar{y})^2} = \frac{s_{y(x)}^2}{s_y^2} \tag{12.119}$$

Es gibt an, welcher Anteil der Streuung mit der Regression beschrieben wird und liegt um Bereich $0 \leq R^2 \leq 1$. Ein Bestimmtheitsmaß $R^2 = 1$ zeigt, dass die Prognosewerte mit den Stichprobenwerten perfekt übereinstimmen. Bei einem Bestimmtheitsmaß von $R^2 = 0$ besteht kein Zusammenhang zwischen den über die Regressionsfunktion geschätzten Werten und den vorliegenden Stichprobenwerten.

Das Bestimmtheitsmaß wächst mit steigender Anzahl von Regressionstermen stetig an. In Abschnitt 11.2.1 wird ausgeführt, dass eine Regression höherer Ordnung nicht zwangsläufig die bessere Regression ist. Deshalb wird das sogenannte adjugierte Bestimmtheitsmaß eingeführt.

$$R_{adj}^2 = 1 - \frac{N-1}{N-M-1} \cdot (1 - R^2) \quad (12.120)$$

Für das Beispiel des Öltemperatursensors aus Tabelle 12.1 ergeben sich folgende Gütekriterien für die unterschiedlichen Regressionen.

Tabelle 12.13: Gütekriterien der unterschiedlichen Regressionen für das Beispiel des Öltemperatursensors mit einem Stichprobenumfang von $N = 11$

Ordnung Regression	FG	a	S_R	R^2	R_{adj}^2
$M = 1$	9	0.0491	0.0739	0.9816	0.9796
$M = 2$	8	0.0443	0.0744	0.9834	0.99793
$M = 3$	9	0.0239	0.0584	0.9911	0.9872

Das Bestimmtheitsmaß und das adjugierte Bestimmtheitsmaß zeigen, dass die Regression in allen Fällen einen sehr großen Anteil der Streuungen abgedeckt. An der Summe der quadratischen Fehler a wird aber auch deutlich, dass der Übergang von linearer Regression auf eine Regression der Ordnung $M = 2$ keinen wesentlichen Einfluss hat.

MATLAB erlaubt die Berechnung des Bestimmtheitsmaßes und des adjungierten Bestimmtheitsmaßes mit dem Befehl `regstats`.

```

1 % MATLAB-Funktionen zur Bewertung der Signifikanz
2 stats = regstats(U,T, 'quadratic')
3 stats.rsquare
4 stats.adjrsquare

```

12.5 Statistische Bewertung des Bestimmtheitsmaßes

Für eine statistische Bewertung des Bestimmtheitsmaßes wird ein Hypothesentest eingeführt. Er geht von der Nullhypothese H_0 aus, dass die Regressionskoeffizienten $\beta_m = 0$ sind. Zur Bewertung wird wie bei der Varianzanalyse eine Zufallsvariable eingeführt, mit der das Verhältnis der erklärten Streuungen zu unerklärten Streuungen bewertet werden kann. Unter Berücksichtigung der Freiheitsgrade ergibt sich das Verhältnis der Varianzen zu

$$v = \frac{\frac{1}{M} \cdot \sum_{n=1}^N (y(x_n) - \bar{y})^2}{\frac{1}{N-M-1} \cdot \sum_{n=1}^N (y_n - \hat{y}_n)^2} = \frac{\frac{1}{M} \cdot \sum_{n=1}^N (y(x_n) - \bar{y})^2}{\frac{1}{N-M-1} \cdot \sum_{n=1}^N r_n^2} = \frac{\frac{1}{M} \cdot \sum_{n=1}^N R^2}{\frac{1}{N-M-1} \cdot (1-R^2)} \quad (12.121)$$

Die Variable weist eine F-Verteilung mit $(M, N - M - 1)$ Freiheitsgraden auf. Falls die Hypothese richtig ist, muss R^2 sehr klein sein. Zur Bewertung wird das Signifikanzniveau α herangezogen.

$$P(v < c) = F(c) = 1 - \alpha \quad (12.122)$$

Die Grenze zur Annahme der Hypothese liegt mit der inversen F-Verteilung mit $(M, N - M - 1)$ Freiheitsgraden damit bei

$$c = F^{-1}(1 - \alpha) \quad (12.123)$$

Die Hypothese wird verworfen, wenn das vorliegende Bestimmtheitsmaß zu einer Kenngröße

$$v_0 = \frac{\frac{1}{M} \cdot \sum_{n=1}^N R^2}{\frac{1}{N-M-1} \cdot (1-R^2)} \quad (12.124)$$

führt, die größer als die Grenze c ist. Alternativ kann eine Überschreitungswahrscheinlichkeit p der Prüfgröße v_0 bestimmt werden und mit dem Signifikanzniveau α verglichen werden. Für einen signifikant von null verschiedenen Regressionskoeffizienten muss die Bedingungen

$$1 - F(v_0) < \alpha \quad (12.125)$$

erfüllt werden. Damit lässt sich der Test des Regressionskoeffizienten in folgenden Prozessschritten zusammenfassen.

Tabelle 12.14: Test der Hypothese $R^2 = 0$ gegen $R^2 > 0$ für das Bestimmtheitsmaß einer linearen Regression

Nr.	Prozessschritt	
1	Wahl eines Signifikanzniveaus α	
2	Bestimmung des zugehörigen Parameters c aus der inversen F-Verteilung ($M, N - M - 1$) Freiheitsgraden $c = F^{-1}(1 - \alpha)$	
3	Berechnung des Bestimmtheitsmaßes der Stichprobe $R^2 = \frac{s_{y(x)}^2}{s_y^2}$	
4	Berechnung der Zufallsvariable für die Stichprobe $v_0 = \frac{\frac{1}{M} \cdot \sum_{n=1}^N R^2}{\frac{1}{N - M - 1} \cdot (1 - R^2)}$	
5	Bestimmung des Annahmebereichs $v_0 \leq c$	Berechnung des p-Values mit der F-Verteilung mit ($M, N - M - 1$) Freiheitsgraden $p = F(v_0)$
6	Für $v_0 \leq c$ wird die Hypothese angenommen, für $v_0 > c$ wird die Hypothese verworfen	Für $p \leq 1 - \alpha$ wird die Hypothese angenommen, für $p > 1 - \alpha$ wird die Hypothese verworfen

12.5.1 Bestimmtheitsmaß und Korrelationskoeffizient

In Kapitel 9 wird der Korrelationskoeffizient r als Maß für die lineare Abhängigkeit zweier Größen x und y eingeführt.

$$r = \frac{s_{xy}}{s_x \cdot s_y} \quad (12.126)$$

Im Fall einer linearen Regression wird erwartet, dass die Stichprobenwerte y_n und die Schätzwerte $y(x_n)$ eine große Korrelation aufweisen. Andererseits wird im vorangegangenen Abschnitt das Bestimmtheitsmaß R^2 als Kenngröße für die Güte der Regression eingeführt.

$$R^2 = \frac{\frac{1}{N-1} \cdot \sum_{n=1}^N (y(x_n) - \bar{y})^2}{\frac{1}{N-1} \cdot \sum_{n=1}^N (y_n - \bar{y})^2} = \frac{s_{y(x)}^2}{s_y^2} \quad (12.127)$$

Bereits durch die Definitionen in Gleichung (12.126) und Gleichung (12.127) kann ein Zusammenhang der beiden Kenngrößen erahnt werden. Ausgehend von der empirischen Varianz der Schätzung

$$s_{y(x)}^2 = \frac{1}{N-1} \cdot \sum_{n=1}^N (y(x_n) - \bar{y})^2 \quad (12.128)$$

kann durch Umformen der vereinfachte Ausdruck in Abhängigkeit der Stichprobenvarianz der Größe x gefunden werden.

$$s_{y(x)}^2 = \frac{1}{N-1} \cdot \sum_{n=1}^N (y(x_n) - \bar{y})^2 = \frac{1}{N-1} \cdot \sum_{n=1}^N (b_1 \cdot (x_n - \bar{x}) + \bar{y} - \bar{y})^2 \quad (12.129)$$

Mit der Definition des Regressionsparameters b_1 aus Gleichung (12.18)

$$b_1 = \frac{s_{xy}}{s_x^2} \quad (12.130)$$

folgt aus Gleichung (12.130)

$$s_{y(x)}^2 = \frac{s_{xy}^2}{s_x^4} \cdot s_x^2 = \frac{s_{xy}^2}{s_x^2} \quad (12.131)$$

Mit der Definitionsgleichung des Bestimmtheitsmaßes R^2 und Gleichung (12.131) ergibt sich bei linearer Regressionsfunktion der Zusammenhang zwischen dem Bestimmtheitsmaß R^2 und dem Korrelationskoeffizienten r aus Kapitel 9

$$R^2 = \frac{s_{y(x)}^2}{s_y^2} = \frac{s_{xy}^2}{s_x^2 \cdot s_y^2} = r^2 \quad (12.132)$$

Bei einer linearen Regression entspricht das Bestimmtheitsmaß R^2 dem Quadrat des Korrelationskoeffizienten r^2 .

12.5.2 Reststreuungsanalyse

Die Reststreuungsanalyse dient dazu, die durch das Modell nicht erklärte Streuung eingehender zu untersuchen. Sie basiert auf den Residuen r_n , also den Abweichungen zwischen den Stichprobenwerten und den entsprechenden Werten der Regressionsfunktion.

$$r_n = y_n - y(x_n) \quad (12.133)$$

Die Residuen spiegeln die in der Regression nicht abgebildeten Abweichungen wieder. Mit ihnen lässt sich entscheiden, ob die Modellannahmen korrekt sind, und es lassen sich eventuelle Besonderheiten erkennen. Bild 12.13 stellt für das Beispiel des Öltemperatursensors die Residuen für eine lineare Regressionsfunktion dar.

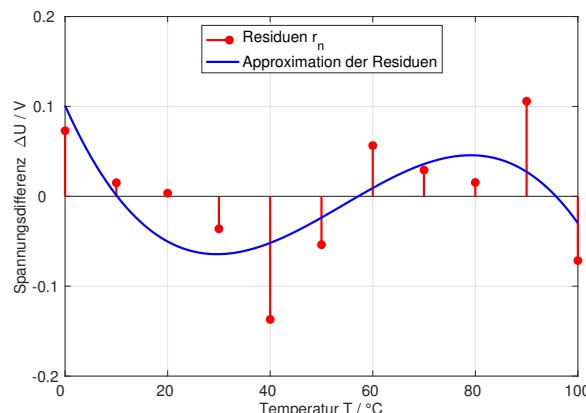


Bild 12.13: Approximation der Residuen für das Beispiel des Öltemperatursensors durch ein Polynom höherer Ordnung

Die Residuen zeigen einen Verlauf, der durch ein Polynom dritter Ordnung beschrieben werden kann. Die Reststreuungsanalyse zeigt damit, dass eine Regressionsfunktion höherer Ordnung das Regressionsergebnis weiter verbessern würde.

Bild 12.14 stellt als weiteres Beispiel die lineare Regression von konstruierten Stichprobenwerten sowie die Residuen der Regression dar. Die Stichprobenwerte könnten zum Beispiel zwei Merkmale x und y eines Produktes als Funktion des Fertigungszeitpunktes beschreiben. Zwischen den Größen x und y soll aus physikalischen Gründen näherungsweise ein linearer Zusammenhang existieren. Für die Stichprobenwerte wird eine lineare Regression durchgeführt und die Residuen werden bestimmt. Bild 12.14 stellt das Ergebnis dar.

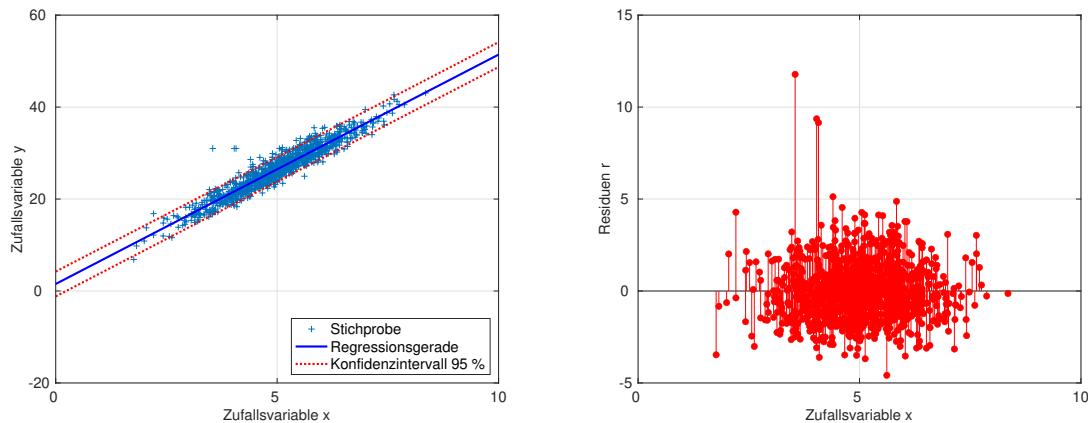


Bild 12.14: Darstellung eines Testdatensatzes mit linearer Regression und den zugehörigen Residuen

Die Residuen in Bild 12.14 werden zunächst plausibilisiert. Die Punkte liegen über den betrachteten Bereich auf ungefähr konstantem Niveau. Würden die Residuen wie bei dem Beispiel zum Öltemperatursensor einen bogenförmigen Verlauf aufweisen, wäre dies ein Hinweis darauf, dass die Ordnung der Regressionsfunktion nicht ausreichend ist. Die Ergebnisse hier zeigen, dass die Wahl der Regressionsfunktion sinnvoll ist. Deshalb wird vermutet, dass auch der physikalische Hintergrund auf einen linearen Zusammenhang zwischen den Zufallsvariablen x und y hinweist.

Nach der ersten Plausibilisierung werden die Residuen auf Ausreißer geprüft. Dazu werden die Residuen in Bild 12.15 als Box-Plot dargestellt.

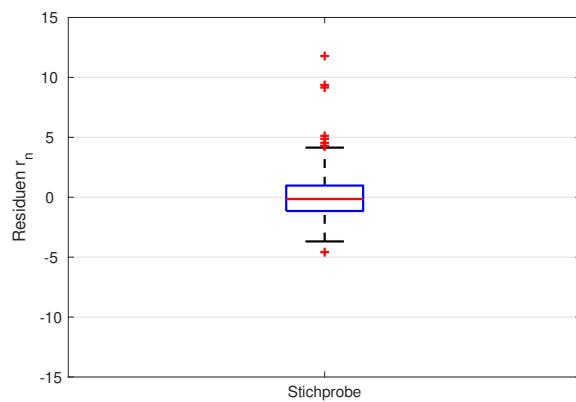


Bild 12.15: Darstellung der Residuen einer Stichprobe als Box-Plot

Bei dem Box-Plot werden einzelne Stichprobenwerte als Ausreißer erkannt. Die Ursachen für diese Abweichungen sind in dem Regressionsmodell nicht abgebildet. Ausreißer sind auf nicht berücksichtigte Einflussfaktoren oder bislang nicht erkannte Störungen in dem Prozessablauf zurückzuführen. Erkannte

Ausreißer helfen damit, Ursachen von Prozessabläufen einzugrenzen. Ihre Analyse ist für die Optimierung des zugrundeliegenden Prozesses von großem Nutzen.

Nach Entfernen der Ausreißer werden die Residuen einem Test auf Normalverteilung unterzogen. Dazu kann ein in [Krey91] beschriebener Goodness-Of-Fit-Test herangezogen. Bei dem Test wird die Hypothese geprüft, dass die Residuen eine Normalverteilung aufweisen. Der p-Wert für diesen Test berechnet sich zu

$$p = 0.0015 \quad (12.134)$$

Da der p-Wert kleiner als das Signifikanzniveau $\alpha = 0.05$ ist, können die Residuen als normalverteilt angesehen werden.

Zur Analyse der Ursachen für Ausreißer werden die Residuen in ihrer zeitlichen Abfolge dargestellt.

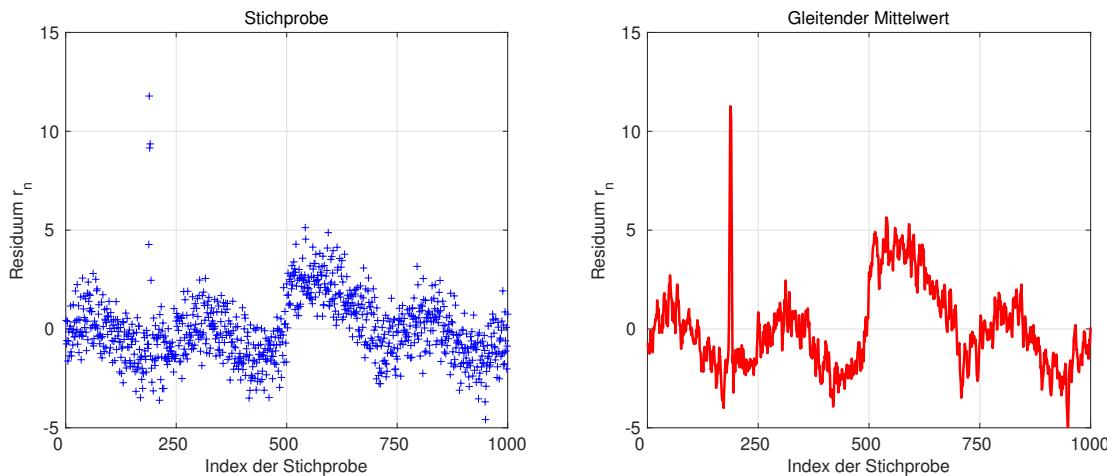


Bild 12.16: Reststreuungen in der zeitlichen Reihenfolge der Stichprobenaufnahme

Bereits bei einem einfachen Plot der Stichprobe über dem Stichprobenindex wird offensichtlich, dass der Prozess, dem diese Stichprobe entnommen wurde, nicht als stationär bezeichnet werden kann. Zur Minimierung der Streuung kann ein gleitender Mittelwert über die Stichprobe gebildet werden. In diesem Fall werden fünf aufeinander folgende Werte gemittelt. Nach der Mittelung ist eine langwellige Schwingung deutlicher zu erkennen. Ursache könnte zum Beispiel eine Abhängigkeit von der Fertigungsschicht oder der Lufttemperatur sein. Außerdem können einige Ausreißer, die im Box-Plot der Residuen auffällig sind, den Indizes 200 - 204 zugeordnet werden. Hier muss sich eine Analyse der Fertigungs- und Messbedingungen für diese Teile anschließen. Weiterhin ist in dem Bereich der Indizes 500 - 700 ein Sprung zu erkennen, der zum Beispiel auf eine auffällige Charge von Zukaufteilen hinweisen kann.

Diese oder andere Filterfunktionen helfen, Anhaltspunkte für die Ursachen von Störungen zu identifizieren. Die Ursache für die Störungen können mit den Zeitangaben gezielt gesucht werden. Diese Informationen erlauben, Prozessschwankungen zeitlich einzugrenzen und gezielt nach den Ursachen für diese Schwankungen zu suchen.

Die Residuen können mit dem Befehl `regstats` berechnet werden:

```

1 % MATLAB-Funktionen zur Bewertung der Signifikanz
2 stats = regstats(U,T, 'linear')
3 stats.r

```

12.6 Sonderformen der zweidimensionalen Regression

In praktischen Aufgabenstellungen wird der Einsatz von Regressionsfunktionen aus verschiedenen Gründen erschwert.

- Zusammenhänge lassen sich nicht zielführend mit Polynomen beschreiben
- Ausreißer verfälschen das Ergebnis maßgeblich

Für beide Einschränkungen werden Lösungsansätze skizziert.

12.6.1 Nichtlineare Regression

Einige technische Aufgabenstellungen der Regression haben die Besonderheit, dass der grundsätzliche funktionale Zusammenhang bekannt ist und nur Parameter des funktionalen Zusammenhangs bestimmt werden müssen. Zum Beispiel lautet die Shockley-Gleichung für den Strom durch eine Diode

$$I_D = I_s \cdot \left(e^{\frac{U_D}{n \cdot U_T}} - 1 \right) \quad (12.135)$$

Dabei ist I_D der Strom durch die Diode, I_s der Sättigungssperrstrom, U_D die an der Diode anliegende Sättigungsspannung, n der Emissionskoeffizient und U_T die temperaturabhängige Temperaturspannung.

Da die Physik des Diodenstroms und seine mathematische Beschreibung bekannt sind, ist es zielführend, dieses Wissen in die Approximation oder Regression einzubinden. Es wird deshalb kein Polynom als Approximationsfunktion verwendet, sondern es wird die Approximationsfunktion verwendet, die sich aus der Physik ergibt.

Um den Diodenstrom I_D als Funktion der anliegenden Spannung U_D darzustellen, werden einige Messwerte aufgenommen. Sie sind in Tabelle 12.15 aufgeführt.

Tabelle 12.15: Messung der Strom-Spannungskennlinie für eine Diode

U_D / V	0	0.1	0.2	0.3	0.4	0.5	0.55	0.6	0.65	0.7
I_D / mA	0	0.000	0.001	0.005	0.089	1.537	6.385	26.54	110.3	458.3

Es liegen 10 Messungen vor, und es müssen die Parameter Sättigungsstrom I_s , Emissionskoeffizient n und Temperaturspannung U_T bestimmt werden. Die Aufgabe ist lösbar, hier sogar überbestimmt.

Auch bei der nichtlinearen Regression kann die Lösung über die Minimierung der Fehlerquadrate erfolgen. Diese Methode führt aber zu einem nichtlinearen Parameteroptimierungsproblem, das im Allgemeinen nur mit numerischen Optimierungsverfahren gelöst werden kann. Zur Lösung wurde hier die MATLAB-Funktion `nlinfit` verwendet. Als Ergebnis ergeben sich die Parameter $I_s = 2nA$, $U_T = 24.7mV$ und $n = 1.423$.

Um den Vorteil des Verfahrens darzustellen, werden die Messwerte einerseits über Polynome der Ordnung $M = 3$ und $M = 5$ dargestellt, andererseits mit den bestimmten Parametern der Diodengleichung berechnet und zur Approximation verwendet.

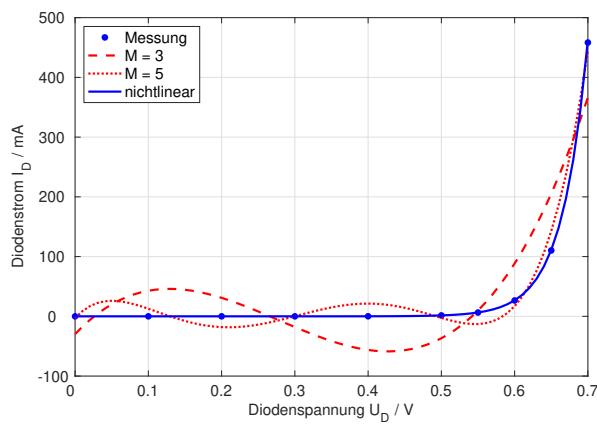


Bild 12.17: Vergleich der Regression von Messwerten mit Polynomen der Ordnung 3 und 5 mit einer physikalisch begründeten nichtlinearen Regression

Es zeigt sich, dass das Polynom höherer Ordnung die Messwerte nicht sinnvoll approximiert, sondern anfängt zu schwingen. Im Gegensatz dazu stellt die physikalisch begründete, nichtlineare Regression eine gute Approximation dar.

Die Berechnung in MATLAB erfordert zunächst die Definition der nichtlinearen Regressionsfunktion:

```

1 % Definition der Funktion idcalc zur Berechnung des Diodenstroms als
2 % Funktion
3 % von Parametern
4 function [ id ] = Idcalc( beta ,u );
5
6 % Auflösen des Parametervektors beta
7 iscalc = beta(1));
8 utcalc = beta(2);
9 ncalc = beta(3));
10
11 % Berechnung des Diodenstroms
12 id = iscalc*(exp(u/ncalc/utcalc )-1)*1000;

```

Dann können mit der Funktion nlinfit die Parameter der nichtlinearen Regressionsfunktion geschätzt werden.

```

1 % Definition der Zahlenwerte
2 ud = [0 0.1 0.2 0.3 0.4 0.5 0.55 0.6 0.65 0.7];
3
4 % Definition der wahren Parameter und des wahren Diodenstroms
5 is = 1e-9;
6 ut = 27e-3;
7 n = 1.3;
8 id = (is*(exp(ud/n/ut )-1))*1000;
9
10 % Aufruf der nichtlinearen Approximation
11 [ betahat ,R,J] = nlinfit(ud, id, @Idcalc , beta );
12
13 % Interpretation des Parametervektors
14 isa = betahat(1);
15 uta = betahat(2);
16 na = betahat(3);
17

```

18 | % Berechnung des Regressionsergebnisses

19 | idapp = (isa*(exp(udint/na/uta)-1))*(1000+10*rand);

Der nichtlineare funktionale Zusammenhang zwischen der Diodenspannung U_D und dem Diodenstrom I_D wird hier mit einer Exponentialfunktion beschrieben. Alternativ könnte die Diodenspannung transformiert werden, sodass ein linearer funktionaler Zusammenhang entsteht. Mit der Linearisierung des Problems können alle statistischen Verfahren zur Berechnung und Bewertung der Regressionsfunktion genutzt werden. Das betrifft sämtliche Testverfahren, wie die Berechnung der Konfidenz- und Prognosebereiche sowie die Bewertung der Regressionsgüte.

12.6.2 Behandlung von Ausreißern - Robuste Regression

Der in Bild 11.14 dargestellte Datensatz weist einige Ausreißer auf. Aufgrund der großen Datenmenge fielen diese Ausreißer nicht ins Gewicht. Liegen nur wenige Stichprobenwerte vor, können Ausreißer das Ergebnis stark verfälschen. Um diesen Effekt zu demonstrieren, wird erneut das Beispiel des Temperatursensors zur Messung der Öltemperatur aufgegriffen. Diesmal wird der Messwert bei 80°C signifikant verändert. Es ergeben sich die in Tabelle 12.16 aufgelisteten Werte.

Tabelle 12.16: Stichprobe für den Zusammenhang zwischen Öltemperatur und Ausgangsspannung eines Temperatursensors, Ausreißer bei der Messung 80°C

Temperatur T / °C	0	10	20	30	40	50
Spannung U / V	2.766	2.862	3.005	3.120	3.173	3.411
Temperatur T / °C	60	70	80	90	100	
Spannung U / V	3.676	3.803	1.944	4.188	4.165	

Bild 12.18 vergleicht die lineare Regression erster Ordnung für die Stichprobe mit und ohne Ausreißer.

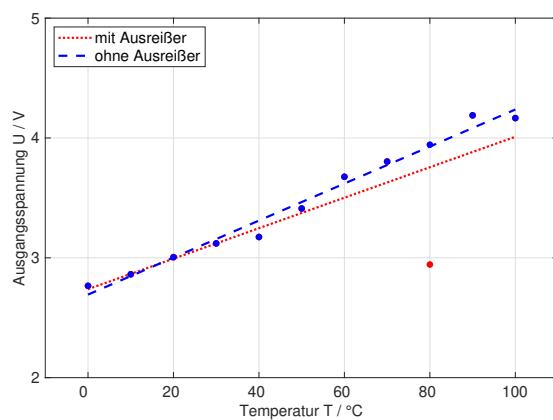


Bild 12.18: Vergleich der Regression von Messwerten mit und ohne Ausreißer

Für die Regression dieser Daten stehen in einigen Programmpaketen Funktionen zur Verfügung, die die unterschiedlichen Stichprobenwerte unterschiedlich gewichtet. Das führt dazu, dass Ausreißer weniger stark in die Summe der Fehlerquadrate eingehen und damit das Regressionsergebnis weniger stark beeinflussen. Bild v stellt diesen Prozess grafisch dar.

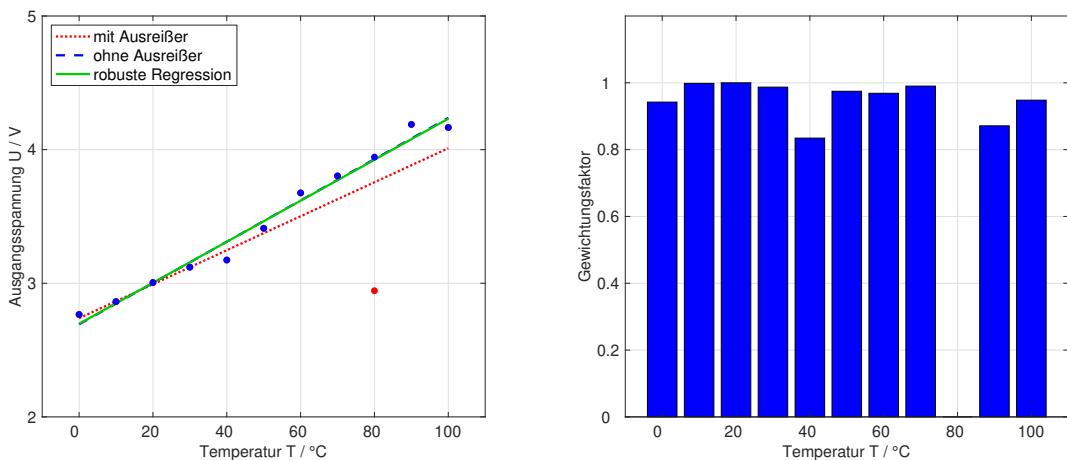


Bild 12.19: Vergleich der Regression von Messwerten mit und ohne Gewichtung der Ausreißer und Darstellung der Gewichtungsfaktoren

Es wird deutlich, dass der Messwert bei $80^\circ C$ mit einem Gewichtungsfaktor nahe null multipliziert wird, sodass dieser Messwert praktisch nicht in das Regressionsergebnis eingeht. Dadurch wird die ursprüngliche Regressionsgerade wieder erreicht.

Die Gewichtungsfaktoren werden dabei iterativ bestimmt. Als Bewertungskriterium werden die Residuen herangezogen. Je größer das Residuum ist, desto weniger stark wird die entsprechende Stelle der Stichprobe zur Berechnung der Regressionsfunktion gewichtet. Dieses Vorgehen ist auch an der Darstellung der Gewichtungsfaktoren in Bild 12.19 zu erkennen. Der Ausreißer bei $80^\circ C$ hat praktisch kein Gewicht. Auch die Stichprobenwerte bei $40^\circ C$ und $90^\circ C$ haben ein reduziertes Gewicht, weil sie vergleichsweise stark von der Regressionsfunktion abweichen.

MATLAB bietet mit dem Befehl `robustfit` eine Möglichkeit, eine robuste Regression durchzuführen und zu bewerten.

12.7 Literatur

- [Krey91] Kreyszig, Erwin: Statistische Methoden und ihre Anwendungen
4., unveränderter Nachdruck der 7. Auflage
Vandenhoeck & Ruprecht, Göttingen, 1991
- [Fahr96] Fahrmeir, Ludwig; Hamerle, Alfred; Tutz, Gerhard: Multivariate statistische Verfahren
2., überarbeitete Auflage
Walter de Gruyter & Co., Berlin
- [Ross06] Ross, M. Sheldon: Statistik für Ingenieure und Naturwissenschaftler
3. Auflage
Spektrum Akademischer Verlag, München, 2006
- [Hart07] Hartung, Joachim; Elpelt, Bärbel: Multivariate Statistik
7., unveränderte Auflage
R. Oldenbourg Verlag, München / Wien
- [Papu01] Papula, Lothar: Mathematik für Ingenieure und Naturwissenschaftler Band 3
4., verbesserte Auflage
Vieweg Teubner, Braunschweig / Wiesbaden, 2008

13 Regression mehrdimensionaler Datensätze

Viele Aufgaben des Design For Six Sigma sind mehrdimensional. Deshalb wird das Vorgehen der eindimensionalen Aufgabenstellungen auf mehrdimensionale Regressionsmodelle erweitert. Dabei wird auch gezeigt, wie Polynome als Regressionsfunktion eingesetzt und bewertet werden können.

13.1 Bestimmung der Regressionsfunktion

Die Herleitung der Regressionsfunktion für mehrdimensionale Datensätze wird genauso durchgeführt wie bei eindimensionalen Datensätzen. Aufgrund der unterschiedlichen Eingangsvariablen ist die Bezeichnung der Variablen komplizierter. Um eine effiziente Darstellungsform zu erhalten, wird deshalb die Matrizenrechnung eingesetzt. Die erforderlichen Grundlagen der linearen Algebra sind im Anhang zusammengefasst.

13.1.1 Modellansatz und Datenstruktur

Ausgangspunkt für mehrdimensionale Regressionsfunktionen ist ein Prozess, dessen Ausgangssignal y durch M Eingangsgrößen x_m und einem Messfehler e beschrieben wird.

$$y = \beta_0 + \beta_1 \cdot x_1 + \dots + \beta_m \cdot x_m + \dots + \beta_M \cdot x_M + e \quad (13.1)$$

Dabei ist e ein zufälliger Fehler, der einen Mittelwert $\mu_e = 0$ und eine unbekannte konstante Varianz σ^2 aufweist. Die Regressionskoeffizienten β_m werden auf Basis einer Stichprobe geschätzt. Da $M + 1$ Regressionsparameter geschätzt werden müssen, sind mindestens $N = M + 1$ Stichproben erforderlich. Jede Stichprobe besteht aus einem Satz von Eingangsgrößen $x_{n1} \dots x_{nM}$ und einem Messwert der Zielgröße y_n . Die Eingangswerte werden in der Matrix X

$$X = \begin{pmatrix} 1 & \cdots & x_{1m} & \cdots & x_{1M} \\ \vdots & \ddots & \vdots & & \vdots \\ 1 & \cdots & x_{nm} & \cdots & x_{nM} \\ \vdots & & \vdots & \ddots & \vdots \\ 1 & \cdots & x_{Nm} & \cdots & x_{NM} \end{pmatrix} \quad (13.2)$$

und die Zielgrößen in einem Vektor y

$$\underline{y} = \begin{pmatrix} y_1 \\ \vdots \\ y_n \\ \vdots \\ y_N \end{pmatrix} \quad (13.3)$$

zusammengefasst. Ziel der Regressionrechnung ist es, Schätzwerte b_m für die Regressionsparameter β_m zu finden.

$$y(\underline{x}^T) = b_0 + b_1 \cdot x_1 + \dots + b_m \cdot x_m + \dots + b_M \cdot x_M \quad (13.4)$$

Dazu wird folgendes Gleichungssystem aufgebaut:

$$\begin{aligned} \underline{\underline{y}} &= \begin{pmatrix} y_1 \\ \vdots \\ y_n \\ \vdots \\ y_N \end{pmatrix} = \begin{pmatrix} b_0 + b_1 \cdot x_{11} + \dots + b_m \cdot x_{1m} + \dots + b_M \cdot x_{1M} + r_1 \\ \vdots \\ b_0 + b_1 \cdot x_{n1} + \dots + b_m \cdot x_{nm} + \dots + b_M \cdot x_{nM} + r_n \\ \vdots \\ b_0 + b_1 \cdot x_{N1} + \dots + b_m \cdot x_{Nm} + \dots + b_M \cdot x_{NM} + r_N \end{pmatrix} \\ &= \begin{pmatrix} 1 & \dots & x_{1m} & \dots & x_{1M} \\ \vdots & \ddots & \vdots & & \vdots \\ 1 & \dots & x_{nm} & \dots & x_{nM} \\ \vdots & & \vdots & \ddots & \vdots \\ 1 & \dots & x_{Nm} & \dots & x_{NM} \end{pmatrix} \cdot \begin{pmatrix} b_0 \\ \vdots \\ b_m \\ \vdots \\ b_M \end{pmatrix} + \begin{pmatrix} r_1 \\ \vdots \\ r_n \\ \vdots \\ r_N \end{pmatrix} = \mathbf{X} \cdot \underline{\underline{b}} + \underline{\underline{r}} \end{aligned} \quad (13.5)$$

Die Regressionsfunktion mit den Schätzwerten $\underline{\underline{b}}$ kann die Stichprobenwerte nicht perfekt abbilden kann. Wie im eindimensionalen Fall werden die Abweichungen als Residuen $\underline{\underline{r}}$ bezeichnet.

13.1.2 Herleitung der Bestimmungsgleichung

Zur Bestimmung der Regressionsgleichung wird die Quadratsumme der Residuen minimiert. Die Residuen errechnen sich aus der Differenz des gemessenen Wertes y_n und dem Wert der Regressionsfunktion an der entsprechenden Stelle

$$\underline{\underline{r}} = \underline{\underline{y}} - \mathbf{X} \cdot \underline{\underline{b}} \quad (13.6)$$

Die Quadratsumme kann als Innenprodukt dargestellt werden.

$$\begin{aligned} a &= \underline{\underline{r}}^T \cdot \underline{\underline{r}} = (\underline{\underline{y}} - \mathbf{X} \cdot \underline{\underline{b}})^T \cdot (\underline{\underline{y}} - \mathbf{X} \cdot \underline{\underline{b}}) = (\underline{\underline{y}}^T - \underline{\underline{b}}^T \cdot \mathbf{X}^T) \cdot (\underline{\underline{y}} - \mathbf{X} \cdot \underline{\underline{b}}) \\ &= \underline{\underline{y}}^T \cdot \underline{\underline{y}} - \underline{\underline{y}}^T \cdot \mathbf{X} \cdot \underline{\underline{b}} - \underline{\underline{b}}^T \cdot \mathbf{X}^T \cdot \underline{\underline{y}} + \underline{\underline{b}}^T \cdot \mathbf{X}^T \cdot \mathbf{X} \cdot \underline{\underline{b}} \end{aligned} \quad (13.7)$$

Wegen der Beziehung

$$\underline{\underline{y}}^T \cdot \mathbf{X} \cdot \underline{\underline{b}} = \underline{\underline{b}}^T \cdot \mathbf{X}^T \cdot \underline{\underline{y}} \quad (13.8)$$

ergibt sich für die Quadratsumme der Residuen

$$\begin{aligned} a &= \underline{\underline{r}}^T \cdot \underline{\underline{r}} = (\underline{\underline{y}} - \mathbf{X} \cdot \underline{\underline{b}})^T \cdot (\underline{\underline{y}} - \mathbf{X} \cdot \underline{\underline{b}}) = (\underline{\underline{y}}^T - \underline{\underline{b}}^T \cdot \mathbf{X}^T) \cdot (\underline{\underline{y}} - \mathbf{X} \cdot \underline{\underline{b}}) \\ &= \underline{\underline{y}}^T \cdot \underline{\underline{y}} - 2 \cdot \underline{\underline{b}}^T \cdot \mathbf{X}^T \cdot \underline{\underline{y}} + \underline{\underline{b}}^T \cdot \mathbf{X}^T \cdot \mathbf{X} \cdot \underline{\underline{b}} \end{aligned} \quad (13.9)$$

Die Regressionsparameter b_m werden so gewählt, dass die Summe a der quadratischen Fehler minimal wird. Notwendige Bedingung für ein Minimum ist, dass die partielle Ableitungen von a nach b_m null werden. Es entstehen $M + 1$ Gleichungen für $M + 1$ unbekannte Parameter b_m . Aus diesem Ansatz folgt die Gleichung

$$\frac{\partial a}{\partial \underline{\underline{b}}} = -2 \cdot \mathbf{X}^T \cdot \underline{\underline{y}} + 2 \cdot \mathbf{X}^T \cdot \mathbf{X} \cdot \underline{\underline{b}} = 0 \quad (13.10)$$

beziehungsweise

$$\mathbf{X}^T \cdot \mathbf{X} \cdot \underline{\underline{b}} = \mathbf{X}^T \cdot \underline{\underline{y}} \quad (13.11)$$

Für den Fall, dass die Inverse von $\mathbf{X}^T \cdot \mathbf{X}$ existiert, ergibt sich die Bestimmungsgleichung

$$\underline{\underline{b}} = (\mathbf{X}^T \cdot \mathbf{X})^{-1} \cdot \mathbf{X}^T \cdot \underline{\underline{y}} \quad (13.12)$$

Dabei sind die Parameter b_m Schätzwerte für die Parameter β_m des zugrundeliegenden Prozesses.

Beispiel: Invertierender Verstärker

Als Beispiel wird ein invertierender Verstärker aufgegriffen. Es wird die Abhängigkeit der Ausgangsspannung U_A von der Eingangsspannung U_E und der Offsetspannung U_{OFF} beschrieben.

$$U_A = -\frac{R_2}{R_1} \cdot U_E + \left(1 + \frac{R_2}{R_1}\right) \cdot U_{OFF} \quad (13.13)$$

Die Funktion soll über eine Regressionsrechnung ermittelt werden. Dazu wird eine Schaltung mit den Widerständen $R_1 = 10k\Omega$ und $R_2 = 1k\Omega$ aufgebaut. Es ergeben sich die in Tabelle 13.1 dargestellten Messwerte.

Tabelle 13.1: Ausgangsspannung eines invertierenden Operationsverstärkers als Beispiel für eine zweidimensionale lineare Regression

Index	U_A / mV	U_{OFF} / mV	U_E / mV
1	5.5954	0	- 50
2	- 5.6012	0	50
3	4.9901	0	- 50
4	- 5.0784	0	50
5	15.1980	10	- 50
6	6.1287	10	50
7	15.4718	10	- 50
8	6.7076	10	50
9	5.0975	5	0

Unter Berücksichtigung der Konstante b_0 ergibt sich die Matrix \mathbf{X}

$$X = \begin{pmatrix} 1 & 0 & -50 \\ 1 & 0 & 50 \\ 1 & 0 & -50 \\ 1 & 0 & 50 \\ 1 & 10 & -50 \\ 1 & 10 & 50 \\ 1 & 10 & -50 \\ 1 & 10 & 50 \\ 1 & 5 & 0 \end{pmatrix} \quad (13.14)$$

und der Vektor \underline{y} zu

$$\underline{y} = \begin{pmatrix} 5.5954 \\ -5.6012 \\ 4.9901 \\ -5.0784 \\ 15.1980 \\ 6.12875 \\ 15.4718 \\ 6.7076 \\ 5.0975 \end{pmatrix} \quad (13.15)$$

Der Vektor \underline{b} der unbekannten Parameter berechnet sich zu

$$\underline{b} = (\underline{X}^T \cdot \underline{X})^{-1} \cdot \underline{X}^T \cdot \underline{y} = \begin{pmatrix} -0.0601 \\ 1.0900 \\ -0.0977 \end{pmatrix} \quad (13.16)$$

Die Regressionsfläche der geschätzten Ausgangsspannung U_A ergibt sich aus

$$U_A(U_E, U_{OFF}) = \underline{X} \cdot \underline{b} = -0.0601 + 1.0900 \cdot U_{OFF} - 0.0977 \cdot U_E \quad (13.17)$$

Die über die Regressionsfunktion berechnete Ausgangsspannung und die über Gleichung (13.13) berechnete Ausgangsspannung sind in Bild 13.1 zum Vergleich dargestellt.

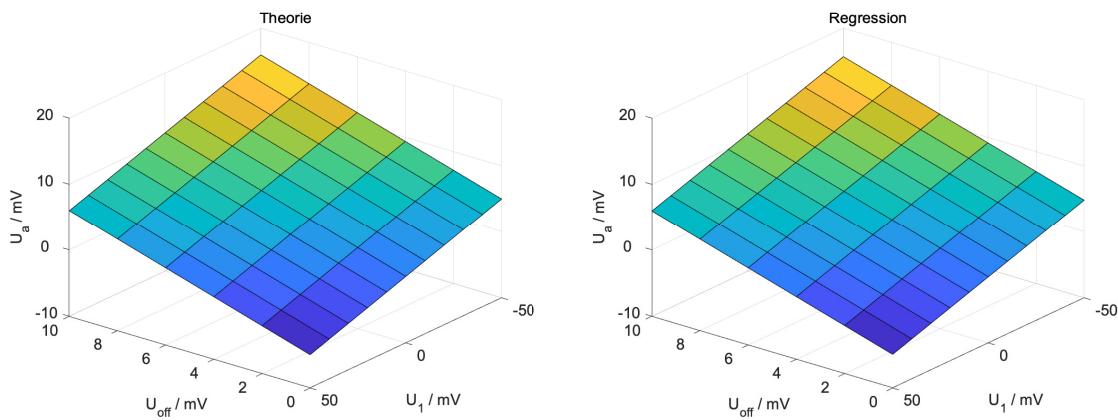


Bild 13.1: Ausgangsspannung eines invertierenden Verstärkers
 a) Theoretische Rechnung
 b) Regressionsrechnung auf Basis der Messwerte in Tabelle 13.1

Aufgrund der Messfehler, die bei der Bestimmung der Ausgangsspannung in Tabelle 13.1 gemacht wurden, stimmt die Regressionsgleichung nicht exakt mit der idealen Gleichung

$$U_A = \left(1 + \frac{R_2}{R_1}\right) \cdot U_{OFF} - \frac{R_2}{R_1} \cdot U_E = 1.1 \cdot U_{OFF} - 0.1 \cdot U_E \quad (13.18)$$

überein. Die bestimmte Regressionsfunktion stellt aber eine gute Näherung der analytischen Gleichung dar.

13.1.3 Geometrische Interpretation der Bestimmungsgleichung

<<< Orthogonale Projektion >>>

13.1.4 Mathematischer Ansatz für mehrdimensionale Regressionsfunktionen

Zur besseren Übersicht wird im vorangegangenen Abschnitt eine zweidimensionale, lineare Funktion verwendet, um eine Regressionsfläche zu berechnen. In der Praxis werden aber Funktionen verwendet, die neben der linearen Abhängigkeit von Eingangsgrößen auch quadratische Abhängigkeiten und Wechselwirkungen abbilden. Die Beschreibung der Zielgröße erfolgt damit für Systeme mit zwei Eingangsgrößen über eine Gleichung der Form

$$y(\underline{x}^T) = b_0 + b_1 \cdot x_1 + b_2 \cdot x_2 + b_3 \cdot x_1 \cdot x_2 + b_4 \cdot x_1^2 + b_5 \cdot x_2^2 \quad (13.19)$$

Dieser quadratische Funktionsansatz geht auf die Taylorreihe zurück. Sie wird verwendet, um Funktionen in der Umgebung definierter Punkte durch Potenzreihen darzustellen. Bei der Taylorreihe handelt es sich um eine lokal gültige Approximation der Funktion. Trotzdem ist die Taylorreihe ein oftmals ausreichender Ansatz, weil die Zielgröße y nur in einem definierten Bereich beschrieben werden soll. Die Funktion kann dabei von mehreren unabhängigen Eingangsgrößen x_m abhängen. Der Funktionsansatz wie in Gleichung (13.19) ist damit mathematisch zulässig und in vielen Fällen wegen der unbekannten Abhängigkeit die einzige Möglichkeit, die Zielgröße y zu beschreiben.

Der Modellansatz in Gleichung (13.19) hat darüber hinaus den Vorteil, dass das Modell mathematisch einfach handhabbar ist. Das erleichtert zum einen für die Bestimmung der Modelfunktion. Zum anderen hat diese einfache mathematische Beschreibung den Vorteil, dass sie auf wenig komplexen mathematischen Operationen aufbaut. Sie kann damit auch in einfachen Systemen wie etwa der Motorsteuerung oder in Steuerungen von Haushaltsgeräten eingesetzt werden.

Der Funktionsansatz in Gleichung (13.19) besteht aus Termen, die bis zur zweiten Ordnung gehen. Er wird deshalb als vollquadratischer Funktionsansatz bezeichnet. Um den funktionalen Zusammenhang

optimieren zu können, muss der Anwender eine Vorstellung von der mathematischen Gestalt unterschiedlicher Funktionsansätze haben. Deshalb werden einige Funktionsansätze kurz vorgestellt. Dabei beschränken sich die Darstellungen auf zwei Eingangsgrößen und eine Zielgröße, um eine räumliche Darstellung zu ermöglichen. Die hier für zwei Eingangsgrößen dargestellten Ansätze lassen sich aber auf beliebig viele Eingangsgrößen erweitern.

Linearer Modellansatz

Ein linearer Ansatz zeichnet sich dadurch aus, dass die Zielgröße über lineare Terme beschrieben werden kann. Die beiden Eingangsgrößen haben keine Wechselwirkung. Mathematisch wird dieser Ansatz beschrieben über

$$y = b_0 + b_1 \cdot x_1 + b_2 \cdot x_2 \quad (13.20)$$

Bild 13.2 stellt die lineare Funktion

$$y = 2.5 + x_1 + x_2 \quad (13.21)$$

grafisch dar. Entlang der Achsen entsprechen die Koeffizienten b_1 und b_2 den Koeffizienten der Steigung der Fläche. Die Zielgröße steigt unabhängig von der Variable x_2 linear mit der Variable x_1 an.

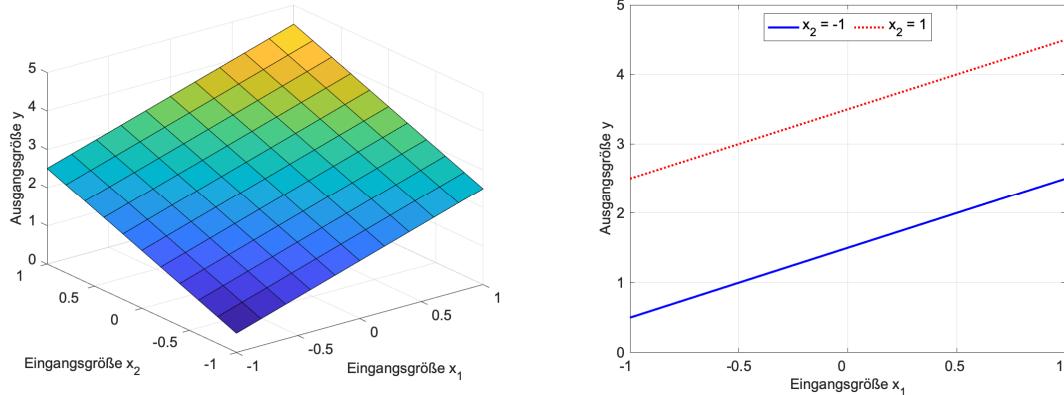


Bild 13.2: Zielgröße y für einen linearen Modellansatz

Dies wird besonders gut deutlich, wenn die Zielgröße wie in dem rechten Teil der Darstellung nur als Funktion einer Variablen, hier der Variablen x_1 , dargestellt wird. Die Linien besitzen immer dieselbe Steigung, sie kreuzen sich nicht. Dieses Bild der parallelen Kennlinien ist charakteristisch für lineare Modelle ohne Wechselwirkung. Die linearen Anteile einer Regression werden als Haupteffekte bezeichnet, da sie in vielen Modellen die wesentlichen Abhängigkeiten abdecken.

Modellansatz mit Wechselwirkungen

Funktionen mit Wechselwirkungen sind an einer Sattelfunktion zu erkennen. Die Wirkung der Eingangsgrößen x_1 ist von der Eingangsgröße x_2 abhängig und umgekehrt. Das kann zu einer gegenseitigen Verstärkung oder Abschwächung führen. Mathematisch Herleitung der Bestimmungsgleichung werden Wechselwirkungen beschrieben durch den Wechselwirkungsterm

$$y = b_0 + b_3 \cdot x_1 \cdot x_2 \quad (13.22)$$

Bild 13.3 stellt die Zielgröße y als Funktion der Variablen x_1 und x_2 für die Funktion

$$y = 2.5 + x_1 \cdot x_2 \quad (13.23)$$

dar.

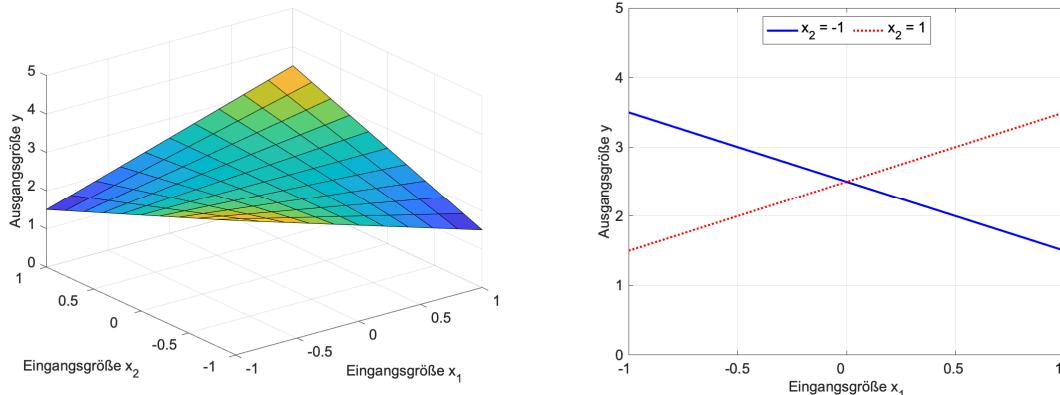


Bild 13.3: Zielgröße y für einen Modellansatz mit Wechselwirkungsterm

Die Steigung in Richtung der Variable x_1 ist von der Variable x_2 abhängig. Diese Abhängigkeit wird im rechten Teil der Darstellung noch deutlicher. In Abhängigkeit der Variable x_2 wechselt die Steigung der Gerade $y(x_1)$ das Vorzeichen.

Quadratischer Modellansatz

Als letzter Ansatz für eine Regressionsfunktion wird der quadratische Ansatz diskutiert. Er wird mathematisch dargestellt über die Gleichung

$$y = 2.5 + b_4 \cdot x_1^2 + b_5 \cdot x_2^2 \quad (13.24)$$

Je nach Vorzeichen der Koeffizienten b_4 und b_5 ergeben sich unterschiedliche Topologien. Bild 13.4 zeigt die Zielgröße für die Funktion

$$y = 2.5 + x_1^2 + x_2^2 \quad (13.25)$$

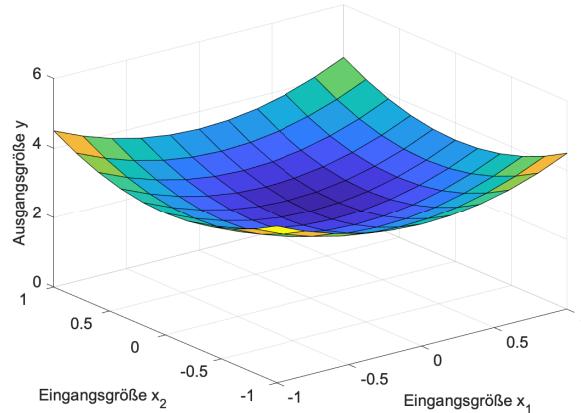


Bild 13.4: Zielgröße y für einen rein quadratischen Modellansatz bei gleichen Vorzeichen

Weil beide Vorzeichen positiv sind, steigt der Zielgröße y , sobald sich die Eingangsgrößen aus dem Koordinatenursprung bewegen. Bild 12.5 zeigt eine Fläche für die Funktion

$$y = 2.5 - x_1^2 + x_2^2 \quad (13.26)$$

IMAGE5

Bild 12.5: Zielgröße y für einen rein quadratischen Modellansatz bei ungleichen Vorzeichen
In Richtung steigender Werte für x_1 wird die Zielgröße y kleiner, während sie für steigende Werte von x_2 ansteigt.

Aufbereitung des Datensatzes

Zur Berechnung der Regressionsfunktion muss die Datenmatrix \mathbf{X} erweitert werden. Durch die elementweise Multiplikation zweier Vektoren \underline{x}_j und \underline{x}_k entsteht ein Vektor \underline{x}_m , der Wechselwirkungen der beiden Größen repäsentiert.

$$x_{nm} = x_{nj} \cdot x_{nk} \quad (13.27)$$

Für quadratische Regressionsterme wird ein Vektor \underline{x}_k elementweise quadriert. Es entsteht ein neuer Vektor \underline{x}_m .

$$x_{nm} = x_{nk}^2 \quad (13.28)$$

Beispiel: Feuchtesensor

Als Beispiel für eine multivariate Regression mit vollquadratischem Ansatz wird die Vermessung der Kapazität eines Feuchtesensors betrachtet. Die aufgenommenen Messergebnisse zeigt Tabelle 13.2. Die Kapazität des Feuchtesensors soll mit einer Regressionsfunktion beschrieben werden. Eingangsgrößen sind die relative Feuchte rF und die Temperatur T , Zielgröße ist die Sensorkapazität C_p .

$$y = C_p = f(rF, T) \quad (13.29)$$

Tabelle 13.2: Urliste einer Vermessung von Feuchtesensoren

Nr.	Kapazität C_p pF	Einfügen Konstante	Temperatur	rel. Feuchte rF %	Neu generierte Datenspalten für vollquadratischen Ansatz		
			T / °C		x_1	x_2	$x_1 \cdot x_2$
1	179.3	1	90.0	9.5	855	8100	90.25
2	182.6	1	90.0	19.8	1782	8100	392.04
3	182.6	1	44.9	17.8	799.22	2016.01	316.84
4	185.6	1	90.1	29.6	2666.96	8118.01	876.16
5	185.6	1	45.2	27.2	1229.44	2043.04	739.84
6	189.8	1	89.7	40.6	3641.82	8046.09	1648.36
7	189.0	1	45.6	36.8	1678.08	2079.36	1354.24
8	188.2	1	22.8	36.7	836.76	519.84	1346.89
9	192.6	1	89.6	50.1	4488.96	8028.16	2510.01
10	192.1	1	45.3	47.4	2147.22	2052.09	2246.76
11	190.9	1	22.7	46.4	1053.28	515.29	2152.96
12	194.6	1	90.0	60.2	5418	8100	3624.04
13	195.2	1	45.2	57.8	2612.56	2043.04	3340.84
14	194.5	1	22.6	57.7	1304.02	510.76	3329.29
15	197.8	1	90.0	70.3	6327	8100	4942.09
16	198.6	1	45.2	68.5	3096.2	2043.04	4692.25
17	198.2	1	22.8	69.1	1575.48	519.84	4774.81
18	201.6	1	89.9	80.1	7200.99	8082.01	6416.01
19	202.6	1	45.2	79.5	3593.4	2043.04	6320.25
20	201.2	1	22.8	78.6	1792.08	519.84	6177.96
21	201.5	1	9.6	75.9	728.64	92.16	5760.81

Um sich einen Eindruck über die Daten zu verschaffen, werden die Stützstellen, an denen die Werte aufgenommen wurden, in einer Ebene aus relativer Feuchte und Temperatur in Bild 12.6 grafisch dargestellt.

IMAGE6

Bild 12.6: Darstellung der Stützstellen, an denen die Messwerte für die Kapazität C_p eines Feuchtesensors gemessen wurde

Der Bereich geringer relativer Feuchte rF und kleiner Temperaturen T wurde bei der Messung ausgespart, da das eingesetzte Messgerät in diesem Bereich keine ausreichende Messfähigkeit aufweist. Bild 12.7 Bild 12.7 stellt die einzelnen Messwerte als dreidimensionales Streudiagramm dar.

IMAGE7

Bild 12.7: Kapazität C_p eines Feuchtesensors als Funktion der Temperatur T und relativen Feuchte rF als Streudiagramm

Die Daten weisen ein weitgehend lineares Verhalten auf. Die Kapazität steigt in Richtung hoher Feuchte an. Das vollquadratische Modell wird angesetzt, um Linearitätsfehler bewerten zu können. Dazu wird jede Variable bis zur zweiten Potenz sowie die einfache Wechselwirkung zwischen relativer Feuchte und Temperatur berücksichtigt. Dadurch ergibt sich als Vektor der Eingangsgrößen zu

$$\underline{x}_n^T = (1 \ rF_n \ T_n \ rF_n \cdot T_n \ rF_n^2 \ T_n^2) \quad (13.30)$$

Die dabei entstehende Erweiterung der Datenmatrix ist in Tabelle 12.2 bereits eingetragen. Mit dem Datensatz wird die Regressionsgleichung berechnet. Für die Koeffizienten b_m ergibt sich der Vektor

$$\underline{b} = (\mathbf{X}^T \cdot \mathbf{X})^{-1} \cdot \mathbf{X}^T \cdot \underline{y} = \begin{pmatrix} 175.2 \\ 0.327 \\ 0.059 \\ -0.246 \cdot 10^{-3} \\ 0.015 \cdot 10^{-3} \\ -0.476 \cdot 10^{-3} \end{pmatrix} \quad (13.31)$$

Die Regressionsgleichung lautet damit

$$C_P = 175.2 + 0.327 \cdot rF + 0.059 \cdot T - 0.246 \cdot 10^{-3} \cdot rF \cdot T + 0.015 \cdot 10^{-3} \cdot rF^2 - 0.476 \cdot 10^{-3} \cdot T^2 \quad (13.32)$$

Das Regressionsergebnis ist in Bild 12.8 für den Temperaturbereich von 10...90°C und den Feuchtebereich von 10...90%rF dargestellt.

IMAGE8

Bild 12.8: Darstellung des Ergebnisses der Regression für einen Temperaturbereich von 10...90°C und einen Feuchtebereich von 10...90%rF

13.1.5 Lösbarkeit

<<< Wird später ergänzt ...

- Regression kann mit einer Taylorreihe verglichen werden
- Koeffizient vor quadratischem Term ist zweite Ableitung, $f'' \cdot \Delta x^2$, deshalb muss zweite Ableitung für ein Minimum > 0 sein
- Umsetzung im mehrdimensionalen Fall führt zur sogenannten Hesse-Matrix, sie muss eine Bedingung ähnlich wie im eindimensionalen Fall erfüllen: Hesse-Matrix muss positiv definit sein
- Kriterium ist aber eher theoretischer Natur, weil die Funktion an sich nicht bekannt ist ...
- Konkrete Bewertungen über Kondition der Matrix = Korrektheit der Regression

>>>

13.2 Statistische Bewertung der Regressionsparameter

Die Schätzung der Regressionsparametern β_m mit den Parametern b_m basiert auf den Messwerten y_n . Sie weisen einen zufälligen Messfehler auf und sind damit Zufallsgrößen. Aus diesem Grund sind auch die mit den Messwerten bestimmten Parameter b_m Zufallsgrößen.

13.2.1 Modell zur statistischen Bewertung der Regression

Die statistische Bewertung beruht auf dem Modellansatz in Gleichung (13.1). Er kann in Vektorschreibweise dargestellt werden als

$$\underline{y} = X \cdot \underline{\beta} + \underline{\epsilon} \quad (13.33)$$

Die Messwerte unterliegen einem Messfehler e . Es wird davon ausgegangen, dass der Messfehler normalverteilt ist. Er weist einen Erwartungswert von

$$E(\underline{y} - \underline{\mu}_y) = E(\underline{\epsilon}) = \underline{0} \quad (13.34)$$

und eine Kovarianzmatrix von

$$E\left(\left(\underline{y} - \underline{\mu}_y\right) \cdot \left(\underline{y} - \underline{\mu}_y\right)^T\right) = E(\underline{\epsilon} \cdot \underline{\epsilon}^T) = I \cdot \sigma^2 \quad (13.35)$$

Dabei ist I die Einheitsmatrix, die nur auf der Hauptdiagonalen Einsen besitzt, alle anderen Werte sind null. Das bedeutet, dass der Messfehler bei allen Messungen konstante Varianz σ^2 besitzt. Alle Kovarianzen σ_{jk} sind null, die Messfehler sind voneinander unabhängig.

Zum Nachweis der Erwartungstreue werden die Erwartungswerte der Regressionsparameter berechnet.

$$\begin{aligned} E(\underline{b}) &= E\left(\left(X^T \cdot X\right)^{-1} \cdot X^T \cdot \underline{y}\right) = E\left(\left(X^T \cdot X\right)^{-1} \cdot X^T \cdot X \cdot \underline{\beta} + \underline{\epsilon}\right) \\ &= E(\underline{\beta}) + \left(X^T \cdot X\right)^{-1} \cdot X^T \cdot E(\underline{\epsilon}) = \underline{\beta} \end{aligned} \quad (13.36)$$

Die Schätzung ist demnach erwartungstreu. Die Kovarianzmatrix berechnet sich mit der Matrix

$$C = \left(X^T \cdot X\right)^{-1} \cdot X^T \quad (13.37)$$

zu

$$\begin{aligned} E\left(\left(\underline{b} - \underline{\beta}\right) \cdot \left(\underline{b} - \underline{\beta}\right)^T\right) &= E\left(\left(C \cdot \underline{y} - C \cdot \underline{\mu}_y\right) \cdot \left(C \cdot \underline{y} - C \cdot \underline{\mu}_y\right)^T\right) \\ &= C \cdot E\left(\left(\underline{y} - \underline{\mu}_y\right) \cdot \left(\underline{y} - \underline{\mu}_y\right)^T\right) \cdot C^T = C \cdot I \cdot \sigma^2 \cdot C^T = C \cdot C^T \cdot \sigma^2 \end{aligned} \quad (13.38)$$

Die Streuung der Koeffizienten ist demnach proportional zur Streuung der Messung. Durch Einsetzen von Gleichung (13.37) ergibt sich wegen

$$C \cdot C^T = \left(\left(X^T \cdot X\right)^{-1} \cdot X^T\right) \cdot \left(\left(X^T \cdot X\right)^{-1} \cdot X^T\right)^T = \left(X^T \cdot X\right)^{-1} \cdot X^T \cdot X \cdot \left(X^T \cdot X\right)^{-1} = \left(X^T \cdot X\right)^{-1} \quad (13.39)$$

der Ausdruck für die Kovarianzmatrix der Regressionsparameter

$$E\left(\left(\underline{b} - \underline{\beta}\right) \cdot \left(\underline{b} - \underline{\beta}\right)^T\right) = C \cdot C^T \cdot \sigma^2 = \left(X^T \cdot X\right)^{-1} \cdot \sigma^2 \quad (13.40)$$

Dabei ist die Varianz der Messung σ^2 unbekannt und muss ebenfalls geschätzt werden. Die Abschätzung erfolgt über die Summe der Fehlerquadrate als Stichprobenvarianz. Aus N Messungen sind $M + 1$ Parameter zu bestimmen. Damit liegen $N - M - 1$ Freiheitsgrade vor, und die Stichprobenvarianz berechnet sich zu

$$\begin{aligned}\sigma^2 \approx s^2 &= \frac{1}{N-M-1} \cdot \sum_{n=1}^N r_n^2 = \frac{1}{N-M-1} \cdot (\underline{y} - X \cdot \underline{b})^T \cdot (\underline{y} - X \cdot \underline{b}) \\ &= \frac{1}{N-M-1} \cdot (\underline{y}^T - \underline{b}^T \cdot X^T) \cdot (\underline{y} - X \cdot \underline{b}) \\ &= \frac{1}{N-M-1} \cdot (\underline{y}^T \cdot \underline{y} - \underline{y}^T \cdot X \cdot \underline{b} - \underline{b}^T \cdot X^T \cdot \underline{y} + \underline{b}^T \cdot X^T \cdot X \cdot \underline{b}) \\ &= \frac{1}{N-M-1} \cdot (\underline{y}^T \cdot \underline{y} - \underline{y}^T \cdot X \cdot \underline{b} - \underline{b}^T \cdot X^T \cdot (\underline{y} - X \cdot \underline{b})) \\ &= \frac{1}{N-M-1} \cdot (\underline{y}^T \cdot \underline{y} - \underline{y}^T \cdot X \cdot \underline{b})\end{aligned}\tag{13.41}$$

Die Summanden in der Klammer sind skalare Größen. Damit gilt

$$\underline{y}^T \cdot X \cdot \underline{b} = \underline{b}^T \cdot X^T \cdot \underline{y}\tag{13.42}$$

und die Varianz der Messgröße kann abgeschätzt werden mit

$$\sigma^2 \approx s^2 = \frac{1}{N-M-1} \cdot (\underline{y}^T \cdot \underline{y} - \underline{b}^T \cdot X^T \cdot \underline{y})\tag{13.43}$$

13.2.2 Bewertung der Regressionsparameter b_m

Mit den Vorüberlegungen in Abschnitt 12.2.1 zur Kovarianzmatrix und zur Varianz der Messgröße können zur Bewertung der Regressionskoeffizienten b_m Zufallsvariablen aufgebaut werden. Mit k_m^2 als Element (m,m) der Matrix $(\mathbf{X}^T \cdot \mathbf{X})^{-1}$ sind die Zufallsvariablen

$$t_m = \frac{b_m - \beta_m}{k_m \cdot s}\tag{13.44}$$

t-verteilt mit $N - M - 1$ Freiheitsgraden. Die Zufallsvariablen t werden genutzt, um den Konfidenzbereich abzuschätzen und die Signifikanz zu bewerten.

Konfidenzintervall für die Regressionskoeffizienten β_m

Zur Berechnung des Konfidenzintervalls des Regressionskoeffizienten β_m wird die Zufallsvariable t_m verwendet. Nach den Ausführungen in Kapitel 5 berechnet sich der Konfidenzbereich aus der Wahrscheinlichkeit

$$P(c_1 < t_m \leq c_2) = F(c_2) - F(c_1) = \gamma\tag{13.45}$$

Durch die Symmetrie des Konfidenzbereichs ergeben sich die Konstanten c_1 und c_2 zu

$$c_1 = F^{-1}\left(\frac{1-\gamma}{2}\right)\tag{13.46}$$

und

$$c_2 = F^{-1}\left(\frac{1+\gamma}{2}\right)\tag{13.47}$$

Durch Umformungen von Gleichung (13.41) ergibt sich ein Ausdruck für den Konfidenzbereich des Regressionskoeffizienten β_m von

$$\gamma = P(c_1 < t_m \leq c_2) = P\left(c_1 < \frac{b_m - \beta_m}{k_m \cdot s} \leq c_2\right) = P(b_m - c_2 \cdot k_m \cdot s \leq \beta_m < b_m - c_1 \cdot k_m \cdot s)\tag{13.48}$$

Die Berechnung wird in Tabelle 13.3 zusammengefasst.

Tabelle 13.3: Vorgehen zur Bestimmung des Konfidenzbereichs für den Regressionskoeffizienten β_m

Nr.	Prozessschritt
1	Wahl einer Konfidenzzahl γ
2	Bestimmung der zugehörigen Parameter c_1 und c_2 aus der inversen t-Verteilung mit $N + M - 1$ Freiheitsgraden $c_1 = F^{-1}\left(\frac{1-\gamma}{2}\right) \quad \text{und} \quad c_2 = F^{-1}\left(\frac{1+\gamma}{2}\right)$
3	Berechnung der Regressionskoeffizienten $\underline{b} = (\underline{X}^T \cdot \underline{X})^{-1} \cdot \underline{X}^T \cdot \underline{y}$
4	Abschätzung der Kovarianzmatrix $(\underline{X}^T \cdot \underline{X})^{-1} \cdot \frac{1}{N - M - 1} \cdot (\underline{y}^T \cdot \underline{y} - \underline{b}^T \cdot \underline{X}^T \cdot \underline{y}) = \begin{pmatrix} k_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & k_M \end{pmatrix}$
5	Bestimmung des Konfidenzintervalls $b_m - c_2 \cdot k_m \cdot s \leq \beta_m < b_m - c_1 \cdot k_m \cdot s$

Beispiel: Feuchtesensor

Die Berechnung der Konfidenzbereiche wird an dem Beispiel des Feuchtesensors angewendet. Für die Koeffizienten b_m ergeben sich für $\gamma = 95\%$ die in Tabelle 13.4 gezeigten Konfidenzbereiche.

Tabelle 13.4: Konfidenzbereiche der Regressionskoeffizienten für die Regression der Kapazität eines Feuchtesensors als Funktion der Temperatur und relativen Feuchte, Konfidenzzahl für $\gamma = 95\%$

Name	physikalische Größe	Regressionskoeffizient b_m	Standardabweichung S_{bm}	Untere Grenze	Obere Grenze
b_0	Konstante	175.2	1.4454	172.1267	178.2881
b_1	rF	0.3279	0.0351	0.2531	0.4026
b_2	T	0.0592	0.0316	-0.008	0.1266
b_3	rF·T	$-0.246 \cdot 10^{-3}$	$0.212 \cdot 10^{-3}$	$-0.6992 \cdot 10^{-3}$	$0.2080 \cdot 10^{-3}$
b_4	rF ²	$-0.015 \cdot 10^{-3}$	$0.279 \cdot 10^{-3}$	$-0.5800 \cdot 10^{-3}$	$0.6111 \cdot 10^{-3}$
b_5	T ²	$-0.47 \cdot 10^{-3}$	$0.215 \cdot 10^{-3}$	$-0.9355 \cdot 10^{-3}$	$-0.01813 \cdot 10^{-3}$

Die Koeffizienten b_0 , b_1 und b_5 schließen den Wert null nicht ein, es wird sich zeigen, dass diese Koeffizienten für die Regression signifikant sind.

Test der Regressionskoeffizienten β_m auf Signifikanz

Zur Reduzierung der Komplexität von Regressionsfunktionen werden die berechneten Koeffizienten Signifikanz geprüft. Es existieren verschiedene Verfahren zur Bewertung der Signifikanz von Regressionskoeffizienten, von denen zwei dargestellt werden sollen. Dies sind der t-Test und die Analyse des Konfidenzbereichs.

Zur Bewertung der Signifikanz des Regressionskoeffizienten β_m wird die Nullhypothese aufgestellt, dass der Korrelationskoeffizient β_m einem Wert $\beta = 0$ entspricht. Trifft diese Hypothese zu, ist die Zielgröße y nicht von der Eingangsgröße x_m abhängig. Wird die Nullhypothese auf Basis der vorliegenden Stichprobe abgelehnt, kann davon ausgegangen werden, dass die Zielgröße y signifikant von der betrachteten Eingangsgröße x_m abhängt.

Damit ein über die Stichprobe geschätzter Regressionskoeffizient b_m mit einer spezifizierten Wahrscheinlichkeit zu der t-Verteilung aus Gleichung (13.39) gehört, muss dieser in dem Intervall $\beta_{mC1} < b_m \leq \beta_{mC2}$ liegen. Wird die Wahrscheinlichkeit dafür mit γ bezeichnet, gilt die Gleichung

$$P(\beta_{mC1} < b_m \leq \beta_{mC2}) = \gamma = 1 - \alpha \quad (13.49)$$

Mit der Verteilung aus Gleichung (13.39) wird die Wahrscheinlichkeit γ , mit der die Variable t innerhalb des Intervalls $c_1 \dots c_2$ liegt, definiert als

$$\gamma = P(c_1 < t_m \leq c_2) = F(c_2) - F(c_1) \quad (13.50)$$

Bei Annahme eines symmetrischen Tests ergeben sich die Konstanten c_1 und c_2 aus den Bedingungen

$$F(c_1) = \frac{1 - \gamma}{2} = \frac{\alpha}{2} \quad (13.51)$$

und

$$F(c_2) = 1 - \frac{1 - \gamma}{2} = 1 - \frac{\alpha}{2} \quad (13.52)$$

Auflösen nach c_1 und c_2 führt zu

$$c_1 = F^{-1}\left(\frac{\alpha}{2}\right) \quad (13.53)$$

und

$$c_2 = F^{-1}\left(1 - \frac{\alpha}{2}\right) \quad (13.54)$$

Durch Umformungen von Gleichung (13.39) und (13.49) ergibt sich ein Ausdruck für den Annahmebereich der Nullhypothese, nämlich dass der geschätzte Regressionskoeffizient b_m mit einer spezifizierten Wahrscheinlichkeit γ zu der angenommenen t-Verteilung gehört.

$$1 - \alpha = P(c_1 < t_m \leq c_2) = P\left(c_1 < \frac{b_m - \beta_m}{k_m \cdot s} \leq c_2\right) = P(c_1 \cdot k_m \cdot s < b_m \leq c_2 \cdot k_m \cdot s) \quad (13.55)$$

Alternativ kann, wie in Kapitel 5 gezeigt wird, eine Unterschreitungswahrscheinlichkeit p der Prüfgröße b_m bestimmt werden und mit dem Signifikanzniveau α verglichen werden. Bei Hypothesentests mit beidseitigem Verwerfungsbereich $\beta_m \neq 0$ müssen für eine Annahme der Nullhypothese die Bedingungen

$$p = F(t_m) > \frac{\alpha}{2} \quad (13.56)$$

und

$$p = F(t_m) \leq 1 - \frac{\alpha}{2} \quad (13.57)$$

erfüllt werden. Damit lässt sich der Test mit der Hypothese $\beta_m = 0$ und der Alternative $\beta_m \neq 0$ in folgenden Prozessschritten zusammenfassen.

Tabelle 13.5: Test der Hypothese $\beta_m = 0$ gegen $\beta_m \neq 0$ für den Regressionskoeffizienten β_m einer linearen Regression

Nr.	Prozessschritt	
1	Wahl eines Signifikanzniveaus α	
2	Bestimmung der zugehörigen Parameter c_1 und c_2 aus der inversen t-Verteilung mit $N - M - 1$ Freiheitsgraden $c_1 = F^{-1}\left(\frac{\alpha}{2}\right)$ und $c_2 = F^{-1}\left(1 - \frac{\alpha}{2}\right)$	
3	Berechnung der Regressionskoeffizienten $\underline{b} = (\mathbf{X}^T \cdot \mathbf{X})^{-1} \cdot \mathbf{X}^T \cdot \underline{y}$	
4	Abschätzung der Kovarianzmatrix $(\mathbf{X}^T \cdot \mathbf{X})^{-1} \cdot \frac{1}{N - M - 1} \cdot (\underline{y}^T \cdot \underline{y} - \underline{b}^T \cdot \mathbf{X}^T \cdot \underline{y}) = \begin{pmatrix} k_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & k_M \end{pmatrix}$	
5	Bestimmung des Annahmebereichs $c_1 \cdot k_m \cdot s < b_m \leq c_2 \cdot k_m \cdot s$	Berechnung des p-Wertes mit der t-Verteilung mit $N - 2$ Freiheitsgraden $p = F\left(\frac{b_m}{k_m \cdot s}\right)$
6	Für $\beta_{mC1} < b_m \leq \beta_{mC2}$ wird die Hypothese angenommen, für $b_m \leq \beta_{mC1}$ oder $b_m > \beta_{mC2}$ wird die Hypothese verworfen	Für $\alpha/2 < p \leq 1 - \alpha/2$ wird die Hypothese angenommen, für $p \leq \alpha/2$ und $p > 1 - \alpha/2$ wird die Hypothese verworfen

Beispiel: Feuchtesensor

Die Signifikanzbewertung wird anhand der Daten des Feuchtesensors durchgeführt. Für die Koeffizienten b_m die Wahrscheinlichkeit für die Nullhypothese bestimmt, dass der Wert b_m null und damit der Einfluss des entsprechenden Terms nicht signifikant ist.

Tabelle 13.6: Bewertung der Regressionskoeffizienten für die Regression der Kapazität eines Feuchtesensors als Funktion der Temperatur und relativen Feuchte

Name	physikalische Größe	Regressionskoeffizient b_m	Standardabweichung S_{bm}	p-Wert	Signifikanz
b_0	Konstante	175.2	1.4454	0	ja
b_1	rF	0.3279	0.0351	$0.12 \cdot 10^{-6}$	ja
b_2	T	0.0592	0.0316	0.080	nein
b_3	rF·T	$-0.246 \cdot 10^{-3}$	$0.212 \cdot 10^{-3}$	0.266	nein
b_4	rF ²	$-0.015 \cdot 10^{-3}$	$0.279 \cdot 10^{-3}$	0.956	nein
b_5	T ²	$-0.47 \cdot 10^{-3}$	$0.215 \cdot 10^{-3}$	0.042	ja

Es zeigt sich, dass der Feuchtesensor neben einem Offset-Wert eine lineare Empfindlichkeit zu relativen Luftfeuchte aufweist. Diese Empfindlichkeit ist der gewünschte Messeffekt des Sensors. Die Nichtlinearität rF^2 und der Temperatureinfluss auf die Empfindlichkeit sind nicht signifikant und damit statistisch gesehen vernachlässigbar. Die Temperatur geht nicht linear, sondern quadratisch in das Messergebnis ein.

Ein Vergleich der Signifikanzbewertung mit dem Konfidenzbereich in Tabelle 13.9 12.9 zeigt, dass ein Term nicht signifikant ist, wenn der Konfidenzbereich des Regressionskoeffizienten den Wert null einschließt. Andernfalls ist er signifikant.

13.2.3 Konfidenzbereich des Erwartungswertes der Regressionsfunktion

Die geschätzte Reaktion y_0 an einer Stelle \underline{x}_0^T ergibt sich aus der Gleichung

$$y(\underline{x}_0^T) = \underline{x}_0^T \cdot \underline{b} \quad (13.58)$$

Dabei werden die Parameter b_m auf Basis des vorliegenden Datensatzes geschätzt, sie sind damit selber Zufallsvariablen. Die Schätzung der Regressionsparameter β_m ist erwartungstreu, sodass sich der Erwartungswert von

$$E(y_0(\underline{x}_0^T)) = E(\underline{x}_0^T \cdot \underline{b}) = \underline{x}_0^T \cdot E(\underline{b}) = \underline{x}_0^T \cdot \underline{\beta} = \mu_y(\underline{x}_0^T) \quad (13.59)$$

noindent ergibt. Die Schätzung des Mittelwertes ist demnach ebenfalls erwartungstreu. Die Varianz der Größe berechnet sich zu

$$\begin{aligned} \sigma_{y_0}^2 &= E\left((y(\underline{x}_0^T) - \mu_y(\underline{x}_0^T)) \cdot (y(\underline{x}_0^T) - \mu_y(\underline{x}_0^T))^T\right) = E\left((\underline{x}_0^T \cdot \underline{b} - \underline{x}_0^T \cdot \underline{\beta}) \cdot (\underline{x}_0^T \cdot \underline{b} - \underline{x}_0^T \cdot \underline{\beta})^T\right) \\ &= \underline{x}_0^T \cdot E\left((\underline{b} - \underline{\beta}) \cdot (\underline{b}^T - \underline{\beta}^T)\right) \cdot \underline{x}_0 \end{aligned} \quad (13.60)$$

Die Kovarianzmatrix der Regressionsparameter kann mit Gleichung (13.40) dargestellt werden als

$$E\left((\underline{b} - \underline{\beta}) \cdot (\underline{b} - \underline{\beta})^T\right) = (X^T \cdot X)^{-1} \cdot \sigma^2 \quad (13.61)$$

Damit ergibt sich

$$\sigma_{y_0}^2 = \underline{x}_0^T \cdot (X^T \cdot X)^{-1} \cdot \underline{x}_0 \cdot \sigma^2 \quad (13.62)$$

Die Varianz σ^2 wird nach Gleichung (13.43) abgeschätzt mit

$$\sigma^2 \approx s^2 = \frac{1}{N - M - 1} \cdot (\underline{y}^T \cdot \underline{y} - \underline{b}^T \cdot X^T \cdot \underline{y}) \quad (13.63)$$

Die geschätzte Varianz von \widehat{y}_0 an der Stelle \underline{x}_0^T lautet damit

$$s_{y_0}^2 = \frac{1}{N - M - 1} \cdot (\underline{y}^T \cdot \underline{y} - \underline{b}^T \cdot X^T \cdot \underline{y}) \cdot \underline{x}_0^T \cdot (X^T \cdot X)^{-1} \cdot \underline{x}_0 \quad (13.64)$$

Damit ist die Größe

$$t = \frac{y(\underline{x}_0^T) - \mu_y(\underline{x}_0^T)}{s_{y_0}} \quad (13.65)$$

t-verteilt mit $N - M - 1$ Freiheitsgraden. Die Zufallsvariable t wird zur Bestimmung des Konfidenzbereichs für den Erwartungswert $\mu_y(\underline{x}_0^T)$ verwendet. Er berechnet sich aus der Wahrscheinlichkeit

$$P(c_1 < t \leq c_2) = F(c_2) - F(c_1) = \gamma \quad (13.66)$$

Durch die Symmetrie des Konfidenzbereichs ergeben sich die Konstanten c_1 und c_2 zu

$$c_1 = F^{-1}\left(\frac{1 - \gamma}{2}\right) \quad (13.67)$$

und

$$c_2 = F^{-1}\left(\frac{1 + \gamma}{2}\right) \quad (13.68)$$

Durch Umformungen ergibt sich ein Ausdruck für den Konfidenzbereich des Mittelwertes.

$$\begin{aligned} \gamma = P(c_1 < t \leq c_2) &= P\left(c_1 < \frac{y(\underline{x}_0^T) - \mu_y(\underline{x}_0^T)}{s_{y_0}} \leq c_2\right) \\ &= P\left(y(\underline{x}_0^T) - c_2 \cdot s_{y_0} < \mu_y(\underline{x}_0^T) \leq y(\underline{x}_0^T) - c_1 \cdot s_{y_0}\right) \end{aligned} \quad (13.69)$$

Die Berechnung wird in Tabelle 13.7 zusammengefasst.

Tabelle 13.7: Vorgehen zur Bestimmung des Konfidenzbereichs für den Erwartungswert $\mu_y(\underline{x}_0^T)$

Nr.	Prozessschritt
1	Wahl einer Konfidenzzahl γ
2	Bestimmung der zugehörigen Parameter c_1 und c_2 aus der inversen t-Verteilung mit $N + M - 1$ Freiheitsgraden $c_1 = F^{-1}\left(\frac{1-\gamma}{2}\right) \quad \text{und} \quad c_2 = F^{-1}\left(\frac{1+\gamma}{2}\right)$
3	Berechnung der Regressionskoeffizienten $\underline{b} = (\mathbf{X}^T \cdot \mathbf{X})^{-1} \cdot \mathbf{X}^T \cdot \underline{y}$
4	Abschätzung der Varianz an der Stelle \underline{x}_0^T $s_{y_0}^2 = \frac{1}{N - M - 1} \cdot (\underline{y}^T \cdot \underline{y} - \underline{b}^T \cdot \mathbf{X}^T \cdot \underline{y}) \cdot \underline{x}_0^T \cdot (\mathbf{X}^T \cdot \mathbf{X})^{-1} \cdot \underline{x}_0$
5	Bestimmung des Konfidenzintervalls an der Stelle \underline{x}_0^T $y(\underline{x}_0^T) - c_2 \cdot s_{y_0} < \mu_y(\underline{x}_0^T) \leq y(\underline{x}_0^T) + c_1 \cdot s_{y_0}$

Beispiel: Feuchtesensor

Das Verfahren wird für das Beispiel Feuchtesensor angewendet. Für die Stützpunkte des berechneten Kennfeldes werden untere und obere Grenze des Konfidenzbereiches berechnet. Das Ergebnis ist in Bild 12.9 dargestellt.

IMAGE 9

Bild 12.9: Darstellung der Regressionfunktion mit Konfidenzbereich (Kennfeld über den gesamten Datenbereich) Schnitt durch die Flächen bei einer konstanten relativen Feuchte $rF = 50\%$

Der Konfidenzbereich umschließt die Regressionsfläche. Ähnlich wie im eindimensionalen Fall weitet sich der Konfidenzbereich nach außen auf. Die Berechnung wurde mit folgendem MATLAB-Programm durchgeführt.

```

1 % Berechnung der Regressionsparameter über Matrizengleichung
2 X = [ones(size(T)) RF T RF.*T RF.\mathrm{\wedge\$2 T.\$}\mathrm{\wedge\$2];
3 bmat = inv(X'*X)*X'*Cp;
4 Cpres = Cp - X*bmat;
5
6 % Erstellen einer Matrix mit den neuen Beobachtungen
7 xneu = [];
8 for rf = 10:10:90
9     for t = 10:10:90
10         xneu = [xneu; [1 rf t rf*t rf\$mathrm{\wedge\$2
11 t\$mathrm{\wedge\$2]];
12     end;
13 end;
14
15 % Berechnung des Erwartungswertes für gegebenes alpha
16 gamma = 0.95;
17 FG = length(X) - length(bmat);
18 CpEmin = [];

```

```

19 CpEmax = [] ;
20 for n = 1:length(xneu)
21 x0 = xneu(n,:)' ;
22 CpEmin = [CpEmin bmat'*x0 + ...
23 tinv((1-gamma)/2,FG)*sqrt(Cpres'*Cpres/FG)*sqrt(x0'*inv(X'*X)*x0)] ;
24 CpEmax = [CpEmax bmat'*x0 + ...
25 tinv((1+gamma)/2,FG)*sqrt(Cpres'*Cpres/FG)*sqrt(x0'*inv(X'*X)*x0)] ;
26 end ;

```

13.2.4 Prognosebereich für zukünftige Stichprobenwerte

Um zukünftige Werte einer Stichprobe vorhersagen zu können, ist es erforderlich, das Prognoseintervall für die Reaktion y_0 an einer Stelle x_0 abschätzen zu können. Die Lage eines zukünftigen Wertes ergibt sich aus dem Mittelwert und dem überlagerten Messfehler.

$$\hat{y}(\underline{x}_0^T) = \underline{x}_0^T \cdot \underline{b} + e \quad (13.70)$$

Der Messfehler e ist mittelwertsfrei, sodass sich der Erwartungswert von

$$E(\hat{y}(\underline{x}_0^T)) = E(\underline{x}_0^T \cdot \underline{b} + e) = \underline{x}_0^T \cdot E(\underline{b}) + E(e) = \underline{x}_0^T \cdot \underline{\beta} \quad (13.71)$$

ergibt. Die Varianz der Größe berechnet sich zu

$$\begin{aligned} \sigma_{\hat{y}_0}^2 &= E\left(\left(\hat{y}(\underline{x}_0^T) - \underline{x}_0^T \cdot \underline{\beta}\right) \cdot \left(\hat{y}(\underline{x}_0^T) - \underline{x}_0^T \cdot \underline{\beta}\right)^T\right) \\ &= E\left(\left(\underline{x}_0^T \cdot \underline{b} + e - \underline{x}_0^T \cdot \underline{\beta}\right) \cdot \left(\underline{x}_0^T \cdot \underline{b} + e - \underline{x}_0^T \cdot \underline{\beta}\right)^T\right) \\ &= \underline{x}_0^T \cdot E\left(\left(\underline{b} - \underline{\beta}\right) \cdot \left(\underline{b}^T - \underline{\beta}^T\right)\right) \cdot \underline{x}_0 + E(e^2) = \underline{x}_0^T \cdot \left(X^T \cdot X\right)^{-1} \cdot \underline{x}_0 \cdot \sigma^2 + \sigma^2 \end{aligned} \quad (13.72)$$

Die Varianz σ^2 wird nach Gleichung (13.73) abgeschätzt mit

$$\sigma^2 \approx s^2 = \frac{1}{N - M - 1} \cdot (\underline{y}^T \cdot \underline{y} - \underline{b}^T \cdot X^T \cdot \underline{y}) \quad (13.73)$$

Die geschätzte Varianz von \hat{y}_0 an der Stelle \underline{x}_0^T lautet damit

$$s_{\hat{y}_0}^2 = \frac{1}{N - M - 1} \cdot (\underline{y}^T \cdot \underline{y} - \underline{b}^T \cdot X^T \cdot \underline{y}) \cdot \left(\underline{x}_0^T \cdot \left(X^T \cdot X\right)^{-1} \cdot \underline{x}_0 + 1\right) \quad (13.74)$$

Damit ist die Größe

$$t = \frac{\hat{y}(\underline{x}_0^T) - \underline{x}_0^T \cdot \underline{b}}{s_{\hat{y}_0}} \quad (13.75)$$

t-verteilt mit $N - M - 1$ Freiheitsgraden. Der Zähler der Variable stellt den Abstand des Schätzwertes \hat{y}_0 an der Stelle \underline{x}_0^T zu der Regressionsfunktion dar, sodass diese Zufallsvariable zur Bestimmung des Konfidenzintervalls verwendet werden kann.

$$P(c_1 < t \leq c_2) = F(c_2) - F(c_1) = \gamma \quad (13.76)$$

Durch die Symmetrie des Konfidenzbereichs ergeben sich die Konstanten c_1 und c_2 zu

$$c_1 = F^{-1}\left(\frac{1-\gamma}{2}\right) \quad (13.77)$$

und

$$c_2 = F^{-1}\left(\frac{1-\gamma}{2}\right) \quad (13.78)$$

Durch Umformungen ergibt sich ein Ausdruck für den Konfidenzbereich des Mittelwertes.

$$\begin{aligned}\gamma &= P(c_1 < t \leq c_2) = P\left(c_1 < \frac{\hat{y}(\underline{x}_0^T) - \underline{x}_0^T \cdot \underline{b}}{s_{\hat{y}_0}} \leq c_2\right) \\ &= P\left(\underline{x}_0^T \cdot \underline{b} + c_1 \cdot s_{\hat{y}_0} < \hat{y}(\underline{x}_0^T) \leq \underline{x}_0^T \cdot \underline{b} + c_2 \cdot s_{\hat{y}_0}\right)\end{aligned}\quad (13.79)$$

Das Verfahren ist in Tabelle 13.8 zusammengefasst.

Tabelle 13.8: Vorgehen zur Bestimmung des Prognosebereich

Nr.	Prozessschritt
1	Wahl einer Konfidenzzahl γ
2	Bestimmung der zugehörigen Parameter c_1 und c_2 aus der inversen t-Verteilung mit $N + M - 1$ Freiheitsgraden $c_1 = F^{-1}\left(\frac{1 - \gamma}{2}\right) \quad \text{und} \quad c_2 = F^{-1}\left(\frac{1 + \gamma}{2}\right)$
3	Berechnung der Regressionskoeffizienten $\underline{b} = (\mathbf{X}^T \cdot \mathbf{X})^{-1} \cdot \mathbf{X}^T \cdot \underline{y}$
4	Abschätzung der Varianz an der Stelle \underline{x}_0^T $s_{y_0}^2 = \frac{1}{N - M - 1} \cdot (\underline{y}^T \cdot \underline{y} - \underline{b}^T \cdot X^T \cdot \underline{y}) \cdot \underline{x}_0^T \cdot (X^T \cdot X)^{-1} \cdot \underline{x}_0$
5	Bestimmung des Konfidenzintervalls an der Stelle \underline{x}_0^T $y(\underline{x}_0^T) - c_2 \cdot s_{y_0} < \mu_y(\underline{x}_0^T) \leq y(\underline{x}_0^T) - c_1 \cdot s_{y_0}$

Beispiel: Feuchtesensor

Für die Stützpunkte des berechneten Kennfeldes werden neben dem Konfidenzbereich auch die untere und obere Grenze des Prognosebereichs berechnet. Das Ergebnis ist in Bild 12.10 zusammen mit dem Konfidenzbereich dargestellt.

IMAGE10

Bild 12.10: Darstellung der Regressionfunktion mit Prognose- und Konfidenzbereich
 a) Kennfeld über den gesamten Datenbereich
 b) Schnitt durch die Flächen bei einer konstanten relativen Feuchte $rF = 50\%$

Die Berechnung wurde mit folgendem MATLAB-Programm durchgeführt.

```

1 % Berechnung der Regressionsparameter über Matrizengleichung
2 X = [ones(size(T)) RF T RF.*T RF.\mathrm{\wedge\$2} T.\mathrm{\wedge\$2}];
3 bmat = inv(X'*X)*X'*Cp;
4 Cpres = Cp - X*bmat;
5
6 % Erstellen einer Matrix mit den neuen Beobachtungen
7 xneu = [];
8 for rf = 10:10:90
9     for t = 10:10:90
10        xneu = [xneu; [1 rf t rf*t rf.\mathrm{\wedge\$2} t.\mathrm{\wedge\$2}]];
11    end;
12 end;
13
14 % Berechnung des Erwartungswertes für gegebenes alpha
15 gamma = 0.95;
16 FG = length(X) - length(bmat);
17 CpEmin = [];
18 CpEmax = [];
19 for n = 1:length(xneu)
20    x0 = xneu(n,:)' ;
21    CpPmin = [CpPmin bmat'*x0 + ...
22               tinv((1-gamma)/2,FG)*sqrt(Cpres'*Cpres/FG)*sqrt(1+x0'*inv(X'*X)*x0)
23               ];
24    CpPmax = [CpPmax bmat'*x0 + ...
25               tinv((1+gamma)/2,FG)*sqrt(Cpres'*Cpres/FG)*sqrt(1+x0'*inv(X'*X)*x0) ];
26 end;

```

Bei dem Schnitt durch die Flächen bei konstanter relativer Feuchte wird deutlich, dass der Prognosebereich größer ist als der Konfidenzbereich. Dies ist auf die größere Varianz bei der Berechnung der Prognosewerte zurückzuführen. Ein Vergleich von Varianzen für Konfidenzbereich

$$s_{y_0}^2 = \frac{1}{N-M-1} \cdot (\underline{y}^T \cdot \underline{y} - \underline{b}^T \cdot X^T \cdot \underline{y}) \cdot \underline{x}_0^T \cdot (X^T \cdot X)^{-1} \cdot \underline{x}_0 \quad (13.80)$$

und Prognosebereich

$$s_{\hat{y}_0}^2 = \frac{1}{N-M-1} \cdot (\underline{y}^T \cdot \underline{y} - \underline{b}^T \cdot X^T \cdot \underline{y}) \cdot \left(\underline{x}_0^T \cdot (X^T \cdot X)^{-1} \cdot \underline{x}_0 + 1 \right) \quad (13.81)$$

zeigt, dass sich die beiden Varianzen bis auf den Summanden + 1 in der letzten Klammer entsprechen. Ein Vergleich der beiden Modelle für Erwartungswert

$$y(\underline{x}_0^T) = \underline{x}_0^T \cdot \underline{b} \quad (13.82)$$

und Prognosewert

$$\hat{y}(\underline{x}_0^T) = \underline{x}_0^T \cdot \underline{b} + e \quad (13.83)$$

zeigt, dass bei der Prognose die statistische Unsicherheit des einzelnen Wertes e für die Vergrößerung verantwortlich ist.

13.2.5 Bewertung der Regression mit Bestimmtheitsmaß und Reststreuungsanalyse

Zur Überprüfung des Modells werden die Residuen der Regression untersucht. Für das Beispiel des Feuchtesensors wird bewertet, ob die Reststreuung einen charakteristischen Verlauf aufweist und das Modell erweitert werden muss. Bild 12.11 stellt die Residuen als Funktion der Temperatur T und als Funktion der Feuchte rF dar. Die Residuen haben keinen charakteristischen Funktionsverlauf, das Modell muss deshalb nicht erweitert werden.

IMAGE11

Bild 12.11: Darstellung der Residuen als Funktion der Temperatur T und als Funktion der Feuchte rF

Zur Vereinfachung des Modells werden die Koeffizienten b_m untersucht. Dazu wird die Standardabweichung s_{bm} berechnet und die Wahrscheinlichkeit für die Nullhypothese bestimmt, dass der Wert b_m null und damit der Einfluss des entsprechenden Terms nicht signifikant ist.

Tabelle 13.9: Bewertung der Regressionskoeffizienten für die Regression der Kapazität eines Feuchtesensors als Funktion der Temperatur und relativen Feuchte

Name	physikalische Größe	Regressionskoeffizient b_m	Standardabweichung S_{bm}	p-Wert
b_0	Konstante	175.2	1.4454	0
b_1	rF	0.3279	0.0351	$0.12 \cdot 10^{-6}$
b_2	T	0.0592	0.0316	0.080
b_3	rF·T	$-0.246 \cdot 10^{-3}$	$0.212 \cdot 10^{-3}$	0.266
b_4	rF ²	$-0.015 \cdot 10^{-3}$	$0.279 \cdot 10^{-3}$	0.956
b_5	T ²	$-0.47 \cdot 10^{-3}$	$0.215 \cdot 10^{-3}$	0.042

Mithilfe der Wahrscheinlichkeit dafür, dass der Koeffizient b_m null ist, lässt sich für die Regressionskoeffizienten eine Signifikanzreihenfolge bestimmen. Der Koeffizient b_0 weist die größte Signifikanz auf und der Koeffizient b_4 die kleinste. Auf Basis dieser Signifikanz werden nacheinander die Koeffizienten b_4 , b_3 und b_2 eliminiert. Alle anderen Koeffizienten sind signifikant. Die Auswirkung auf das adjugierte Bestimmtheitsmaß ist in Bild 12.12 dargestellt.

IMAGE12

Bild 12.12: Darstellung des adjungierte Bestimmtheitsmaßes als Funktion des entfernten Regressionskoeffizienten (kumulativ)

Das adjugierte Bestimmtheitsmaß ändert sich durch das Weglassen der Regressionskoeffizienten nur wenig. Generell ist das adjugierte Bestimmtheitsmaß mit Werten größer 0.99 sehr hoch. Nach dieser Analyse lässt sich der Datensatz mit dem reduzierten Satz von Größen beschreiben, die in Tabelle 12.10 dargestellt sind.

Tabelle 12.10: Bewertung Regressionskoeffizienten für die Regression der Kapazität eines Feuchtesensors als Funktion der Temperatur und relativen Feuchte bei einem reduzierten Modell

Tabelle 13.10: Bewertung der Regressionskoeffizienten für die Regression der Kapazität eines Feuchtesensors als Funktion der Temperatur und relativen Feuchte

Name	physikalische Größe	Regressionskoeffizient b_m	Standardabweichung s_{bm}	p-Wert
b_0	Konstante	175.2	0.3379	0
b_1	rF	0.3116	$5.21 \cdot 10^{-3}$	0
b_5	T^2	$-0.084 \cdot 10^{-3}$	$0.033 \cdot 10^{-3}$	0.020

Zur Verifikation wird für das reduzierte Modell die Reststreuungsanalyse durchgeführt. Dazu werden zunächst die Residuen als Funktion der Temperatur und als Funktion der Feuchte dargestellt und mit den Residuen der ersten Regression verglichen.

IMAGE13

Bild 12.13: Darstellung der Residuen als Funktion der Temperatur und als Funktion der Feuchte für das voll quadratische Modell und das reduzierte Modell

Die Residuen des reduzierten Modells weichen von denen des voll quadratischen Modells nur geringfügig ab. Dadurch wird die These, dass die reduzierte Regression ausreichend genau ist, weiter gestützt. Als letzte Plausibilisierung der Messwerte wird der Box-Plot der Residuen eingesetzt, er ist in Bild 12.14Bild 12.14 dargestellt.

IMAGE14

Bild 12.14: Prüfung der Residuen für das reduzierte Modell mit einem Box-Plot

Es existieren keine Messwerte, die im Box-Plot als Ausreißer gekennzeichnet sind. Eine Auffälligkeit hinsichtlich der Stichprobe kann in diesem Fall deshalb nicht gefunden werden. Mit einem Hypothesentest kann die Hypothese, dass es sich bei den Residuen um eine Normalverteilung handelt, nicht verworfen werden.

13.3 Korrektheit der Regression und schlecht gestellte Probleme

<<< Wird später ergänzt ... <<<

13.3.1 Vorüberlegung zur Invertierbarkeit

Bei der Berechnung der Regressionsfunktionen wird ein Datensatz verwendet, bei dem N Messungen von M Eingangsgrößen und jeweils einer Zielgröße durchgeführt werden. Die Eingangswerte werden in der Matrix X

$$\underline{X} = \begin{pmatrix} 1 & \cdots & x_{1m} & \cdots & x_{1M} \\ \vdots & \ddots & \vdots & & \vdots \\ 1 & \cdots & x_{nm} & \cdots & x_{nM} \\ \vdots & & \vdots & \ddots & \vdots \\ 1 & \cdots & x_{Nm} & \cdots & x_{NM} \end{pmatrix} \quad (13.84)$$

und die Zielgrößen in einem Vektor y

$$\underline{y} = \begin{pmatrix} y_1 \\ \vdots \\ y_n \\ \vdots \\ y_N \end{pmatrix} \quad (13.85)$$

Ziel der Regression ist es, eine Funktion für den Zusammenhang zwischen den Eingangsgrößen \underline{x}^T und der Zielgrößen y anzugeben.

$$y(\underline{x}^T) = b_0 + b_1 \cdot x_1 + \dots + b_m \cdot x_m + \dots + b_M \cdot x_M \quad (13.86)$$

Die über die Bestimmungsgleichung

$$\underline{b} = (\underline{X}^T \cdot \underline{X})^{-1} \cdot \underline{X}^T \cdot \underline{y} \quad (13.87)$$

bestimmten Parameter b_m sind Schätzwerte für die Parameter β_m des zugrundeliegenden Prozesses.

$$y = \beta_0 + \beta_1 \cdot x_1 + \dots + \beta_m \cdot x_m + \dots + \beta_M \cdot x_M + e \quad (13.88)$$

Bei der Bestimmung der Regressionskoeffizienten b_m wird eine Invertierung vorgenommen. Für eine erfolgreiche Invertierung von Datensätzen müssen unterschiedliche Kriterien erfüllt werden, die in den folgenden Abschnitten hergeleitet werden. Vor der mathematischen Diskussion der Invertierbarkeit werden unterschiedliche Aspekte herausgegriffen und grafisch analysiert.

Existenz der Lösung

Notwendige Voraussetzung für die Invertierung einer Abbildung ist die Existenz der Lösung. Um diese Bedingung zu verdeutlichen, vergleicht Bild 12.15 zwei Stichproben.

IMAGE15

Bild 12.15: Existenz der Lösung am Beispiel einer

In dem linken Diagramm liegen die Stichprobenwerte exakt auf der Funktion

$$y_n = \sqrt{x_n} \quad (13.89)$$

Damit existiert eine inverse Abbildung

$$x_n = y_n^2 \quad (13.90)$$

Im Gegensatz dazu weichen die Stichprobenwerte in dem rechten Diagramm von der idealen Funktion um einen Differenz e_n ab.

$$y_n = \sqrt{x_n} + e_n \quad (13.91)$$

Es existiert zunächst keine eindeutige inverse Funktion. Die hier gezeigte Problematik wird in den vorangegangenen Abschnitten dadurch umgangen, dass keine Invertierung durchgeführt wird, sondern eine Optimierung. Es wird die Lösung gesucht, die im Sinne der quadratischen Fehler

$$a = \sum_{n=1}^N r_n^2 = \sum_{n=1}^N (y_n - b_0 - b_1 \cdot x_n)^2 \quad (13.92)$$

die beste Approximation liefert. Damit existiert auch in diesem Fall eine inverse Abbildung.

Eindeutigkeit der Invertierung

Die verwendeten Regressionsmodelle bestehen teilweise aus Polynomen höherer Ordnung. Ein Beispiel ist die Parabel

$$y = x^2 \quad (13.93)$$

Die Invertierung dieser Polynome muss nicht eindeutig sein. Bild 12.16 verdeutlicht die Problematik grafisch.

IMAGE16

Bild 12.16: Diskussion der Invertierbarkeit an einem Polynom höherer Ordnung

Der Funktionswert y kann nicht eindeutig auf x umgerechnet werden. Es ergibt sich eine Lösung für positive und eine Lösung für negative Werte von x . Die Mehrdeutigkeit der Lösung wird mathematisch dargestellt als

$$x = \pm \sqrt{y} \quad (13.94)$$

Das Beispiel verdeutlicht, dass eine eindeutige Invertierung nur dann erreicht wird, wenn Zusatzinformationen vorliegen. Die Parabel wäre eindeutig invertierbar, wenn die Zusatzinformation $x > 0$ vorliegt. In dem Fall wäre die Lösung

$$x = \sqrt{y} \quad (13.95)$$

Es wird sich zeigen, dass auch bei in der Regressionsrechnung schlecht gestellter Probleme zusätzliche Annahmen getroffen werden müssen, um eindeutige Lösung zu erzielen.

Lineare Abhängigkeit und Kolinearität

Durch die Regressionsrechnung werden funktionale Zusammenhänge zwischen Stichprobenwerten beschrieben. Im Fall zweier Eingangs- und einer Zielgrößen wird eine Fläche bestimmt, die das Verhalten von Datenpunkten mathematisch zusammenfassen beschreiben sollen. Bild 12.17 zeigt im linken Diagramm eine Konstellation von Datenpunkten im Raum.

IMAGE17

Bild 12.17: Lineare Abhängigkeit der Datenpunkte im dreidimensionalen Raum

Bei dieser Konstellation liegen alle Datenpunkte, die zur Bestimmung einer Regressionsfläche verwendet werden können, auf einer Linie. Solche Stichproben werden als kolinear bezeichnet. Dadurch sind alle Flächen eine Lösung, die um diese Linie herum rotieren. In Bild 12.17 sind exemplarisch zwei mögliche Regressionsflächen eingezeichnet. Aufgrund der linearen Abhängigkeit der Datenpunkte existiert keine eindeutige Lösung.

Selbst wenn die Stichprobenwerte nicht exakt auf einer Geraden liegen, aber eine Streuung aufweisen, kann die Regressionsfläche nicht eindeutig bestimmt werden. Die Lösung hängt außerdem stark von den vorliegenden Stichprobenwerten und ihrer Streuung ab. Eine für die Regression geeigneter Datensatz besitzt deshalb Datenspalten, die voneinander linear unabhängig sind.

Stabilität der Lösung

Existiert eine eindeutige inverse Abbildung, muss die inverse Abbildung stabil sein. Unter einer stabilen Lösung wird eine Lösung verstanden, die auf kleine Änderungen Δx der Eingangsgröße

$$x = x_0 + \Delta x \quad (13.96)$$

mit kleinen Änderungen Δy der Zielgröße reagiert.

$$y = y_0 + \Delta y \quad (13.97)$$

Bild 12.18 verdeutlicht diesen Gedanken mit einem Gegenbeispiel, bei dem die inverse Abbildung einen Sprung aufweist.

IMAGE18

Bild 12.18: Instabilität der Interpolation

Da die Inverse an der Stelle x_0 einen Sprung in ihrem Verlauf aufweist, führen selbst beliebig kleine Änderungen Δx zu erheblichen Änderungen der Größe y . Aufgrund von Prozessstreuungen wird das Regressionsergebnis streuen.

Hadamarsche Korrektheitsdefinition

Die Vorüberlegungen zur Invertierbarkeit werden in der Hadamarschen Korrektheitsdefinition zusammengefasst. Eine Abbildung der Form

$$\underline{b} = (\mathbf{X}^T \cdot \mathbf{X})^{-1} \cdot \mathbf{X}^T \cdot \underline{y} \quad (13.98)$$

ist nach Hadamar korrekt, wenn die oben beschriebenen Bedingungen an Existenz, Eindeutigkeit und Stabilität erfüllt sind. Die hier verbal beschriebenen Anforderungen werden im folgenden mathematisch begründet, und es werden entsprechende Kriterien für den Nachweis hergeleitet.

13.3.2 Lösung und Lösbarkeit linearer Gleichungssysteme

Liegen $N = M + 1$ Messungen für die Bestimmung von $M + 1$ Regressionskoeffizienten vor, handelt es sich bei der Aufgabe

$$\underline{y} = \mathbf{X} \cdot \underline{b} \quad (13.99)$$

um ein lineares Gleichungssystem mit $M + 1$ Unbekannten und $M + 1$ Gleichungen. Im Normalfall liegen mehr als $M + 1$ Messungen zur Bestimmung der Regressionskoeffizienten vor. Trotzdem wird an dieser Stelle auf die Lösbarkeit von llnraren Gleichungssystemen eingegangen, weil sich die dabei gewonnenen Erkenntnisse auf die Berechnung von Regressionsfunktion übertragen lassen.

Da es sich bei der Matrix um einen quadratische Matrix handelt, kann die Gleichung durch Multiplikation der inversen Matrix \mathbf{X}^{-1} gelöst werden, wenn sie existiert. Der Koeffizientenvektor \underline{b} berechnet sich dann zu

$$\mathbf{X}^{-1} \cdot \underline{y} = \mathbf{X}^{-1} \cdot \mathbf{X} \cdot \underline{b} = \underline{b} \quad (13.100)$$

Auf die dazu erforderlichen Bedingungen wird in diesem Abchnitt eingegangen.

Determinante einer Matrix

Die Determinante D einer $N \times N$ Matrix \mathbf{X}

$$D = \det(\mathbf{X}) \quad (13.101)$$

lässt sich nach dem Laplaceschen Entwicklungssatz durch die Entwicklung nach der n-ten Zeile

$$D = \sum_{m=1}^M x_{nm} \cdot (-1)^{n+m} \cdot D_{nm} \quad (13.102)$$

oder der m-ten Spalte

$$D = \sum_{n=1}^M x_{nm} \cdot (-1)^{n+m} \cdot D_{nm} \quad (13.103)$$

berechnen. Dabei ist D_{nm} die $(N - 1)$ - reihige Unterdeterminante von D. Für Determinanten gilt das Multiplikationstheorem

$$\det(\mathbf{X} \cdot Y) = \det(\mathbf{X}) \cdot \det(Y) \quad (13.104)$$

Die Determinanten der Matrix \mathbf{X} und \mathbf{X}^T sind identisch.

$$\det(\mathbf{X}) = \det(\mathbf{X}^T) \quad (13.105)$$

Die Determinante der Einheitsmatrix \mathbf{E} ist eins.

$$\det(\mathbf{E}) = 1 \quad (13.106)$$

Determinanten sind null, wenn eine Zeile beziehungsweise eine Spalte als Linearkombination der übrigen Zeilen beziehungsweise Spalten dargestellt werden kann. In dem Fall sind die Zeilen beziehungsweise Spalten voneinander linear abhängig.

Weitere Rechneregeln zu Determinanten sind zum Beispiel in [] zusammengestellt.

Beispiel: Unterdeterminanten

<<< wird später ergänzt >>>

Reguläre Matrix und Inverse einer Matrix

Eine $N \times N$ Matrix \mathbf{X} ist regulär, wenn sie eine Determinante $\det(\mathbf{X}) \neq 0$ besitzt. Die inverse Matrix \mathbf{X}^{-1} existiert, wenn die Matrix regulär ist. In dem Fall ist die Determinante $\det(\mathbf{X})$ von null verschieden und die Inverse berechnet sich über

$$\mathbf{X}^{-1} = \frac{1}{\det(\mathbf{X})} \cdot \begin{pmatrix} X_{11} & X_{21} & \cdots & X_{N1} \\ X_{12} & X_{22} & \cdots & X_{N2} \\ \vdots & \vdots & \ddots & \vdots \\ X_{1N} & X_{2N} & \cdots & X_{NN} \end{pmatrix} \quad (13.107)$$

Dabei berechnen sich die X_{nm} mit der Determinanten D

$$D = \det(\mathbf{X}) \quad (13.108)$$

über

$$X_{nm} = (-1)^{n+m} \cdot D_{nm} \quad (13.109)$$

D_{nm} ist wieder die $(N - 1)$ - reihige Unterdeterminante von D. Für die Inverse einer Matrix existieren folgende Rechenregeln.

Tabelle 13.11: Rechenregeln zur Inversen einer Matrix

Regel	Gleichung
Inverse einer Inversen	$\mathbf{X} = (\mathbf{X}^{-1})^{-1}$
Inverse des Produktes zweier Matrizen	$(\mathbf{X} \cdot \mathbf{Y})^{-1} = \mathbf{Y}^{-1} \cdot \mathbf{X}^{-1}$
Inverse der Transponierten	$(\mathbf{X}^T)^{-1} = (\mathbf{X}^{-1})^T$

Beispiel: Inverse einer Matrix

<<< wird später ergänzt >>>

Orthogonale Matrizen

Im Rahmen der Versuchsplanung werden orthogonale Matrizen zur Regressionsrechnung eingesetzt. Eine $N \times N$ Matrix \mathbf{X} ist orthogonal, wenn

$$\mathbf{X}^T \cdot \mathbf{X} = E \quad (13.110)$$

gilt. Für orthogonale Matrizen gilt mit den Rechenregeln für Determinanten

$$\det(\mathbf{X}^T \cdot \mathbf{X}) = \det(\mathbf{X}^T) \cdot \det(\mathbf{X}) = \det(\mathbf{X}) \cdot \det(\mathbf{X}) = \det(E) = 1 \quad (13.111)$$

Daraus folgt, dass die Determinante der Matrix \mathbf{X} den Wert $\det(\mathbf{X}) = \pm 1$ aufweist und damit immer regulär ist. Damit existiert die Inverse \mathbf{X}^{-1} , und es gilt:

$$\mathbf{X} \cdot \mathbf{X}^{-1} = E = \mathbf{X} \cdot \mathbf{X}^T \quad (13.112)$$

Für orthogonale Matrizen sind inverse Matrix \mathbf{X}^{-1} und transponierte Matrix \mathbf{X}^T identisch.

$$\mathbf{X}^{-1} = \mathbf{X}^T \quad (13.113)$$

Rang einer Matrix

Für nichtquadratische $N \times M$ Matrizen können keine Determinanten bestimmt werden. Es können aber Unterdeterminanten berechnet werden. Dazu werden $N - P$ Zeilen und $M - P$ Spalten gestrichen, sodass quadratische $P \times P$ Matrix entstehen. Die Determinanten der P -reihigen Restmatrix werden Unterdeterminanten P -ter Ordnung genannt.

Unter dem Rang einer $N \times M$ Matrix \mathbf{X} wird die höchste Ordnung R aller von Null verschiedenen Unterdeterminanten von \mathbf{X} verstanden.

$$R = Rg(\mathbf{X}) \quad (13.114)$$

Der Rang ist immer kleiner als M und N . Er ändert sich nicht, wenn

- zwei Spalten miteinander getauscht werden
- eine Zeile oder eine Spalte mit einem Faktor $k \neq 0$ multipliziert wird
- zu einer Zeile oder Spalte ein beliebiges Vielfaches einer anderen Zeile oder Spalte addiert wird

Zur Bestimmung des Rangs einer Matrix kann die Matrix mithilfe elementarer Umformungen auf ein Trapezform gebracht werden.

$$\mathbf{T} = \left(\begin{array}{ccccccc} y_{11} & y_{12} & \cdots & y_{1R} & y_{1R+1} & \cdots & y_{1N} \\ 0 & y_{22} & \cdots & y_{2R} & y_{2R+1} & \cdots & y_{2N} \\ \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & y_{RR} & y_{RR+1} & \cdots & y_{RN} \\ 0 & 0 & \cdots & 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 0 & 0 & \cdots & 0 \end{array} \right) \quad (13.115)$$

Der Rang der Matrix entspricht dann der Anzahl nicht verschwindender Zeilen R . Diese Anzahl von Zeilen entspricht der Zahl unabhängiger Zeilen- beziehungsweise Spaltenvektoren.

Beispiel: Unterdeterminanten und Rang einer Matrix

<<< wird später ergänzt >>>

Lösung linearer Gleichungssysteme

Lineare Gleichungssysteme können zum Beispiel mithilfe des Gauß-Algorithmus gelöst werden. Dabei zeigt sich, dass sie nur eine Lösung haben, wenn die Koeffizientenmatrix

$$\mathbf{X} = \begin{pmatrix} 1 & \cdots & x_{1m} & \cdots & x_{1M} \\ \vdots & \ddots & \vdots & & \vdots \\ 1 & \cdots & x_{nm} & \cdots & x_{nM} \\ \vdots & & \vdots & \ddots & \vdots \\ 1 & \cdots & x_{Nm} & \cdots & x_{NM} \end{pmatrix} \quad (13.116)$$

und die erweiterte Koeffizientenmatrix

$$(\mathbf{X} \underline{y}) = \begin{pmatrix} 1 & \cdots & x_{1m} & \cdots & x_{1M} & y_1 \\ \vdots & \ddots & \vdots & & \vdots & \vdots \\ 1 & \cdots & x_{nm} & \cdots & x_{nM} & y_n \\ \vdots & & \vdots & \ddots & \vdots & \vdots \\ 1 & \cdots & x_{Nm} & \cdots & x_{NM} & y_N \end{pmatrix} \quad (13.117)$$

denselben Rang R besitzen. Für $R = N$ existiert genau eine Lösung, für $R < N$ existieren unendlich viele Lösungen. Die Lösungen haben $N - R$ frei wählbare Parameter. Zu einem Gleichungssystem

$$\underline{y} = \mathbf{X} \cdot \underline{b} \quad (13.118)$$

gehört ein homogenes Gleichungssystem

$$\underline{0} = \mathbf{X} \cdot \underline{b}_H \quad (13.119)$$

Das homogene Gleichungssystem hat entweder die Lösung $\underline{b}_H = 0$ oder unendlich viele Lösungen. Falls es unendlich viele Lösungen besitzt, wird der Raum, in dem diese Lösungen liegen, als Nullraum bezeichnet. Er hat $N - R$ Dimensionen.

Beispiel: Lösung linearer Gleichungssysteme

<<< wird später ergänzt >>>

Eigenwerte und Eigenvektoren

Durch die Matrizengleichung

$$(X - \lambda \cdot E) \cdot \underline{b} = 0 \quad (13.120)$$

wird ein Eigenwertproblem beschrieben. Dabei ist λ der Eigenwert der Matrix \mathbf{X} und \underline{b} ein Eigenvektor der Matrix \mathbf{X} . Die Matrix $\mathbf{A} - \lambda \cdot \mathbf{E}$ wird als charakteristische Matrix bezeichnet.

Die Eigenwerte errechnen sich durch Lösen der charakteristischen Gleichung

$$\det(X - \lambda \cdot E) = 0 \quad (13.121)$$

Sie ist bei einer $N \times N$ Matrix ein Polynom N -ter Ordnung, das N Lösungen λ_n aufweist. Zu den Eigenwerten λ_n gehören Eigenvektoren \underline{b}_n , die die Lösungsvektoren des Eigenwertproblems darstellen.

$$(X - \lambda_n \cdot E) \cdot \underline{b}_n = 0 \quad (13.122)$$

Sind die Eigenwerte λ_n voneinander verschieden, sind die Eigenvektoren linear unabhängig. Bei mehrfachen Eigenwerten sind die Eigenvektoren voneinander linear abhängig.

Die Determinante von \mathbf{X} ist gleich dem Produkt der Eigenwerte.

$$\det(\mathbf{X}) = \lambda_1 \cdot \lambda_2 \cdot \dots \cdot \lambda_N \quad (13.123)$$

Damit ist die Forderung nach Invertierbarkeit $\det(\mathbf{X}) \neq 0$ identisch mit der Forderung, dass alle $\lambda_n \neq 0$ sein müssen.

Beispiel: Eigenwerte und Eigenvektoren

<<< wird später ergänzt >>>

Diagonalform einer Matrix

Für jeden Eigenwert λ_n und jeden Eigenvektor \underline{b}_n gilt die Gleichung

$$(\mathbf{X} - \lambda_n \cdot \mathbf{E}) \cdot \underline{b}_n = 0 \quad (13.124)$$

beziehungsweise

$$\mathbf{X} \cdot \underline{b}_n = \lambda_n \cdot \mathbf{E} \cdot \underline{b}_n = \lambda_n \cdot \underline{b}_n = \underline{b}_n \cdot \lambda_n \quad (13.125)$$

Werden alle Eigenvektoren in einer Matrix zusammengefasst, ergibt sich die Gleichung

$$\mathbf{X} \cdot \begin{pmatrix} \underline{b}_1 & \dots & \underline{b}_N \end{pmatrix} = \begin{pmatrix} \underline{b}_1 & \dots & \underline{b}_N \end{pmatrix} \cdot \begin{pmatrix} \lambda_1 & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & \lambda_N \end{pmatrix} \quad (13.126)$$

Eine Matrix \mathbf{X} kann mithilfe der Eigenwerte λ_n und der Eigenvektoren \underline{b}_n dargestellt werden als

$$\mathbf{X} = \begin{pmatrix} \underline{b}_1 & \dots & \underline{b}_N \end{pmatrix} \cdot \begin{pmatrix} \lambda_1 & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & \lambda_N \end{pmatrix} \cdot \begin{pmatrix} \underline{b}_1 & \dots & \underline{b}_N \end{pmatrix}^{-1} \quad (13.127)$$

Mit dieser Darstellung lassen sich zum einen Potenzen von Matrizen wegen der Beziehung

$$\begin{pmatrix} \underline{b}_1 & \dots & \underline{b}_N \end{pmatrix} \cdot \begin{pmatrix} \underline{b}_1 & \dots & \underline{b}_N \end{pmatrix}^{-1} = \mathbf{E} \quad (13.128)$$

vergleichsweise einfach berechnen. Zum Beispiel ergibt sich für \mathbf{X}^2

$$\begin{aligned} \mathbf{X}^2 &= \mathbf{X} \cdot \mathbf{X} = \begin{pmatrix} \underline{b}_1 & \dots & \underline{b}_N \end{pmatrix} \cdot \begin{pmatrix} \lambda_1 & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & \lambda_N \end{pmatrix} \cdot \mathbf{E} \cdot \begin{pmatrix} \lambda_1 & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & \lambda_N \end{pmatrix} \cdot \begin{pmatrix} \underline{b}_1 & \dots & \underline{b}_N \end{pmatrix}^{-1} \\ &= \begin{pmatrix} \underline{b}_1 & \dots & \underline{b}_N \end{pmatrix} \cdot \begin{pmatrix} \lambda_1 & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & \lambda_N \end{pmatrix}^2 \cdot \begin{pmatrix} \underline{b}_1 & \dots & \underline{b}_N \end{pmatrix}^{-1} \\ &= \begin{pmatrix} \underline{b}_1 & \dots & \underline{b}_N \end{pmatrix} \cdot \begin{pmatrix} \lambda_1^2 & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & \lambda_N^2 \end{pmatrix}^2 \cdot \begin{pmatrix} \underline{b}_1 & \dots & \underline{b}_N \end{pmatrix}^{-1} \end{aligned} \quad (13.129)$$

Zum anderen lässt sich die Inverse der Matrix \mathbf{X} berechnen über

$$\begin{aligned}
 \mathbf{X}^{-1} &= \left(\begin{pmatrix} \underline{b}_1 & \cdots & \underline{b}_N \end{pmatrix} \begin{pmatrix} \lambda_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \lambda_N \end{pmatrix} \begin{pmatrix} \underline{b}_1 & \cdots & \underline{b}_N \end{pmatrix}' \right)^{-1} \\
 &= \left(\begin{pmatrix} \underline{b}_1 & \cdots & \underline{b}_N \end{pmatrix} \begin{pmatrix} \lambda_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \lambda_N \end{pmatrix} \begin{pmatrix} \underline{b}_1 & \cdots & \underline{b}_N \end{pmatrix}' \right)^{-1} \\
 &= \left(\begin{pmatrix} \underline{b}_1 & \cdots & \underline{b}_N \end{pmatrix} \begin{pmatrix} \frac{1}{\lambda_1} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \frac{1}{\lambda_N} \end{pmatrix} \begin{pmatrix} \underline{b}_1 & \cdots & \underline{b}_N \end{pmatrix}' \right)^{-1}
 \end{aligned} \tag{13.130}$$

Eine Invertierung der Matrix \mathbf{X} mit den Eigenwerten λ_n führt zur Inversen \mathbf{X}^{-1} mit den Eigenwerten $1/\lambda_n$. Eine Invertierung ist nur möglich, wenn alle λ_n von null verschieden sind.

Beispiel: Diagonalform einer Matrix

<<< wird später ergänzt >>>

Eigensysteme und Optimierung

<<< wird später ergänzt >>>

Beispiel: Eigensystem und Optimierung

<<< wird später ergänzt >>>

13.3.3 Singulärwertzerlegung von Matrizen (13.06.2015)

Die in Abschnitt 12.3.2 beschriebenen Eigenschaften und Verfahren gelten für N-dimensionale Gleichungssysteme mit N x N Matrizen. Im Fall der Regression liegen typischerweise mehr als N = M + 1 Messungen vor, so dass die in dem Abschnitt beschriebenen Verfahren in der Regressionsrechnung nicht direkt angewendet werden können. Wesentlicher Grund ist die Tatsache, dass die Matrix \mathbf{X} keine quadratische N x N Matrix sondern eine N x M + 1 Matrix ist.

Die Berechnung der Regressionskoeffizienten über die Gleichung

$$\underline{b} = (\mathbf{X}^T \cdot \mathbf{X})^{-1} \cdot \mathbf{X}^T \cdot \underline{y} \quad (13.131)$$

beruht auf der Minimierung der Summe quadratischer Fehler. Dabei wird die Pseudoinverse $(\mathbf{X}^T \cdot \mathbf{X})^{-1}$ gebildet. Die statistische Bewertung der Regression hat gezeigt, dass die Pseudoinverse in die Berechnung der Kovarianzmatrix von den Regressionskoeffizienten \underline{b} und damit in das Konfidenzintervall sowie das Prognoseintervall der Regression eingeht. Deshalb wird statt der Matrix \mathbf{X} die Matrix $\mathbf{X}^T \cdot \mathbf{X}$ diskutiert.

Definition der Singulärwertzerlegung

Eine N x M Matrix \mathbf{X} kann als Produkt

$$\mathbf{X} = \mathbf{U} \cdot \mathbf{S} \cdot \mathbf{V}^T \quad (13.132)$$

dargestellt werden. Die Zerlegung der Matrix in die angegebenen Produkte wird als Singulärwertzerlegung bezeichnet (Singular Value Decomposition). Die Matrix \mathbf{S} ist eine Diagonalmatrix. Die Elemente der Diagonalmatrix werden als singuläre Werte oder Singulärwerte bezeichnet. Sie sind die positiven Quadratwurzeln der Eigenwerte von $\mathbf{X} \cdot \mathbf{X}^T$ und $\mathbf{X}^T \cdot \mathbf{X}$. Sie sind typischerweise abfallend angeordnet.

Die Matrix \mathbf{U} besteht aus den Eigenvektoren von $\mathbf{X} \cdot \mathbf{X}^T$ und die Matrix \mathbf{V} besteht aus den Eigenvektoren der Matrix $\mathbf{X}^T \cdot \mathbf{X}$. Die Vektoren der Matrix \mathbf{U} werden als linke und die der rechten Matrix \mathbf{V} als rechte Singulärvektoren bezeichnet. Um einen Eindruck von der Dimension der Matrizen \mathbf{U} , \mathbf{S} und \mathbf{V} zu bekommen, sind die Matrizen in Gleichung exemplarisch für eine 3 x 2 Matrix \mathbf{X} dargestellt.

$$\mathbf{X} = \begin{pmatrix} x_{11} & x_{12} \\ x_{21} & x_{22} \\ x_{31} & x_{32} \end{pmatrix} = \begin{pmatrix} u_{11} & u_{12} & u_{13} \\ u_{21} & u_{22} & u_{23} \\ u_{31} & u_{32} & u_{33} \end{pmatrix} \cdot \begin{pmatrix} s_1 & 0 \\ 0 & s_2 \\ 0 & 0 \end{pmatrix} \cdot \begin{pmatrix} v_{11} & v_{21} \\ v_{12} & v_{22} \end{pmatrix} = \mathbf{U} \cdot \mathbf{S} \cdot \mathbf{V}^T \quad (13.133)$$

Auf die Berechnung der Singulärwertzerlegung wird an dieser Stelle nicht näher eingegangen. Sie ist zum Beispiel in [] beschrieben und kann mit dem MATLAB-Befehl svd durchgeführt werden.

Die Matrizen \mathbf{U} und \mathbf{V} sind orthogonal. Für orthogonale Matrizen gilt die Beziehung

$$\mathbf{V}^{-1} = \mathbf{V}^T \quad (13.134)$$

beziehungsweise

$$\mathbf{U}^{-1} = \mathbf{U}^T \quad (13.135)$$

Damit kann Gleichung (13.132) umgeformt werden zu

$$\mathbf{X} \cdot \mathbf{V} = \mathbf{U} \cdot \mathbf{S} \cdot \mathbf{V}^T \cdot \mathbf{V} = \mathbf{U} \cdot \mathbf{S} \cdot \mathbf{V}^{-1} \cdot \mathbf{V} = \mathbf{U} \cdot \mathbf{S} \quad (13.136)$$

Die ersten R Elemente s_n der Diagonalmatrix \mathbf{S} sind von null verschieden. Sie geben den Rang der Matrix \mathbf{X} an. Für $n \leq R$ gilt:

$$\mathbf{X} \cdot \underline{v}_n = s_n \cdot \underline{u}_n \quad (13.137)$$

Das Produkt der Matrix X mit einer Spalte von V ist ein Vielfaches der entsprechenden Spalte von U.
Für $R + 1 \leq n \leq N$ gilt

$$\mathbf{X} \cdot \underline{v}_n = \underline{0} \quad (13.138)$$

Die Vektoren \underline{v}_n liegen im Nullraum von \mathbf{X} .

Singulärwertzerlegung in der Bildverarbeitung

09.06.

Singulärwertzerlegung in der Regressionsrechnung

11.06.

13.4 Korrektheit und Kondition einer Matrix

$\det(X) \neq 0$ bedeutet Korrektheit nach Hadamard

13.5 Regularisierung von Matrizen (21.06.2015)

<<< Wird später ergänzt ... >>>

14 Transformation von Zufallsvariablen

15 Anhang 2: Grundlagen der linearen Algebra

15.1 Matrizenalgebra

In vielen multivariaten Fragestellungen in der Statistik wird Matrixalgebra benötigt. Insbesondere werden bei der multiple Regression die Variablen in einer sogenannten Design-Matrix \underline{X} zusammengefasst. Sie hat M Zeilen, die die Anzahl der Beobachtungen repräsentiert, und N Spalten, die die Anzahl der an der Regression beteiligten Variablen repräsentiert. Ein Eintrag dieser Matrix entspricht der Messung eines Merkmals bei einer Beobachtung. Eine Zeile der Matrix enthält die Messergebnisse aller Merkmale für eine Beobachtung. Die Werte der Zielgröße werden in einem M -dimensionalen Vektor \underline{Y} zusammengefasst. Damit gilt die Gleichung

$$\underline{X} \cdot B = \underline{Y} \quad (15.1)$$

Es wird sich zeigen, dass zur Schätzung des Regressionskoeffizienten die Gleichung

$$B = (\underline{X}^T \cdot \underline{X})^{-1} \cdot \underline{X}^T \cdot \underline{Y} \quad (15.2)$$

gelöst werden muss. Zur Berechnung werden also Matrixmultiplikation, die Inverse einer Matrix und die Transponierte einer Matrix benötigt. Auch in vielen anderen Anwendungen in der Statistik ist die Matrixalgebra von zentraler Bedeutung. Die folgenden Abschnitte stellen die wesentlichen Rechenregeln der Matrixalgebra zusammen.

15.1.1 Vektoren und Matrizen

Ein Vektor \underline{X} ist eine Anordnung von Elementen in einer Spalte mit M Zeilen.

$$\underline{X} = \begin{pmatrix} X_1 \\ X_2 \\ \vdots \\ X_M \end{pmatrix} \quad (15.3)$$

Ein Zeilenvektor \underline{X}^T ist eine Anordnung von Elementen in einer Zeile mit N Spalten.

$$\underline{X}^T = (X_1 \ X_1 \ \dots \ X_N) \quad (15.4)$$

Eine Matrix \underline{X} ist eine rechteckförmige Anordnung von Elementen mit M Zeilen und N Spalten. Stimmt die Anzahl von Zeilen M und die Anzahl von Spalten N überein, wird sie als quadratische Matrix bezeichnet.

$$\underline{X} = \begin{pmatrix} X_{11} & X_{12} & \dots & X_{1N} \\ X_{21} & X_{22} & \dots & X_{2N} \\ \dots & \dots & \dots & \dots \\ X_{M1} & X_{M2} & \dots & X_{MN} \end{pmatrix} \quad (15.5)$$

Matrizen können aus verschiedenen Spaltenvektoren $X_1 \dots X_N$ zusammengesetzt werden. Alternativ können sie als Spaltenvektoren $X_1^T \dots X_M^T$ zusammengesetzt werden.

$$\underline{X} = \begin{pmatrix} X_{11} & X_{12} & \dots & X_{1N} \\ X_{21} & X_{22} & \dots & X_{2N} \\ \dots & \dots & \dots & \dots \\ X_{M1} & X_{M2} & \dots & X_{MN} \end{pmatrix} = (X_1 \dots X_N) = \begin{pmatrix} X_1^T \\ \vdots \\ X_M^T \end{pmatrix} \quad (15.6)$$

Die Transponierte \underline{X}^T einer Matrix \underline{X} entsteht durch Vertauschen von Zeilen und Spalten der Matrix \underline{X} .

$$\underline{X} = \begin{pmatrix} X_{11} & X_{12} & \dots & X_{1N} \\ X_{21} & X_{22} & \dots & X_{2N} \\ \dots & \dots & \dots & \dots \\ X_{M1} & X_{M2} & \dots & X_{MN} \end{pmatrix}^T = \begin{pmatrix} X_{11} & X_{12} & \dots & X_{1M} \\ X_{21} & X_{22} & \dots & X_{2M} \\ \dots & \dots & \dots & \dots \\ X_{N1} & X_{N2} & \dots & X_{NM} \end{pmatrix} \quad (15.7)$$

Für das Transponieren von Matrizen gelten die in Tabelle 15.1 zusammengefassten Rechenregeln.

Tabelle 15.1: Rechenregeln für das Transponieren von Matrizen

Regel	Gleichung
Doppeltes Transponieren	$(\underline{X}^T)^T = \underline{X}$
Transponieren einer Summe von Matrizen	$(\underline{X} + \underline{Y})^T = \underline{X}^T + \underline{Y}^T$
Transponieren des Produktes von Matrizen	$(\underline{X} \cdot \underline{Y})^T = \underline{X}^T \cdot \underline{Y}^T$

Für das Rechnen mit Matrizen sind einige Matrizen von besonderer Bedeutung. Sie sind Tabelle 15.2 zusammengefasst.

Tabelle 15.2: Matrizen mit besonderer Bedeutung

Befehl	Beschreibung
Nullmatrix	$\underline{0} = \begin{pmatrix} 0 & 0 & \dots & 0 \\ 0 & 0 & \dots & 0 \\ \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & 0 \end{pmatrix}$
Matrix mit Einsen	$\underline{I} = \begin{pmatrix} 1 & 1 & \dots & 1 \\ 1 & 1 & \dots & 1 \\ \dots & \dots & \dots & \dots \\ 1 & 1 & \dots & 1 \end{pmatrix}$
n-ter Einheitsvektor	$E_n = (0, \dots, 1, \dots, 0)$
Diagonalmatrix	$\underline{X} = \text{diag}(x_{11} \dots x_{NN}) = \begin{pmatrix} x_{11} & 0 & \dots & 0 \\ 0 & x_{22} & \dots & 0 \\ \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & x_{NN} \end{pmatrix}$
Einheitsmatrix	$\underline{I} = \begin{pmatrix} 1 & 0 & \dots & 0 \\ 0 & 1 & \dots & 0 \\ \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & 1 \end{pmatrix}$

Eine Matrix kann aus Teilmatrizen unterteilt oder partitioniert werden.

$$\underline{X} = \begin{pmatrix} X_{11} & X_{12} & \dots & X_{1N} \\ X_{21} & X_{22} & \dots & X_{2N} \\ \dots & \dots & \dots & \dots \\ X_{M1} & X_{M2} & \dots & X_{MN} \end{pmatrix} = \begin{pmatrix} \underline{X}_{11} & \underline{X}_{12} \\ \underline{X}_{21} & \underline{X}_{22} \end{pmatrix} \quad (15.8)$$

Bei Transponieren der Matrizen gilt für partitionierte Matrizen:

$$\underline{X}^T = \begin{pmatrix} \underline{X}_{11} & \underline{X}_{12} \\ \underline{X}_{21} & \underline{X}_{22} \end{pmatrix}^T = \begin{pmatrix} \underline{X}_{11}^T & \underline{X}_{12}^T \\ \underline{X}_{21}^T & \underline{X}_{22}^T \end{pmatrix} \quad (15.9)$$

15.1.2 Matrizenoperationen

Addition von Matrizen

Die Summe $\underline{X} + \underline{Y}$ zweier $M \times N$ Matrizen berechnet sich aus dem komponentenweise Addition der einzelnen Elemente.

$$\begin{aligned}\underline{X} + \underline{Y} &= \begin{pmatrix} X_{11} & X_{12} & \dots & X_{1N} \\ X_{21} & X_{22} & \dots & X_{2N} \\ \dots & \dots & \dots & \dots \\ X_{M1} & X_{M2} & \dots & X_{MN} \end{pmatrix} + \begin{pmatrix} Y_{11} & Y_{12} & \dots & Y_{1N} \\ Y_{21} & Y_{22} & \dots & Y_{2N} \\ \dots & \dots & \dots & \dots \\ Y_{M1} & Y_{M2} & \dots & Y_{MN} \end{pmatrix} \\ &= \begin{pmatrix} X_{11} + Y_{11} & X_{12} + Y_{12} & \dots & X_{1N} + Y_{1N} \\ X_{21} + Y_{21} & X_{22} + Y_{22} & \dots & X_{2N} + Y_{2N} \\ \dots & \dots & \dots & \dots \\ X_{M1} + Y_{M1} & X_{M2} + Y_{M2} & \dots & X_{MN} + Y_{MN} \end{pmatrix}\end{aligned}\tag{15.10}$$

Die Summe kann nur ausgeführt werden, wenn die Anzahl von Zeilen und Spalten der beiden Matrizen identisch ist. Für die Addition von Matrizen gelten die in Tabelle 15.3 aufgeführten Rechenregeln. Sie ergeben sich aus der komponentenweise Addition der Matrixelemente.

Tabelle 15.3: Rechenregeln für das Transponieren von Matrizen

Operation	Mathematische Beschreibung
Assoziativgesetz	$(\underline{X} + \underline{Y}) + \underline{Z} = \underline{X} + (\underline{Y} + \underline{Z})$
Kommutativgesetz	$\underline{X} + \underline{Y} = \underline{Y} + \underline{X}$
Neutrales Element	$\underline{X} + 0 = \underline{X}$
Inverses Element	$\underline{X} + (-\underline{X}) = \underline{X} - \underline{X} = 0$

Produkt von einer skalaren Größe und einer Matrix

Das Produkt skalaren Größe a und einer Matrix \underline{X} ist definiert als

$$\lambda \cdot \underline{X} = \begin{pmatrix} \lambda \cdot X_{11} & \lambda \cdot X_{12} & \dots & \lambda \cdot X_{1N} \\ \lambda \cdot X_{21} & \lambda \cdot X_{22} & \dots & \lambda \cdot X_{2N} \\ \dots & \dots & \dots & \dots \\ \lambda \cdot X_{M1} & \lambda \cdot X_{M2} & \dots & \lambda \cdot X_{MN} \end{pmatrix}\tag{15.11}$$

Für die skalare Multiplikation gelten die in Tabelle 15.4 aufgeführten Rechenregeln. Sie ergeben sich aus der komponentenweise skalaren Multiplikation der Matrixelemente.

Tabelle 15.4: Rechenregeln für die skalare Multiplikation

Operation	Mathematische Beschreibung
Distributivgesetz	$(\lambda + \mu) \cdot \underline{X} = \lambda \cdot \underline{X} + \mu \cdot \underline{X}$ $\lambda \cdot (\underline{X} + \underline{Y}) = \lambda \cdot \underline{X} + \lambda \cdot \underline{Y}$
Assoziativgesetz	$(\lambda \cdot \mu) \cdot \underline{X} = \lambda \cdot (\mu \cdot \underline{X})$
Transposition	$(\lambda \cdot \underline{X})^T = \lambda \cdot \underline{X}^T$

Produkt von Matrizen

Das Produkt der $M \times N$ Matrix \underline{X} und der $N \times P$ Matrix \underline{Y} ist die $M \times P$ Matrix \underline{Z}

$$uX \cdot \underline{Y} = \underline{Z} \quad (15.12)$$

mit den Elementen

$$z_{mp} = \sum_{n=1}^N x_{mn} \cdot y_{np} \quad (15.13)$$

Ausführlich ergibt sich die Matrix \underline{Z} zu

$$\begin{aligned} \underline{X} \cdot \underline{Y} &= \begin{pmatrix} X_{11} & X_{12} & \dots & X_{1N} \\ X_{21} & X_{22} & \dots & X_{2N} \\ \dots & \dots & \dots & \dots \\ X_{M1} & X_{M2} & \dots & X_{MN} \end{pmatrix} \cdot \begin{pmatrix} Y_{11} & Y_{12} & \dots & Y_{1P} \\ Y_{21} & Y_{22} & \dots & Y_{2P} \\ \dots & \dots & \dots & \dots \\ Y_{N1} & Y_{N2} & \dots & Y_{NP} \end{pmatrix} \\ &= \begin{pmatrix} \sum_{n=1}^N x_{1n} \cdot y_{n1} & \sum_{n=1}^N x_{1n} \cdot y_{n2} & \dots & \sum_{n=1}^N x_{1n} \cdot y_{nP} \\ \sum_{n=1}^N x_{2n} \cdot y_{n1} & \sum_{n=1}^N x_{2n} \cdot y_{n2} & \dots & \sum_{n=1}^N x_{2n} \cdot y_{nP} \\ \dots & \dots & \dots & \dots \\ \sum_{n=1}^N x_{Mn} \cdot y_{n1} & \sum_{n=1}^N x_{Mn} \cdot y_{n2} & \dots & \sum_{n=1}^N x_{Mn} \cdot y_{nP} \end{pmatrix} = \begin{pmatrix} Z_{11} & Z_{12} & \dots & Z_{1P} \\ Z_{21} & Z_{22} & \dots & Z_{2P} \\ \dots & \dots & \dots & \dots \\ Z_{M1} & Z_{M2} & \dots & Z_{MP} \end{pmatrix} = \underline{Z} \end{aligned} \quad (15.14)$$

Die Produkt der Matrizen kann nur ausgeführt werden, wenn die Anzahl von Spalten der ersten Matrix mit der Anzahl von Zeilen der zweiten Matrix identisch ist. Für die Addition von Matrizen gelten die in Tabelle 15.3 aufgeführten Rechenregeln. Sie ergeben sich aus der komponentenweise Addition der Matrixelemente.

Tabelle 15.5: Rechenregeln für die Addition von Matrizen

Operation	Mathematische Beschreibung
Distributivgesetz	$(\underline{X} + \underline{Y}) \cdot \underline{Z} = \underline{X} \cdot \underline{Z} + \underline{Y} \cdot \underline{Z}$
Assoziativgesetz	$(\underline{X} \cdot \underline{Y}) \cdot \underline{Z} = \underline{X} \cdot (\underline{Y} \cdot \underline{Z})$
Neutrales Element	$\underline{X} \cdot \underline{I} = \underline{X}$
Transposition	$(\underline{X} \cdot \underline{Y})^T = \underline{X}^T \cdot \underline{Y}^T$

Das Produkt von Matrizen ist nicht kommutativ, auch nicht bei quadratischen Matrizen gleicher Größe.

15.1.3 Kenngrößen von Matrizen

Für das Lösen von linearen Gleichungssystemen sind einige Kenngrößen von Matrizen von Bedeutung.

Lineare Abhängigkeit von Spalten und Zeilenvektoren

Die Spaltenvektoren einer Matrix X_1, \dots, X_N sind linear unabhängig, wenn für jede Linearkombination der Vektoren die Gleichung

$$b_1 \cdot X_1 + b_2 \cdot X_2 + b_3 \cdot X_3 + \dots + b_N \cdot X_N = 0 \quad (15.15)$$

nur für

$$b_1 = b_2 = b_3 = \dots = b_N = 0 \quad (15.16)$$

erfüllt ist.

Rang einer Matrix

Die maximale Anzahl linear unabhängiger Spaltenvektoren einer $M \times N$ Matrix heißt Spaltenrang der Matrix $rgs(\underline{X})$. Die maximale Anzahl linear unabhängiger Zeilenvektoren einer $M \times N$ Matrix heißt Zeilenrang der Matrix $rgz(\underline{X})$. Spaltenrang und Zeilenrang einer Matrix sind gleich groß. Sie werden als Rang $rg(\underline{X})$ der Matrix \underline{X} bezeichnet.

$$rg(\underline{X}) = rg(\underline{X}) = rg(\underline{X}) \leq \min(N, M) \quad (15.17)$$

Für den Rang einer Matrix gelten die in Tabelle 15.6 zusammengefassten Rechenregeln.

Tabelle 15.6: Rechenregeln für den Rang von Matrizen

Operation	Mathematische Beschreibung
Negative Matrix	$rg(\underline{X}) = rg(-\underline{X})$
Transponierte Matrix	$rg(\underline{X}) = rg(\underline{X})^T$
Rang der Summe zweier Matrizen	$rg(\underline{X}) - rg(\underline{Y}) \leq rg(\underline{X} + \underline{Y}) \leq rg(\underline{X}) + rg(\underline{Y})$
Rang des Produktes zweier Matrizen	$rg(\underline{X} \cdot \underline{Y}) \leq \min(rg(\underline{X}), rg(\underline{Y}))$
Rang der Einheitsmatrix $N \times N$	$rg(\underline{I}) = N$

Inverse einer quadratischen Matrix

Die Matrix \underline{X} sei eine $N \times N$ Matrix. Die Matrix \underline{X}^{-1} ist die Inverse zur Matrix \underline{X} , wenn die Beziehung

$$\underline{X} \cdot \underline{X}^{-1} = \underline{X}^{-1} \cdot \underline{X} = \mathbf{I} \quad (15.18)$$

gilt. Die Matrix \underline{X} ist genau dann invertierbar, wenn für den Rang der Matrix $rg(\underline{X})$ der Anzahl von Spaltenvektoren N entspricht.

$$rg(\underline{X}) = N \quad (15.19)$$

Falls die Inverse existiert, ist sie eindeutig. Weist die Matrix \underline{X} den Rang $rg(\underline{X}) = N$ auf, besitzt sie einen vollen Rang und wird als regulär bezeichnet. Ist der Rang der Matrix \underline{X} kleiner als die Anzahl der Spaltenvektoren N , wird sie als singuläre Matrix bezeichnet.

Determinante einer Matrix

Die Determinante $det(\underline{X})$ einer quadratischen Matrix \underline{X} ist für die Lösung von linearen Gleichungssystem und die Berechnung von der Inversen einer Matrix von Bedeutung. Die Determinante einer 2×2 Matrix berechnet sich zu

$$det(\underline{X}) = det \begin{pmatrix} x_{11} & x_{12} \\ x_{21} & x_{22} \end{pmatrix} = x_{11} \cdot x_{22} - x_{12} \cdot x_{21} \quad (15.20)$$

Für eine 3×3 Matrix ergibt sich die Determinante aus

$$\begin{aligned} det(\underline{X}) &= det \begin{pmatrix} x_{11} & x_{12} & x_{13} \\ x_{21} & x_{22} & x_{23} \\ x_{31} & x_{32} & x_{33} \end{pmatrix} \\ &= x_{11} \cdot x_{22} \cdot x_{33} + x_{12} \cdot x_{23} \cdot x_{31} + x_{13} \cdot x_{21} \cdot x_{32} - x_{13} \cdot x_{22} \cdot x_{31} - x_{12} \cdot x_{21} \cdot x_{33} - x_{11} \cdot x_{23} \cdot x_{32} \end{aligned} \quad (15.21)$$

Die Determinante einer quadratischen Matrix mit $N > 3$ wird über die Entwicklung nach der n -ten Spalte

$$det(\underline{X}) = \sum_{n=1}^N (-1)^{n+m} \cdot x_{mn} \cdot det(\underline{X}_{mn}) \quad (15.22)$$

oder m -ten Zeile

$$det(\underline{X}) = \sum_{m=1}^N (-1)^{n+m} \cdot x_{mn} \cdot det(\underline{X}_{mn}) \quad (15.23)$$

berechnet. Dabei ist \underline{X}_{mn} die Teilmatrix, die durch Streichen der m -ten Zeile und der n -ten Spalte entsteht. Die Determinante einer Matrix ist eindeutig bestimmt. Die Determinante einer quadratischen $N \times N$ Matrix kann unter den in Tabelle 15.7 zusammengefassten Bedingungen vereinfacht werden.

Tabelle 15.7: Vereinfachte Berechnung von Determinanten einer Matrix

Bedingung	Mathematische Beschreibung
Spalte oder Reihe der Matrix \underline{X} besteht nur aus Nullen	$\det(\underline{X}) = 0$
Matrix \underline{X} besteht aus zwei identischen Zeilen oder zwei identischen Spalten	$\det(\underline{X}) = 0$
Matrix \underline{X} besitzt Dreiecksform	$\det(\underline{X}) = x_{11} \cdot x_{22} \cdots \cdot x_{NN}$
Matrix \underline{X} ist die Einheitsmatrix $\underline{X} = \underline{I}$	$\det(\underline{X}) = \det(\underline{I}) = 1$

Außerdem gelten für die Berechnung von Determinanten einer Matrix die in Tabelle 15.8 zusammengestellten Rechenregeln.

Tabelle 15.8: Rechenregeln für die Determinante einer Matrix

Bedingung	Mathematische Beschreibung
Transponierte Matrix	$\det(\underline{X}^T) = \det(\underline{X})$
Multiplikation mit einem Skalar λ	$\det(\lambda \cdot \underline{X}^T) = \lambda^N \cdot \det(\underline{X})$
Zusammenhang zwischen Rang und Determinante	$\det(\underline{X}) \neq 0$ äquivalent zu $\text{rg}(\underline{X}) = N$
Produkt zweier Matrizen	$\det(\underline{X} \cdot \underline{Y}) = \det(\underline{X}) \cdot \det(\underline{Y})$
Inverse der Matrix	$\det(\underline{X}^{-1}) = \frac{1}{\det(\underline{X})}$

Der Zusammenhang zwischen Rang und Determinante kann zum Nachweis der Invertierbarkeit einer Matrix \underline{X} verwendet werden. Ist die Determinante der Matrix von null verschieden ($\det(\underline{X}) \neq 0$), ist der Rang der Matrix so groß wie die Anzahl der Spaltenvektoren N ($\text{rg}(\underline{X}) = N$) und damit invertierbar. Ist die Determinante null ($\det(\underline{X}) = 0$), ist die Matrix nicht invertierbar.

Die Determinante einer Matrix ist das Produkt ihrer Eigenwerte.

Spur einer Matrix

Die Matrix \underline{X} sei eine $N \times N$ Matrix. Die Summe aller Diagonalelemente der Matrix wird als Spur der Matrix bezeichnet.

$$\text{sp}(\underline{X}) = \sum_{n=1}^N x_{nn} \quad (15.24)$$

Für die Berechnung der Spur einer Matrix gelten die in Tabelle 15.9 zusammengestellten Rechenregeln.

Tabelle 15.9: Rechenregeln für die Spur einer Matrix

Bedingung	Mathematische Beschreibung
Summer zweier Matrizen	$sp(\underline{X} + \underline{Y}) = sp(\underline{X}) + sp(\underline{Y})$
Tronsponierte Matrix	$sp(\underline{X}^T) = sp(\underline{X})$
Multiplikation mit einem Skalar	$sp(\lambda \cdot \underline{X}) = \lambda \cdot sp(\underline{X})$
Produkt zweier Matrizen	$sp(\underline{X} \cdot \underline{Y}) = sp(\underline{Y} \cdot \underline{X})$

Die Spur einer Matrix ist die Summe ihrer Eigenwerte.

15.1.4 Lösung linearer Gleichungssysteme

Darstellung des Gleichungssystems in Matrix-Schreibweise

Dreieckform des Gleichungssystems

Bewertung der Lösbarkeit von linearen Gleichungssystemen

Sonderfall quadratischen Koeffizientenmatrizen

Berechnung der Inverse einer quadratischen Matrix