

Mathematische Statistik für Informatiker und Ingenieure

Skript zur gleichnamigen Vorlesung aus dem Wintersemester 2006/07

Friedrich Graef*

Stand 12. Februar 2007

*Institut für Angewandte Mathematik, FAU Erlangen-Nürnberg

Inhaltsverzeichnis

1. Statistische Datenanalyse	5
1.1. Beispiele	5
1.2. Statistische Experimente	7
1.3. Inhalt der Vorlesung	8
1.4. Statistik-Software	9
1.5. Literatur	9
2. Der statistische Raum	11
2.1. Beispiele	11
2.1.1. Die Normalverteilungsannahme	12
2.1.2. Bernoulli-Versuchsreihen	12
2.1.3. Vergleich zweier Mittelwerte	13
2.2. Allgemeine Voraussetzungen	13
2.2.1. Parametrische Statistik	14
2.2.2. Absolutstetige und diskrete Verteilungsannahme	14
2.2.3. Unabhängige Wiederholungen	14
2.3. Statistiken	15
2.3.1. Die Verteilung einer Statistik	15
2.3.2. Suffiziente Statistiken	16
2.3.3. Die Koordinatenvariablen	17
2.4. Parameterfunktionen	18
3. Parameterschätzung	19
3.1. Die Maximum-Likelihood-Methode	19
3.1.1. Die Normalverteilung	20
3.1.2. Umrechnungsformeln	21
3.1.3. Suffiziente Statistiken	22
3.1.4. Binomialverteilung	22
3.2. Gütekriterien	23
3.2.1. Die Risikozerlegung	23
3.2.2. Erwartungstreue	24
3.2.3. Die Wirksamkeit	26
4. Konfidenzbereiche	27
4.1. Ein Konfidenzintervall für den Mittelwert	27
4.2. Die Studentsche t-Verteilung	29
4.3. Testverteilungen	31
4.3.1. Die Chi-Quadrat-Verteilung	31
4.3.2. Die t-Verteilung	31
4.4. Konfidenzintervalle für die Parameter der Normalverteilung	32
4.4.1. Konfidenzintervall für den Mittelwert	32
4.4.2. Konfidenzintervall für die Varianz	32

4.5.	Konfidenzbereiche	33
4.5.1.	Die statistische Methode	33
4.5.2.	Die Methode der Pivotvariablen	35
4.5.3.	Pivotvariable für die Exponentialverteilung	36
4.5.4.	Berechnung mit R	37
4.6.	Asymptotische Pivotvariable	39
4.6.1.	Konfidenzintervalle für die Binominalverteilung	39
5.	Signifikanztests	41
5.1.	Tests	41
5.2.	Signifikanztests	42
5.3.	Tests für den Mittelwert der Normalverteilung	43
5.3.1.	Einseitige Hypothesen	44
5.3.2.	Der p-Wert	45
5.3.3.	Einfache Nullhypothese	46
5.3.4.	Berechnung mit R	48
5.4.	Konstruktionsverfahren für kritische Bereiche	49
5.5.	Vergleich zweier Mittelwerte	50
5.5.1.	Der Studentsche t-Test	51
5.5.2.	Vergleich zweier Varianzen	54
5.5.3.	Der Welch-Test	56
5.5.4.	Berechnung mit R	56
6.	Lineare Modelle	59
6.1.	Lineare Modelle	59
6.2.	Die Methode der kleinsten Quadrate	61
6.3.	Lösung der Normalgleichungen mit R	63
7.	Lineare Regression	65
7.1.	Die Maximum-Likelihood-Schätzfunktionen	66
7.2.	Die Verteilung von B und Q	67
7.3.	Lineare Parameterfunktionen	71
7.4.	Konfidenzintervalle und Tests	73
7.5.	Berechnung mit R	74
8.	Lineare Hypothesen	76
8.1.	Die Datenstruktur	76
8.2.	Die Nullhypothese	78
8.2.1.	Regressionsansatz als Nullhypothese	78
8.2.2.	Die einfache Varianzanalyse	79
8.2.3.	Lineare Hypothesen	80
8.3.	Schätzung des Abstands	81
8.4.	Berechnung des Abstandsmaßes	83
8.5.	Statistische Eigenschaften	85

8.6. Kritischer Bereich	88
8.7. Die einfache Varianzanalyse	89
8.8. Test eines Regressionsansatzes	91
A. Statistische Berechnungen mit R	95
A.1. Vektoren	95
A.2. Wahrscheinlichkeitsrechnung	97
A.3. Statistische Kenngrößen von Stichproben	98
A.4. Tabellen	103
A.5. Faktoren	105
B. Beispiel-Datensätze	108
B.1. <code>bw.txt</code> : Bremsweg in Abhängigkeit von der Geschwindigkeit	108
B.2. <code>kk.txt</code> : Länge von Kuckuckseiern bei verschiedenen Vogelarten	108
B.3. <code>a10.txt</code> : Test eines Regressionsansatzes	109

1. Statistische Datenanalyse

In dieser Vorlesung beschäftigen wir uns mit einem Teilgebiet der mathematischen Statistik, die als statistische Datenanalyse bezeichnet wird, d.h. die Auswertung und Interpretation von experimentell gewonnenen fehlerbehafteten Messwerten, wobei die Fehler zufälligen Charakter besitzen.

Zur Veranschaulichung der Probleme, die wir im folgenden behandeln, beginnen wir mit einigen typischen Beispielen.

1.1. Beispiele

Beispiel 1.1 *In Tabelle 1 sind die Temperaturen in Grad Celsius aufgelistet, bei denen 59 Proben einer bestimmten Wachssorte zu schmelzen begannen. Wir gehen davon aus, dass es einen für diese Wachssorte charakteristischen Schmelzpunkt μ gibt und die gemessenen Werte von der Form $x_k = \mu + e_k$, $k = 1 \dots 59$, sind, wobei die Abweichungen e_k die von einer Messung zur anderen zufällig variierenden und für ein Experiment dieser Art unvermeidlichen Messungenauigkeiten sind.*

Aus diesen Messwerten soll ein Schätzwert für den Parameters μ zusammen mit einer Fehlerabschätzung berechnet werden.

63.78	63.45	63.58	63.08	63.40	64.42	63.27	63.10
63.34	63.50	63.83	63.63	63.27	63.30	63.83	63.50
63.36	63.86	63.34	63.92	63.88	63.36	63.36	63.51
63.51	63.84	64.27	63.50	63.56	63.39	63.78	63.92
63.92	63.56	63.43	64.21	64.24	64.12	63.92	63.53
63.50	63.30	63.86	63.93	63.43	64.40	63.61	63.03
63.68	63.13	63.41	63.60	63.13	63.69	63.05	62.85
63.31	63.66	63.60					

Tabelle 1: Schmelzpunkte von Wachsproben

Die alleinige Kenntnis der Messwerte reicht dazu offensichtlich nicht aus, wenn sowohl μ als auch die Fehler e_k vollständig unbekannt sind. Um eine solche Berechnung durchführen zu können, benötigt man zusätzlich Informationen über die e_k . In der Statistik geht man davon aus, dass diese Fehler Ergebnisse eines Zufallsmechanismus sind und dass das zugehörige Wahrscheinlichkeitsgesetz wenigstens teilweise bekannt ist.

Beispiel 1.2 *Die typische Situation, in der sich der Empfänger einer großen Anzahl eines bestimmten Warenartikels (z.B. eines DVD-Laufwerks) befindet, ist die folgende: Der Lieferant des Artikels garantiert, dass der Ausschussanteil höchstens 4% beträgt. Der Empfänger entnimmt der Lieferung 100 Artikel und findet 9 Ausschusstücke. Kann er daraus schließen, dass der Ausschussanteil in dieser Lieferung größer als 4% ist?*

Auf der Basis dieses Experiments kann diese Frage nicht definitiv beantwortet werden, denn aufgrund des Zufallscharakters dieses Verfahrens besteht durchaus die Möglichkeit, dass man bei 4% Ausschussanteil 9 Ausschussstücke unter den 100 untersuchten findet, wenn auch mit einer ziemlich kleinen Wahrscheinlichkeit. Lieferant und Empfänger müssen sich also zuvor noch darüber einigen, wie groß die Wahrscheinlichkeit sein darf, dass eine Lieferung irrtümlicherweise reklamiert wird.

Beispiel 1.3 Kurz nach der Einführung des metrischen Maßsystems Anfang der 1970er Jahre in Australien führte ein Statistikprofessor das folgende Experiment durch um zu testen, ob seine Studenten Längen in der Einheit „Meter“ korrekt einschätzen können. Eine Gruppe von 44 Studenten sollte die Breite des Hörsaals in Meter schätzen, eine zweite Gruppe von 69 Studenten sollte die Schätzung in „Fuss“ angeben. Tabelle 2 zeigt das Resultat. Frage: Wird die Breite dabei unterschiedlich eingeschätzt?

unit	width	unit	width	unit	width	unit	width	unit	width
metres	8	metres	15	feet	30	feet	40	feet	48
metres	9	metres	15	feet	30	feet	40	feet	50
metres	10	metres	15	feet	30	feet	40	feet	50
metres	10	metres	15	feet	30	feet	41	feet	50
metres	10	metres	16	feet	30	feet	41	feet	51
metres	10	metres	16	feet	32	feet	42	feet	54
metres	10	metres	16	feet	32	feet	42	feet	54
metres	10	metres	17	feet	33	feet	42	feet	54
metres	11	metres	17	feet	34	feet	42	feet	55
metres	11	metres	17	feet	34	feet	43	feet	55
metres	11	metres	17	feet	34	feet	43	feet	60
metres	11	metres	18	feet	35	feet	44	feet	60
metres	12	metres	18	feet	35	feet	44	feet	63
metres	12	metres	20	feet	36	feet	44	feet	70
metres	13	metres	22	feet	36	feet	45	feet	75
metres	13	metres	25	feet	36	feet	45	feet	80
metres	13	metres	27	feet	37	feet	45	feet	94
metres	14	metres	35	feet	37	feet	45		
metres	14	metres	38	feet	40	feet	45		
metres	14	metres	40	feet	40	feet	45		
metres	15	feet	24	feet	40	feet	46		
metres	15	feet	25	feet	40	feet	46		
metres	15	feet	27	feet	40	feet	47		
metres	15	feet	30	feet	40	feet	48		

Tabelle 2: Längenschätzung in Meter und Fuss

Beispiel 1.4 Ein häufig auftretendes Problem ist die Berechnung einer Kurve, die möglichst gut zu einem vorgegebenen Diagramm von Messpunkten passt. Die Abbildung 1 stellt für eine Reihe von Bremsversuchen mit einem bestimmten Fahrzeugtyp die zu verschiedenen Geschwindigkeiten gemessenen Bremswege dar. Gesucht ist eine Formel $w = g(v)$, die den Zusammenhang zwischen der Geschwindigkeit v und dem mittleren Bremsweg w möglichst gut wiedergibt.

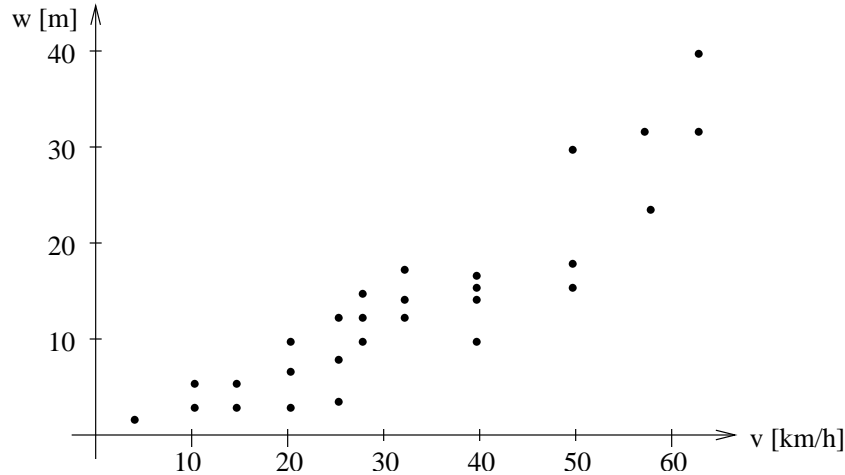


Abbildung 1: Bremsweg in Abhängigkeit von der Geschwindigkeit

Es gibt natürlich kein Computerprogramm, das auf die Eingabe der Koordinaten der Messpunkte hin eine Formel ausspuckt. Man kann sich lediglich eine eingeschränkte Menge von Funktionen vorgeben und daraus diejenige bestimmen, die am besten passt. In unserem Fall kann man sich davon leiten lassen, dass sich der Bremsweg zusammensetzt aus der Strecke, die während der Schrecksekunde zurückgelegt wird und der des eigentlichen Bremsvorgangs. Erstere ist proportional zur Geschwindigkeit v und letztere zur kinetischen Energie des Fahrzeugs bzw. zum Quadrat v^2 der Geschwindigkeit. Man wird also von dem Funktionentyp

$$g(v) = \alpha v + \beta v^2$$

mit noch offenen *Parametern* α und β ausgehen.

1.2. Statistische Experimente

Bei den oben beschriebenen Beispielen werden jeweils Zufallsexperimente an einer *statistischen Grundgesamtheit* durchgeführt. Diese Grundgesamtheit wird durch einen oder mehrere *Zustandsparameter* charakterisiert, deren Werte nicht bekannt sind und durch das Experiment ermittelt werden sollen.

Ein Rückschluss vom Ergebnis des Experiments auf den Zustand der Grundgesamtheit ist natürlich nur dann möglich, wenn zwischen diesen Größen ein Zusammenhang besteht.

Wenn dieser Zusammenhang dadurch gegeben ist, dass das Wahrscheinlichkeitsgesetz, nach dem das Ergebnis anfällt, vom Zustand der Grundgesamtheit abhängt, sprechen wir von einem **statistischen Experiment**.

Im ersten Beispiel kann man mit dem zentralen Grenzwertsatz argumentieren, dass die Messfehler Werte von normalverteilten Zufallsvariablen sind. Wenn man zusätzlich davon ausgehen kann, dass bei der Messung kein systematischer Fehler begangen wird, sondern die Fehler zufällig positiv oder negativ sind, so besitzen diese Zufallsvariablen den Erwartungswert Null. Wird jeweils das gleiche Experiment zur Bestimmung des Schmelzpunkts benutzt, so sind die Fehler etwa von der gleichen Größenordnung. Man kann daher annehmen, dass die Fehlervariablen alle die gleiche, wenn auch unbekanntere Varianz σ^2 besitzen. Schließlich wird man auch davon ausgehen können, dass die Experimente im stochastischen Sinne unabhängig voneinander durchgeführt werden.

Dies bedeutet, dass die Messwerte x_k die Ergebnisse einer n -fachen unabhängigen Wiederholung eines Experiments mit $\mathcal{N}(\mu, \sigma^2)$ -Verteilung sind. Die Parameter μ und σ^2 dieser Verteilung charakterisieren somit den Zustand der Grundgesamtheit. Sie sind nicht bekannt und man möchte aus den gewonnenen Messwerten Schätzwerte und Fehlerabschätzungen für diese Parameter berechnen.

Im zweiten Beispiel ist die Grundgesamtheit die Artikellieferung und ihr Zustand der Anteil p der Ausschussartikel ($0 \leq p \leq 1$). Ist der Umfang der Lieferung groß gegenüber der Anzahl $n = 100$ der zur Inspektion ausgewählten Artikel, so ändert sich der Ausschussanteil durch die Entnahme von Prüfaxemplaren praktisch nicht. Man kann daher davon ausgehen, dass das Experiment als Bernoulli-Versuchsreihe der Länge 100 mit der Erfolgswahrscheinlichkeit p für das Entnehmen eines defekten Artikels modelliert werden kann. Die Anzahl der defekten Exemplare unter den entnommenen ist daher binomialverteilt mit den Parametern $n = 100$ und p .

Im dritten Beispiel hat man es mit zwei Gruppen von Messwerten der Form $x_j = \mu_1 + e_j$ bzw. $y_k = \mu_2 + f_k$ zu tun, die analog zum ersten Beispiel als normalverteilt mit Parametern μ_1, σ_1^2 bzw. μ_2, σ_2^2 angenommen werden, wobei μ_1 bzw. μ_2 einen „mittleren Schätzwert“ der jeweiligen Studentengruppe darstellt.

Zu entscheiden ist in diesem Fall auf der Basis der Schätzwerte, ob μ_1 und μ_2 sich unterscheiden oder nicht.

Im letzten Beispiel schließlich wird man von dem Ansatz

$$w_k = \alpha v_k + \beta v_k^2 + e_k$$

für den Bremsweg w_k und die Geschwindigkeit v_k beim k -ten Versuch mit normalverteilten Fehlern e_k ausgehen und daraus Schätzwerte für die Parameter α und β berechnen.

1.3. Inhalt der Vorlesung

Die unmittelbar folgenden vier Kapitel stellen den theoretischen Teil der Vorlesung dar. Als Erstes führen wir das Konzept des **statistischen Raums** und der **Statistik** ein, die den Begriffen *Wahrscheinlichkeitsraum* und *Zufallsgröße* in der Wahrscheinlichkeitsrechnung entsprechen. Der statistische Raum ist das allgemeine mathematische Modell

für ein statistisches Experiment und eine Statistik repräsentiert den Vorgang, mit dem sich die praktische Statistik im wesentlichen beschäftigt: Der Zusammenfassung großer Datenmengen zu einigen wenigen aussagekräftigen Kenngrößen.

Daran anschließend werden die drei wichtigsten **Methoden** der induktiven Statistik behandelt: **Parameterschätzung**, **Konfidenzbereiche** und **Signifikanztests**. Der Begriff Parameterschätzung spricht für sich selbst, Konfidenzbereiche sind statistische Fehlerabschätzungen und Signifikanztests statistische Entscheidungsverfahren, wie sie in den Beispielen 1.2 und 1.3 vorgestellt wurden.

Der zweite mehr anwendungsbezogene Teil der Vorlesung befasst sich mit **Linearen Modellen**, d.h. der Anpassung von Funktionen an Messdaten, wie etwa im Beispiel 1.4. Die Parameterschätzung heißt in diesem Zusammenhang **Regression** und die Frage, ob eine vorgegebene Funktionenklasse überhaupt zu vorgegebenen Messwerten passt, wird im Kapitel über **Lineare Hypothesen** untersucht.

1.4. Statistik-Software

Für die Manipulation von Daten, die Berechnung von Schätzwerten und Konfidenzbereichen und die Durchführung von Signifikanztests gibt es eine große Menge von Softwarepaketen, z.B. SPSS, SAS, S-Plus und viele andere, die aufgrund des Funktionsumfangs und gemäß dem Gesetz von Angebot und Nachfrage zum Teil stolze Preise haben.

Im Rahmen dieser Vorlesung und für die zugehörigen Übungen wird das Programm R der „R Foundation for Statistical Computing“ benutzt, das unter der *GNU General Public License* steht und für viele Betriebssysteme von <http://www.r-project.org> heruntergeladen werden kann. Es entspricht im wesentlichen dem kommerziellen Paket S-PLUS.

Damit die Vorlesung auch etwas statistische Praxis vermittelt, wird in den meisten Kapiteln dargestellt, wie man die beschriebenen Verfahren mittels R anwendet. Die verwendeten Beispieldatensätze sind im Anhang B abgedruckt und können außerdem von der Vorlesungs-Webseite <http://www.am.uni-erlangen.de/home/graef/wr3> heruntergeladen werden.

Anhang A enthält eine sehr kurze und auf diese Vorlesung bezogene Einführung in R. Eine sehr schöne Beschreibung dieser Programmiersprache [17] steht auf der oben genannten R-Webseite zum Download.

1.5. Literatur

Die Literatur zu dem in dieser Vorlesung behandelten Thema ist unüberschaubar.

Zur Vorbereitung des ersten Teils der Vorlesung wurden die Bücher von Beaumont [1], Mood, Graybill und Boes [11] und Rice [13] verwendet, für den zweiten Teil wurden Schach, Schäfer [14] und Scheffé [15] herangezogen. Die Anwendung von R bei statistischen Verfahren wird von Everitt und Hothorn [5] beschrieben. Außerdem findet man jede Menge R-Tutorials im Internet.

Eine eher nichtmathematische Darstellung der Grundgedanken und Methoden der Statistik findet man in Bosch [2], dem noch aus der „Vor-PC-Epoche“ stammenden Buch

von Kreyszig [7] und bei Menges [9].

Die eigentliche und strenge *Mathematische* Statistik wird z.B. von Witting [19], Witting, Nölle [18] und Schmetterer [16] behandelt.

Aus Zeitgründen konnten in dieser Vorlesung leider die Themen Varianzanalyse und Verallgemeinerte Lineare Modelle nicht angesprochen werden. Dafür verweisen wir auf die oben genannten Bücher von Schach, Schäfer und Scheffé sowie Dobson [4].

2. Der statistische Raum

Analog zum Konzept des Wahrscheinlichkeitsraums ist der statistische Raum, den wir in diesem Abschnitt einführen, ein abstrakter Rahmen zur mathematischen Formulierung eines statistischen Experiments. Dabei müssen die folgenden drei Bestandteile konkretisiert werden: Die möglichen Zustände der Grundgesamtheit, die bei Durchführung des Experiments möglichen Ergebnisse und das Wahrscheinlichkeitsgesetz, das Ergebnisse und Zustände miteinander verbindet. Ein entsprechendes mathematisches Modell besteht daher aus den folgenden Größen:

1. **Die Grundgesamtheit:** Die statistische Grundgesamtheit, an der das Experiment durchgeführt wird, befindet sich in einem bestimmten **Zustand**, der mit dem griechischen Buchstaben θ bezeichnet wird. Die statistische Grundgesamtheit wird mathematisch durch die Menge aller Zustände charakterisiert, in der sie sich befinden kann. Diese Menge heißt der **Zustandsraum** und wird mit Θ bezeichnet.
2. **Die Stichprobe:** Das Ergebnis des Zufallsexperiments, das an der Grundgesamtheit durchgeführt wird, heißt die **Stichprobe** und wird mit dem Buchstaben x bezeichnet. Die Menge \mathcal{X} der möglichen Ergebnisse dieses Zufallsexperiments heißt der **Stichprobenraum**. Es wird ohne weitere Erwähnung vorausgesetzt, dass auf dem Stichprobenraum eine σ -Algebra $\mathcal{A}(\mathcal{X})$ von Teilmengen festgelegt wird.
3. **Die Verteilungsannahme:** Das Wahrscheinlichkeitsgesetz, gemäß dem die Stichprobe zustande kommt, hängt vom Zustand θ ab, in dem sich die Grundgesamtheit befindet. Es wird mit P_θ bezeichnet. Die Menge $\mathcal{P} = \{P_\theta ; \theta \in \Theta\}$ aller dieser Wahrscheinlichkeitsgesetze heißt die **Verteilungsannahme**.

Definition 2.1 *Ein Tripel $(\Theta, \mathcal{X}, \mathcal{P})$ von Größen der oben beschriebenen Art heißt ein statistischer Raum.*

Interpretation: Man geht davon aus, dass das Ergebnis x des statistischen Experiments gemäß dem Wahrscheinlichkeitsgesetz $P_{\hat{\theta}}$ anfällt, wenn sich die Grundgesamtheit im Zustand $\hat{\theta}$ befindet.

Die Ausgangssituation der Statistik ist aber die, dass uns dieser Zustand nicht bekannt ist, sondern lediglich das Ergebnis x des entsprechenden Zufallsexperiments vorliegt. Das **Problem** der Statistik besteht daher ganz allgemein darin zu **entscheiden, welcher Zustand θ bzw. welche Wahrscheinlichkeit P_θ am besten zu dem vorliegenden Ergebnis x passt.**

2.1. Beispiele

Wir stellen die statistischen Räume zu den vier Beispielen aus dem ersten Kapitel auf.

2.1.1. Die Normalverteilungsannahme

Beispiel 1.1 ist *das* klassische statistische Experiment der statistischen Datenanalyse: die n-fache Wiederholung eines Zufallsexperiments mit $\mathcal{N}(\mu, \sigma^2)$ -Verteilung. Die unbekanntes Größen sind der Mittelwert μ und die Varianz σ^2 . Der Zustandsraum ist daher

$$\Theta = \{\theta = (\mu, \sigma^2) ; \mu \in \mathbb{R}, \sigma^2 > 0\}$$

Die Stichprobe besteht aus n reellwertigen Messwerten x_1, x_2, \dots, x_n , die im angenommenen Zusammenhang eigentlich nur positiv sein können. Da die Normalverteilung mit Mittelwert im Wertebereich um 60 negativen Ergebnissen eine vernachlässigbar kleine Wahrscheinlichkeit zuordnet, nehmen wir an, dass die einzelnen Messwerte gemäß einer $\mathcal{N}(\mu, \sigma^2)$ -Normalverteilung mit der Dichte

$$f_1(t; \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(t-\mu)^2}$$

anfallen. Kann man davon ausgehen, dass die einzelnen Messungen stochastisch unabhängig voneinander stattfinden und fasst man die n Messwerte zu einem Vektor zusammen, so ergibt sich als Stichprobenraum die Menge

$$\mathcal{X} = \mathbb{R}^n = \{x = (x_1, \dots, x_n) ; x_k \in \mathbb{R}\}$$

und die Verteilungsannahme besteht aus den (Produkt-)Wahrscheinlichkeiten $P_{(\mu, \sigma^2)}$ mit den Dichten

$$\begin{aligned} f_n(x; \theta) &= f_n(x_1, x_2, \dots, x_n; \mu, \sigma^2) \\ &= f_1(x_1; \mu, \sigma^2) f_1(x_2; \mu, \sigma^2) \cdots f_1(x_n; \mu, \sigma^2) \\ &= \prod_{k=1}^n \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(x_k - \mu)^2} \\ &= \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n e^{-\frac{1}{2\sigma^2} \sum_{k=1}^n (x_k - \mu)^2} \end{aligned} \tag{2.1}$$

2.1.2. Bernoulli-Versuchsreihen

Beim Beispiel 1.2 besteht die statistische Grundgesamtheit aus einer großen Lieferung gleicher Artikel, die entweder intakt oder defekt sein können. Im dargestellten Zusammenhang ist der Zustand dieser Grundgesamtheit charakterisiert durch den Anteil an defekten Artikeln, also einer reellen Zahl θ mit $0 \leq \theta \leq 1$:

$$\Theta = [0, 1] = \{\theta \in \mathbb{R} ; 0 \leq \theta \leq 1\}$$

Zur Prüfung wird n-mal hintereinander ein Exemplar zufällig aus der Lieferung entnommen und inspiziert, d.h. es werden n Bernoulli-Experimente durchgeführt mit den Ergebnissen $\delta_k = 0$, falls das k-te Exemplar intakt und $\delta_k = 1$, falls das k-te Exemplar defekt ist.

Bei der Entnahme des ersten Artikels ist die Erfolgswahrscheinlichkeit für das Ziehen eines defekten Stücks gleich dem Anteil θ an defekten Stücken insgesamt. Durch die Entnahme von Probeexemplaren ändert sich dieser Anteil. Falls aber der Umfang der Lieferung groß gegenüber der Anzahl zu prüfender Stücke ist, kann man diese Änderung vernachlässigen und annehmen, dass es sich bei dem gesamten Vorgang um eine Bernoulli-Versuchsreihe der Länge n mit dem unbekanntem Anteil θ als Erfolgswahrscheinlichkeit handelt. Der Stichprobenraum ist demgemäß

$$\mathcal{X} = \{\delta = (\delta_1, \delta_2, \dots, \delta_n) ; \delta_k \in \{0, 1\}\}$$

und die Wahrscheinlichkeiten P_θ der Verteilungsannahme sind durch die Wahrscheinlichkeitsfunktionen

$$f_n(\delta; \theta) = f_n(\delta_1, \delta_2, \dots, \delta_n; \theta) = \theta^{\delta_1 + \delta_2 + \dots + \delta_n} (1 - \theta)^{n - (\delta_1 + \delta_2 + \dots + \delta_n)} \quad (2.2)$$

charakterisiert.

Man fragt sich in diesem Zusammenhang natürlich, ob es eine Rolle spielt, in welcher Reihenfolge die Nullen und Einsen anfallen bzw., ob es nicht reicht, zur Schätzung von θ nur die Anzahl $k = \delta_1 + \delta_2 + \dots + \delta_n$ der gezogenen defekten Stücke heranzuziehen, wo doch die Wahrscheinlichkeitsfunktion nur von dieser Größe abhängt. Wir werden diese Frage im Abschnitt über suffiziente Statistiken wieder aufgreifen.

2.1.3. Vergleich zweier Mittelwerte

Im Beispiel 1.3 werden zwei Messreihen $x = (x_1, x_2, \dots, x_{n_1})$ und $y = (y_1, y_2, \dots, y_{n_2})$ ermittelt. Wenn man analog zum Beispiel 1.1 annimmt, dass die x_j gemäß einer $\mathcal{N}(\mu_1, \sigma_1^2)$ - und die y_k gemäß einer $\mathcal{N}(\mu_2, \sigma_2^2)$ -Verteilung entstanden sind und die Einzelexperimente stochastisch unabhängig voneinander stattgefunden haben, ist der Parameterraum

$$\Theta = \{\theta = (\mu_1, \mu_2, \sigma_1^2, \sigma_2^2) ; \mu_i \in \mathbb{R}, \sigma_i^2 > 0\},$$

der Stichprobenraum für das Gesamtexperiment

$$\mathcal{X} = \{z = (x, y) = (x_1, x_2, \dots, x_{n_1}, y_1, y_2, \dots, y_{n_2}) ; x_j, y_k \in \mathbb{R}\} = \mathbb{R}^{n_1 + n_2}$$

und die Dichte der Verteilung $P_\theta = P_{(\mu_1, \mu_2, \sigma_1^2, \sigma_2^2)}$ gleich

$$f(z; \theta) = \left(\frac{1}{\sqrt{2\pi\sigma_1^2}} \right)^{n_1} \left(\frac{1}{\sqrt{2\pi\sigma_2^2}} \right)^{n_2} e^{-\frac{1}{2\sigma_1^2} \sum_{j=1}^{n_1} (x_j - \mu_1)^2 - \frac{1}{2\sigma_2^2} \sum_{k=1}^{n_2} (y_k - \mu_2)^2} \quad (2.3)$$

2.2. Allgemeine Voraussetzungen

Wir befassen uns in dieser Vorlesung nicht in voller Allgemeinheit mit der mathematischen Statistik. Um mit den Hilfsmitteln der klassischen Analysis arbeiten zu können, setzen wir voraus, dass die statistischen Räume die nachstehend beschriebenen Eigenschaften besitzen.

2.2.1. Parametrische Statistik

Wir nehmen stets an, dass der Zustand der Grundgesamtheit wie in den obigen Beispielen durch einen reellen Vektor

$$\theta = (\theta_1, \theta_2, \dots, \theta_p)$$

beschrieben werden kann. Man spricht in diesem Fall von **parametrischer Statistik**. θ bzw. seine Komponenten θ_j heißen **Parameter** und der Zustandsraum $\Theta \subset \mathbb{R}^p$ der **Parameterraum**.

Ein Fall von nichtparametrischer Statistik wäre z.B., dass die Verteilungsannahme aus allen Wahrscheinlichkeiten mit stetiger Verteilungsfunktion besteht. Der Zustand ist dann die jeweilige Verteilungsfunktion.

2.2.2. Absolutstetige und diskrete Verteilungsannahme

Wir betrachten im folgenden nur Verteilungsannahmen mit Wahrscheinlichkeiten, die durch Dichten oder Wahrscheinlichkeitsfunktionen charakterisiert werden können.

1. Absolutstetige Verteilungsannahmen. Hier ist der Stichprobenraum die Menge $\mathcal{X} = \mathbb{R}^n$ der n-dimensionalen reellen Vektoren

$$x = (x_1, x_2, \dots, x_n)$$

zusammen mit der σ -Algebra \mathcal{B}_n der Borelschen Mengen und die Wahrscheinlichkeiten P_θ sind absolutstetig mit Dichten $f(\cdot; \theta)$:

$$P_\theta(B) = \int 1_B(x) f(x; \theta) dx = \int_B f(x; \theta) dx$$

2. Diskrete Verteilungsannahmen. Der Stichprobenraum ist eine endliche oder abzählbare Menge mit der Menge aller Teilmengen als σ -Algebra und die Wahrscheinlichkeiten P_θ sind durch ihre Wahrscheinlichkeitsfunktionen $f(\cdot; \theta)$ charakterisiert:

$$P_\theta(B) = \sum_{x \in B} f(x; \theta)$$

Wir werden daher häufig die Verteilungsannahme in der Form

$$\mathcal{P} = \{f(x; \theta) ; \theta \in \Theta\}$$

darstellen, wobei die Funktionen f die zu den P_θ gehörigen Dichten bzw. Wahrscheinlichkeitsfunktionen sind.

2.2.3. Unabhängige Wiederholungen

Die am häufigsten vertretene Art von statistischem Experiment ist die n-malige unabhängige Wiederholung eines *Einzelexperiments* beschrieben durch einen statistischen Raum $(\Theta, \mathcal{X}_0, \mathcal{P}_0)$ mit einer Verteilungsannahme $\mathcal{P}_0 = \{P_\theta^0 ; \theta \in \Theta\} = \{f_\theta(\cdot; \theta) ; \theta \in \Theta\}$.

Der Stichprobenraum ist dann das kartesische Produkt

$$\mathcal{X} = \mathcal{X}_0^n = \{x = (x_1, x_2, \dots, x_n) ; x_k \in \mathcal{X}_0\}$$

mit der Produkt- σ -Algebra und die Verteilungsannahme ist dadurch charakterisiert, dass für jedes θ der Wahrscheinlichkeitsraum $(\mathcal{X}, \mathcal{A}(\mathcal{X}), P_\theta)$ das Produkt von n Exemplaren des Wahrscheinlichkeitsraums $(\mathcal{X}_0, \mathcal{A}(\mathcal{X}_0), P_\theta^0)$ ist, d.h. dass P_θ die Dichte bzw. Wahrscheinlichkeitsfunktion

$$f(x; \theta) = f(x_1, x_2, \dots, x_n; \theta) = f_0(x_1; \theta) f_0(x_2; \theta) \cdots f_0(x_n; \theta)$$

besitzt.

Um diese Situation mit wenigen Worten zu beschreiben, werden wir im folgenden häufig die Formulierung „**n-malige Wiederholung eines P_θ^0 -Experiments**“ verwenden.

2.3. Statistiken

Im normalen Sprachgebrauch versteht man unter einer Statistik eine Tabelle oder eine graphische Darstellung, die die Stichprobe in einer übersichtlichen und anschaulichen Form zusammenfasst. Aus mathematischer Sicht ist eine **Statistik** daher einfach eine messbare Abbildung

$$T : \mathcal{X} \rightarrow \mathcal{T}$$

des Stichprobenraums \mathcal{X} in einen Raum \mathcal{T} mit (i.a. sehr viel) kleinerer Dimension. Wesentlich ist dabei, dass die Vorschrift zur Berechnung des Funktionswertes $T(x)$ **nicht** den (unbekannten) Parameter θ enthält, sondern nur die Messergebnisse x_k .

Praktisch gesprochen ist eine Statistik eine Rechenvorschrift, die nach Durchführung des Experiments mit den dann bekannten Daten ausgeführt werden kann.

2.3.1. Die Verteilung einer Statistik

Befindet sich die Grundgesamtheit im Zustand θ , so wird das statistische Experiment durch den Wahrscheinlichkeitsraum $(\mathcal{X}, \mathcal{A}(\mathcal{X}), P_\theta)$ beschrieben. Eine Statistik T ist dann eine Zufallsgröße auf diesem Wahrscheinlichkeitsraum, deren Verteilung im Gegensatz zur Rechenformel für $T(x)$ sehr wohl vom Zustand abhängt:

$$P_\theta^T(B) = P_\theta(T \in B) = P_\theta\{x \in \mathcal{X} ; T(x) \in B\} \quad (2.4)$$

Ist die Statistik reellwertig: $\mathcal{T} \subset \mathbb{R}$, so kann man — soweit existent — den Erwartungswert und die Varianz betrachten. Diese Größen hängen natürlich vom Zustand der Grundgesamtheit, d.h. vom Parameter θ ab:

$$\mathcal{E}_\theta T = \int T(x) P_\theta(dx) = \begin{cases} \int T(x) f(x; \theta) dx & \text{bei absolutstetiger Verteilung} \\ \sum_{x \in \mathcal{X}} T(x) f(x; \theta) & \text{bei diskreter Verteilung} \end{cases} \quad (2.5)$$

$$\text{var}_\theta(T) = \mathcal{E}_\theta(T - \mathcal{E}_\theta T)^2 \quad (2.6)$$

Auf der Bernoulli-Versuchsreihe des Beispiels 1.2 ist bekanntermaßen die Statistik $T(\delta) = \delta_1 + \delta_2 + \dots + \delta_n$ als die *Summe der Erfolge* binomialverteilt mit Parametern n und θ :

$$P_\theta^T\{t\} = f^T(t; \theta) = \binom{n}{t} \theta^t (1 - \theta)^{n-t}$$

mit dem Erwartungswert $\mathcal{E}_\theta T = n\theta$ und der Varianz $\text{var}_\theta(T) = n\theta(1 - \theta)$

2.3.2. Suffiziente Statistiken

Der Zweck einer Statistik besteht im allgemeinen darin, die große Menge an Einzeldaten x_k , aus denen das Ergebnis x besteht, auf einige wenige Kenngrößen zu komprimieren, ohne dass Information über den Zustand, die in den Werten x_k enthalten ist, verlorengeht.

Eine Statistik, die diese Eigenschaft besitzt, heißt **suffizient**. Die exakte mathematische Definition der Suffizienz beinhaltet das Konzept einer *bedingten Wahrscheinlichkeitsverteilung*, mit dem wir uns im Rahmen dieser Vorlesung nicht beschäftigen wollen. In sehr vager aber hoffentlich intuitiv verständlicher Form heißt eine Statistik $T(x)$ suffizient für θ , wenn unter der Bedingung, dass der Wert $t = T(x)$ der Statistik bekannt ist, die bedingte Wahrscheinlichkeitsverteilung des Ergebnisses nicht mehr vom Zustand θ abhängt:

$$P_\theta(B|T = t) = p(B|t)$$

Denn aus einer Funktion, die nicht vom Argument θ abhängt, lässt sich keine Information über dieses Argument entnehmen. Infolgedessen muss die gesamte (statistische) Information über θ in der Verteilung P_θ^T der Statistik T stecken.

Aus dieser Definition ergibt sich die folgende Charakterisierung von suffizienten Statistiken:

Satz 2.1 Sei $(\Theta, \mathcal{X}, \mathcal{P})$ ein statistischer Raum mit der Verteilungsannahme

$\mathcal{P} = \{f(\cdot; \theta) ; \theta \in \Theta\}$ und $T : \mathcal{X} \rightarrow \mathcal{T}$ eine Statistik.

Die Statistik T ist suffizient für θ , wenn es eine Funktion $h : \mathcal{X} \rightarrow \mathbb{R}$ und eine Funktion $g : \mathcal{T} \times \Theta \rightarrow \mathbb{R}$ gibt, so dass für alle $x \in \mathcal{X}$ und alle $\theta \in \Theta$ gilt

$$f(x; \theta) = h(x)g(T(x), \theta)$$

Beispiele

Die Wahrscheinlichkeitsfunktion (2.2) zu unserem Wareneingangsbeispiel hat die Form

$$f_n(\delta; \theta) = \theta^{\delta_1 + \delta_2 + \dots + \delta_n} (1 - \theta)^{n - (\delta_1 + \delta_2 + \dots + \delta_n)} = h(\delta)g(T(\delta), \theta)$$

mit $T(\delta) = \delta_1 + \delta_2 + \dots + \delta_n$, $h(\delta) = 1$ und $g(t, \theta) = \theta^t (1 - \theta)^{n-t}$. T ist also suffizient.

Für die Dichte (2.1) der n -fachen Wiederholung eines $\mathcal{N}(\mu, \sigma^2)$ -Experiments,

$$f_n(x; \theta) = \left(\frac{1}{\sqrt{2\pi\sigma^2}} \right)^n e^{-\frac{1}{2\sigma^2} \sum_{k=1}^n (x_k - \mu)^2},$$

formt man den Term $q(x, \mu) = \sum_{k=1}^n (x_k - \mu)^2$ im Exponenten etwas um:

$$q(x, \mu) = \sum_{k=1}^n x_k^2 - 2\mu \sum_{k=1}^n x_k + n\mu^2 = T_2(x) - 2\mu T_1(x) + n\mu^2$$

Dann hat die Dichte f_n die Gestalt

$$f_n(x; \theta) = h(x)g((T_1(x), T_2(x)), (\mu, \sigma^2)) = h(x)g(T(x), \theta)$$

mit $h(x) = 1$ und der zweidimensionalen Statistik

$$T(x) = (T_1(x), T_2(x)) = \left(\sum_{k=1}^n x_k, \sum_{k=1}^n x_k^2 \right)$$

Eine suffiziente Statistik ist weder eindeutig bestimmt, noch enthält die Definition der Suffizienz die Forderung, dass das Ergebnis komprimiert werden muss. Die Funktion $T(x) = x$, die gar nichts komprimiert, ist gemäß obigem Satz suffizient.

Die Suche nach weitestgehender Komprimierung, d.h. nach einer **minimal suffizienten Statistik** bleibt als zusätzliche Aufgabe bestehen. Ohne näher darauf einzugehen, begnügen wir uns mit der Faustregel, dass eine Statistik dann minimal suffizient ist, wenn die Dimension des Werts $T(x)$ mit der Dimension des Parametervektors θ übereinstimmt.

2.3.3. Die Koordinatenvariablen

Eine wichtige Rolle spielen im folgenden die sogenannten Koordinatenvariablen. Ist der Stichprobenraum $\mathcal{X} = \mathcal{X}_0^n$ das kartesische Produkt von n Exemplaren der Menge \mathcal{X}_0 , d.h. die Ergebnisse sind von der Form $x = (x_1, x_2, \dots, x_n)$ mit $x_k \in \mathcal{X}_0$, so heißt die durch

$$X_k(x_1, x_2, \dots, x_n) = x_k \tag{2.7}$$

definierte Statistik $X_k : \mathcal{X} \rightarrow \mathcal{X}_0$ die **k-te Koordinatenvariable**.

Satz 2.2 *Ist der statistische Raum $(\Theta, \mathcal{X}, \mathcal{P})$ die n -fache Wiederholung von $(\Theta, \mathcal{X}_0, \mathcal{P}_0)$ mit $\mathcal{P}_0 = \{P_\theta^0 ; \theta \in \Theta\}$, so gilt für jedes $\theta \in \Theta$:*

- (a) *Die Koordinatenvariablen X_1, \dots, X_n sind auf $(\mathcal{X}, \mathcal{A}(\mathcal{X}), P_\theta)$ stochastisch unabhängig.*
- (b) *Die Verteilung von X_k bezüglich $(\mathcal{X}, \mathcal{A}(\mathcal{X}), P_\theta)$ ist P_θ^0 :*

$$P_\theta(X_k \in B) = P_\theta^0(B)$$

Der Beweis dieses Satzes ergibt sich unmittelbar aus den in der Vorlesung Wahrscheinlichkeitsrechnung I [6] hergeleiteten Sätzen über mehrdimensionale Verteilungen.

Beispiel: Ist $(\Theta, \mathcal{X}, \mathcal{P})$ die n -fache Wiederholung eines $\mathcal{N}(\mu, \sigma^2)$ -Experiments, so sind die Koordinatenvariablen auf $(\mathbb{R}^n, \mathcal{B}_n, P_{(\mu, \sigma^2)})$ stochastisch unabhängig und $\mathcal{N}(\mu, \sigma^2)$ -verteilt.

Anstatt einen statistischen Raum $(\Theta, \mathcal{X}, \mathcal{P})$ als die n -fache Wiederholung von $(\Theta, \mathcal{X}_0, \mathcal{P}_0)$ zu charakterisieren, werden wir häufig diese Situation mit der Formulierung „**Die Koordinatenvariablen X_k sind stochastisch unabhängig und P_θ^0 -verteilt.**“ beschreiben.

2.4. Parameterfunktionen

Meist interessiert man sich nicht für den ganzen Parametervektor $\theta = (\theta_1, \dots, \theta_p)$, sondern nur für eine einzelne Komponente θ_j oder für eine Funktion der Komponenten, z.B. für $\mu + 3\sqrt{\sigma^2}$ bei $\theta = (\mu, \sigma^2)$. Wir definieren daher

Definition 2.2 Eine Abbildung $\tau : \Theta \rightarrow \mathbb{R}$ heißt eine **Parameterfunktion**.

3. Parameterschätzung

Um auf der Basis der Stichprobe Aussagen über den unbekannt Parameter $\theta = (\theta_1, \theta_2, \dots, \theta_p)$ treffen zu können, benötigt man als Erstes ein Rechenverfahren, d.h. eine Statistik $T : \mathcal{X} \rightarrow \mathbb{R}^p$ so, dass man die Werte $T(x)$ als *gute* Schätzwerte für den unbekannt Parameter θ bzw. einer seiner Komponenten θ_j ansehen kann.

Vor der Frage, was *gute* Schätzfunktionen sind, bzw. der Frage nach passenden Gütekriterien steht dabei die nach einem Konstruktionsprinzip, das zu jeder Verteilungsannahme *vernünftige* Schätzfunktionen für den Zustand liefert.

3.1. Die Maximum-Likelihood-Methode

Der Grundgedanke der von R.A. Fisher vorgeschlagenen Maximum-Likelihood-Methode besteht in der Umkehrung des Problems der Wahrscheinlichkeitsrechnung, bei bekanntem Wahrscheinlichkeitsgesetz den Ausgang des Experiments, d.h. das Ergebnis zu prognostizieren:

Ist der Zustand $\hat{\theta}$ der Grundgesamtheit und damit die Wahrscheinlichkeit $P_{\hat{\theta}}$ bekannt, so wird man erwarten, dass ein Ergebnis zustande kommt, das bezüglich $P_{\hat{\theta}}$ eine große Wahrscheinlichkeit besitzt.

Bei absolutstetigen Verteilungen kann man dazu aber nicht die Wahrscheinlichkeit $P_{\hat{\theta}}\{x\}$ des Elementarereignisses x verwenden, denn die ist für alle x gleich Null. Aus der Beziehung

$$P_{\hat{\theta}}(B) = \int_B f(x; \hat{\theta}) dx \approx f(x; \hat{\theta})|B|$$

bzw. der Interpretation des Integrals als „Fläche“ unter der durch $f(x; \hat{\theta})$ beschriebenen „Kurve“ wird man unter einem Ergebnis x , das eine große Wahrscheinlichkeit besitzt, eines verstehen, für das der Funktionswert $f(x; \hat{\theta})$ groß ist.

Wenn man sich bei bekanntem Zustand $\hat{\theta}$ auf eine eindeutige Prognose festlegen soll, wird man daher ein \hat{x} mit der Eigenschaft

$$f(\hat{x}, \hat{\theta}) \geq f(x; \hat{\theta}) \text{ für alle } x \in \mathcal{X}$$

wählen.

Die Situation in der Statistik ist umgekehrt die, dass das Ergebnis, d.h. die Stichprobe \hat{x} bekannt ist und der Zustand gesucht wird, der zu diesem Ergebnis am Besten passt. In Umkehrung des obigen Prognoseprinzips wird man sich vernünftigerweise für einen Zustand $\hat{\theta}$ entscheiden, der dem bekannten Ergebnis \hat{x} eine möglichst große Wahrscheinlichkeit zuordnet, d.h. ein $\hat{\theta}$ mit der Eigenschaft

$$f(\hat{x}; \hat{\theta}) \geq f(\hat{x}; \theta) \text{ für alle } \theta \in \Theta$$

Daraus ergibt sich die **Maximum-Likelihood-Methode**:

Gegeben sei eine Verteilungsannahme $\mathcal{P} = \{P_{\theta} ; \theta \in \Theta\}$ mit Verteilungen P_{θ} , die die Dichten oder Wahrscheinlichkeitsfunktionen $f(x; \theta)$ besitzen. Die Funktion $L(\theta; x) =$

$f(x; \theta)$, d.h. die Dichte bzw. Wahrscheinlichkeitsfunktion als Funktion des Parameters θ , heißt die **Likelihood-Funktion**.

Ein $\hat{\theta}_x \in \mathbb{R}^p$ mit der Eigenschaft

$$L(\hat{\theta}_x; x) \geq L(\theta; x) \text{ für alle } \theta \in \Theta$$

heißt **Maximum-Likelihood-Schätzwert** für θ zur Stichprobe x .

Die Statistik $T : \mathcal{X} \rightarrow \mathbb{R}^p$ mit $T(x) = \hat{\theta}_x$ für alle $x \in \mathcal{X}$ heißt eine **Maximum-Likelihood-Schätzfunktion** für den Parameter θ .

3.1.1. Die Normalverteilung

Die Likelihood-Funktion zur n-fachen Wiederholung eines Experiments mit $\mathcal{N}(\mu, \sigma^2)$ -Verteilung lautet

$$L(\mu, \sigma^2) = \left(\frac{1}{\sqrt{2\pi\sigma^2}} \right)^n e^{-\frac{1}{2\sigma^2} \sum_{k=1}^n (x_k - \mu)^2}$$

Zur Bestimmung des Maximums kann man an Stelle der Funktion L selbst auch den natürlichen Logarithmus $L^*(\mu, \sigma^2) = \ln L(\mu, \sigma^2)$ heranziehen. Da die Logarithmusfunktion monoton steigt:

$$\ln x < \ln y \quad \text{für } 0 < x < y$$

ist $L(\mu_1, \sigma_1^2) < L(\mu_2, \sigma_2^2)$ genau dann, wenn $L^*(\mu_1, \sigma_1^2) < L^*(\mu_2, \sigma_2^2)$. Die beiden Funktionen besitzen daher die gleichen Maximalstellen.

Die Anwendung der Rechenregeln $\ln(ab) = \ln(a) + \ln(b)$, $\ln(a^x) = x \ln(a)$ und $\ln(e^x) = x$ ergibt für die Funktion L^* die Gestalt

$$L^*(\mu, \sigma^2) = -\frac{n}{2} \ln(2\pi) - \frac{n}{2} \ln(\sigma^2) - \frac{1}{\sigma^2} Q(\mu)$$

mit

$$Q(\mu) = \sum_{k=1}^n (x_k - \mu)^2$$

1.Schritt: Bei festgehaltenem $\sigma^2 > 0$ wird $L^*(\mu, \sigma^2)$ in Abhängigkeit von μ umso größer, je kleiner der Wert von $Q(\mu)$ ist.

Die notwendige Bedingung für eine Minimalstelle der Funktion Q :

$$\frac{dQ}{d\mu} = -2 \left(\sum_{k=1}^n x_k - n\mu \right) = 0$$

liefert als einzige Lösung die Zahl

$$\hat{\mu} = \frac{x_1 + x_2 + \dots + x_n}{n} =: \bar{x}$$

die wegen $\frac{d^2Q}{d\mu^2} = 2n > 0$ eine Minimalstelle ist.

Für beliebige $\mu \in \mathbb{R}$ und $\sigma^2 > 0$ gilt daher

$$L^*(\mu, \sigma^2) \leq L^*(\bar{x}, \sigma^2)$$

und wir müssen daher nur noch die Maximalstelle der Funktion

$$S^*(\sigma^2) = L^*(\bar{x}, \sigma^2) = -\frac{n}{2} \ln(2\pi) - \frac{n}{2} \ln(\sigma^2) - \frac{1}{2\sigma^2} Q(\bar{x})$$

in Abhängigkeit von der Variablen σ^2 suchen.

Zur Erinnerung: σ^2 ist hier **nicht** als das Quadrat einer Zahl σ zu verstehen, sondern als Symbol für eine positive Zahl.

Die notwendige Bedingung für eine Maximalstelle der Funktion S^* lautet

$$\frac{dS^*}{d\sigma^2}(\sigma^2) = -\frac{n}{2\sigma^2} + \frac{1}{2(\sigma^2)^2} Q(\bar{x}) = 0$$

mit der Lösung

$$\hat{\sigma}^2 = \frac{1}{n} Q(\bar{x}) = \frac{1}{n} \sum_{k=1}^n (x_k - \bar{x})^2$$

Da die Ableitung der Funktion S^* nach σ^2 für $\sigma^2 < \hat{\sigma}^2$ positiv und für $\sigma^2 > \hat{\sigma}^2$ negativ ist, handelt es sich um eine Maximalstelle.

Als **Maximum-Likelihood-Schätzfunktion** für den Parameter $\theta = (\mu, \sigma^2)$ der Normalverteilung enthalten wir damit die Funktion

$$T(x) = \left(\bar{x}, \frac{1}{n} \sum_{k=1}^n (x_k - \bar{x})^2 \right) \quad (3.8)$$

Die beiden Komponenten

$$T_1(x) = \bar{x} \quad (3.9)$$

und

$$T_2(x) = \frac{1}{n} \sum_{k=1}^n (x_k - \bar{x})^2 \quad (3.10)$$

nennt man naheliegenderweise die Maximum-Likelihood-Schätzfunktion für den Mittelwert μ bzw. die Varianz σ^2 .

3.1.2. Umrechnungsformeln

Für die praktische Berechnung und die folgenden theoretischen Überlegungen muss der Ausdruck $\sum_{k=1}^n (x_k - \bar{x})^2$ etwas umgeformt werden:

$$\begin{aligned} \sum_{k=1}^n (x_k - \bar{x})^2 &= \sum_{k=1}^n x_k^2 - 2\bar{x} \sum_{k=1}^n x_k + n\bar{x}^2 = \sum_{k=1}^n x_k^2 - 2\bar{x} \cdot n\bar{x} + n\bar{x}^2 \\ &= \sum_{k=1}^n x_k^2 - n\bar{x}^2 \end{aligned} \quad (3.11)$$

$$= \sum_{k=1}^n x_k^2 - \frac{1}{n} \left(\sum_{k=1}^n x_k \right)^2 \quad (3.12)$$

3.1.3. Suffiziente Statistiken

Besitzt die Dichte oder Wahrscheinlichkeitsfunktion die Gestalt

$$f(x; \theta) = h(x)g(T(x), \theta)$$

mit $h(x) > 0$, so erhält man einen Maximum-Likelihood-Schätzwert durch Maximieren von $g(T(x), \theta)$ bezüglich θ . Aufgrund der Gestalt der Funktion g hängt dieser Schätzwert nur von $T(x)$ ab, woraus wir schließen:

Satz 3.1 *Eine Maximum-Likelihood-Schätzfunktion ist stets Funktion einer suffizienten Statistik.*

Dass im obigen Beispiel die Statistik (3.8) suffizient ist, ersieht man aus der folgenden Umformung des Exponenten der Dichte:

$$\begin{aligned} \sum_{k=1}^n (x_k - \mu)^2 &= \sum_{k=1}^n [(x_k - \bar{x}) + (\bar{x} - \mu)]^2 \\ &= \sum_{k=1}^n (x_k - \bar{x})^2 + 2(\bar{x} - \mu) \sum_{k=1}^n (x_k - \bar{x}) + n(\bar{x} - \mu)^2 \end{aligned}$$

Wegen

$$\sum_{k=1}^n (x_k - \bar{x}) = \sum_{k=1}^n x_k - n\bar{x} = \sum_{k=1}^n x_k - \sum_{k=1}^n x_k = 0$$

erhält man

$$\begin{aligned} \sum_{k=1}^n (x_k - \mu)^2 &= \sum_{k=1}^n (x_k - \bar{x})^2 + n(\bar{x} - \mu)^2 \\ &= T_2(x) + n(T_1(x) - \mu)^2 \end{aligned}$$

Im Fall der Normalverteilung sind daher $\tilde{T}(x) = (\sum_{k=1}^n x_k, \sum_{k=1}^n x_k^2)$ und $T(x) = (\bar{x}, \sum_{k=1}^n (x_k - \bar{x})^2)$ minimal suffizient. Auch minimal suffiziente Statistiken sind nicht eindeutig bestimmt. man kann sie aber ineinander umrechnen, wie der vorherige Abschnitt zeigt.

3.1.4. Binomialverteilung

Da die Statistik $T(\delta) = \delta_1 + \dots + \delta_n$ für den Parameter der Bernoulli-Versuchsreihe minimal suffizient ist, kann man sich bei der statistischen Analyse bei der Wareneingangskontrolle auf die Werte und die Verteilung dieser Statistik zurückziehen, d.h. auf den statistischen Raum $([0, 1], \{0, 1, \dots, n\}, \mathcal{P}^T)$ mit den Wahrscheinlichkeits- bzw. Likelihoodfunktionen

$$f(x; \theta) = L(\theta; x) = \binom{n}{x} \theta^x (1 - \theta)^{n-x}$$

mit $0 \leq \theta \leq 1$ und $x = 0, 1, \dots, n$.

Fall $\mathbf{x} = \mathbf{0}$: $L(\theta; 0) = (1 - \theta)^n$ hat auf $[0, 1]$ die Maximalstelle $\hat{\theta}_0 = 0$.

Fall $\mathbf{x} = \mathbf{n}$: $L(\theta; n) = \theta^n$ wird maximal bei $\hat{\theta}_n = 1$.

Fall $\mathbf{0} < \mathbf{x} < \mathbf{n}$: Hier betrachten wir den Logarithmus der Likelihoodfunktion:

$$L^*(\theta; x) = \ln \binom{n}{x} + x \ln(\theta) + (n - x) \ln(1 - \theta)$$

mit der Ableitung

$$\frac{d}{d\theta} L^*(\theta; x) = \frac{x}{1 - \theta} - \frac{n - x}{1 - \theta} = \frac{x(1 - \theta) - (n - x)\theta}{\theta(1 - \theta)} = \frac{x - n\theta}{\theta(1 - \theta)}$$

Die Ableitung wird Null an der Stelle

$$\hat{\theta}_x = \frac{x}{n}$$

Sie ist links von dieser Stelle positiv und rechts davon negativ. Es handelt sich also um die Maximalstelle.

3.2. Gütekriterien

Sei $(\Theta, \mathcal{X}, \mathcal{P})$ ein statistischer Raum und $\tau : \Theta \rightarrow \mathbb{R}$ eine Parameterfunktion. Wir befassen uns in diesem Abschnitt mit der Frage, welche Eigenschaften eine Statistik $D : \mathcal{X} \rightarrow \mathbb{R}$ besitzen sollte, damit man sie als *gute Schätzfunktion* für den Wert $\tau(\theta)$ bezeichnen kann.

3.2.1. Die Risikozerlegung

Als Abstand zwischen dem Schätzwert $D(x)$ und dem zu schätzenden Wert $\tau(\theta)$ wählen wir die quadratische Differenz $(D(x) - \tau(\theta))^2$. Um eine Gütemaß für die Funktion D zu erhalten, bildet man ein gewichtetes Mittel dieser Abstände über alle Ergebnisse x mit den Wahrscheinlichkeiten ihres Auftretens bei Zustand θ . Die Größe

$$R(D, \theta) = \int (D(x) - \tau(\theta))^2 P_\theta(dx) = \mathcal{E}_\theta(D - \tau(\theta))^2$$

heißt das **Risiko** der Schätzfunktion D bei Zustand θ (für die Parameterfunktion τ).

Es gibt im allgemeinen allerdings keine Schätzfunktion D^* , die gleichmäßig für alle Zustände θ ein geringeres Risiko als jede andere Schätzfunktion besäße. Man muss sich bei der Suche nach einer besten Schätzfunktion auf eine Klasse von „vernünftigen“ Funktionen beschränken. Einen Ansatz für die Wahl einer solchen Klasse liefert die folgende Umformung des Risikos:

$$\begin{aligned} R(D, \theta) &= \mathcal{E}_\theta(D - \tau(\theta))^2 = \mathcal{E}_\theta [(D - \mathcal{E}_\theta D) + (\mathcal{E}_\theta D - \tau(\theta))]^2 \\ &= \mathcal{E}_\theta [(D - \mathcal{E}_\theta D)^2 + 2(\mathcal{E}_\theta D - \tau(\theta))(D - \mathcal{E}_\theta D) + (\mathcal{E}_\theta D - \tau(\theta))^2] \\ &= \mathcal{E}_\theta (D - \mathcal{E}_\theta D)^2 + 2(\mathcal{E}_\theta D - \tau(\theta))\mathcal{E}_\theta(D - \mathcal{E}_\theta D) + (\mathcal{E}_\theta D - \tau(\theta))^2 \end{aligned}$$

Wegen $\mathcal{E}_\theta(D - \mathcal{E}_\theta D)^2 = \text{var}_\theta(D)$ und $\mathcal{E}_\theta(D - \mathcal{E}_\theta D) = \mathcal{E}_\theta D - \mathcal{E}_\theta D = 0$ erhält man daraus die sogenannte **Risikozerlegung**

$$R(D, \theta) = (\mathcal{E}_\theta D - \tau(\theta))^2 + \text{var}_\theta(D) \quad (3.13)$$

in einen *systematischen* Anteil $(\mathcal{E}_\theta D - \tau(\theta))^2$ und einen rein *zufallsbedingten* Anteil $\text{var}_\theta(D)$.

Demgemäß betrachten wir im folgenden nur solche Schätzfunktionen, deren systematischer Anteil verschwindet und suchen in dieser Klasse nach Funktionen mit minimalem zufallsbedingten Anteil.

3.2.2. Erwartungstreue

Definition 3.1 Eine Statistik $D : \mathcal{X} \rightarrow \mathbb{R}$ heißt eine **erwartungstreue Schätzfunktion** für eine Parameterfunktion $\tau(\theta)$, wenn für **alle** $\theta \in \Theta$ gilt

$$\mathcal{E}_\theta D = \int D(x) P_\theta(dx) = \tau(\theta) \quad (3.14)$$

$\mathcal{E}_\theta D$ ist der Erwartungswert der Statistik D bezüglich der Wahrscheinlichkeit \mathcal{P}_θ . Falls diese die Dichte $f(x; \theta)$ besitzt, ist

$$\mathcal{E}_\theta D = \int D(x) f(x; \theta) dx \quad (3.15)$$

Die Normalverteilung

Die n-fache Wiederholung eines Experiments mit $\mathcal{N}(\mu, \sigma^2)$ -Verteilung kann man wie in Abschnitt 2.3.3 beschrieben mittels der Koordinatenvariablen X_k charakterisieren: Für jeden Parametervektor $\theta = (\mu, \sigma^2)$ sind die Statistiken $X_k(x_1, x_2, \dots, x_n) = x_k$ auf dem Wahrscheinlichkeitsraum $(\mathbb{R}^n, \mathcal{B}_n, P_{(\mu, \sigma^2)})$ stochastisch unabhängig und $\mathcal{N}(\mu, \sigma^2)$ -verteilt. Insbesondere gilt also $\mathcal{E}_{(\mu, \sigma^2)} X_k = \mu$ und $\text{var}_{(\mu, \sigma^2)}(X_k) = \sigma^2$.

Schätzfunktion für den Mittelwert μ . Die Maximum-Likelihood-Schätzfunktion (3.9) für den Mittelwert $\tau_1(\mu, \sigma^2) = \mu$ lässt sich mit Hilfe der Koordinatenvariablen darstellen:

$$\begin{aligned} T_1(x) &= \bar{x} = \frac{1}{n} (x_1 + x_2 + \dots + x_n) \\ &= \frac{1}{n} (X_1(x) + X_2(x) + \dots + X_n(x)) \end{aligned}$$

Die Statistik T_1 wird üblicherweise mit \bar{X} bezeichnet und heißt das **Stichprobenmittel** (sample mean) :

$$\bar{X} = \frac{1}{n} (X_1 + X_2 + \dots + X_n) \quad (3.16)$$

mit $\overline{X}(x) = \bar{x}$. Wegen der Linearität des Erwartungswertoperators gilt

$$\begin{aligned}\mathcal{E}_{(\mu, \sigma^2)} \overline{X} &= \frac{1}{n} (\mathcal{E}_{(\mu, \sigma^2)} X_1 + \mathcal{E}_{(\mu, \sigma^2)} X_2 + \dots + \mathcal{E}_{(\mu, \sigma^2)} X_n) \\ &= \frac{1}{n} \underbrace{(\mu + \mu + \dots + \mu)}_{n\text{-mal}} \\ &= \mu = \tau_1(\mu, \sigma^2)\end{aligned}$$

woraus folgt, dass das Stichprobenmittel \overline{X} eine erwartungstreue Schätzfunktion für die Parameterfunktion τ_1 , d.h. den Mittelwert ist.

Schätzfunktion für die Varianz σ^2 .

Wegen der Umrechnungsformel (3.11) besitzt die Maximum-Likelihood-Schätzfunktion (3.10) die beiden Darstellungen

$$T_2 = \frac{1}{n} \sum_{k=1}^n (X_k - \overline{X})^2$$

und

$$T_2 = \frac{1}{n} \sum_{k=1}^n X_k^2 - (\overline{X})^2 \quad (3.17)$$

Ferner ist

$$\mathcal{E}_{(\mu, \sigma^2)}(X_k^2) = \text{var}_{(\mu, \sigma^2)}(X_k) + (\mathcal{E}_{(\mu, \sigma^2)} X_k)^2 = \sigma^2 + \mu^2$$

und wegen der stochastischen Unabhängigkeit der Koordinatenvariablen

$$\begin{aligned}\text{var}_{(\mu, \sigma^2)}(\overline{X}) &= \text{var}_{(\mu, \sigma^2)}\left(\frac{1}{n} (X_1 + \dots + X_n)\right) \\ &= \left(\frac{1}{n}\right)^2 (\text{var}_{(\mu, \sigma^2)}(X_1) + \dots + \text{var}_{(\mu, \sigma^2)}(X_n)) \\ &= \left(\frac{1}{n}\right)^2 (\sigma^2 + \dots + \sigma^2) = \frac{1}{n} \sigma^2\end{aligned}$$

so dass

$$\mathcal{E}_{(\mu, \sigma^2)}(\overline{X}^2) = \text{var}_{(\mu, \sigma^2)}(\overline{X}) + (\mathcal{E}_{(\mu, \sigma^2)} \overline{X})^2 = \frac{1}{n} \sigma^2 + \mu^2$$

Für den Erwartungswert der Statistik T_2 ergibt sich daraus

$$\mathcal{E}_{(\mu, \sigma^2)} T_2 = \frac{1}{n} \sum_{k=1}^n (\sigma^2 + \mu^2) - \left(\frac{1}{n} \sigma^2 + \mu^2\right) = \sigma^2 - \frac{1}{n} \sigma^2 = \frac{n-1}{n} \sigma^2$$

Die Statistik T_2 ist demnach keine erwartungstreue Schätzfunktion für $\tau_2(\mu, \sigma^2) = \sigma^2$. Die Erwartungstreue lässt sich aber hier ganz einfach dadurch erreichen, dass man T_2 mit dem Faktor $\frac{n}{n-1}$ multipliziert.

Die so transformierte Statistik

$$S^2 = \frac{n}{n-1} T_2 = \frac{1}{n-1} \sum_{k=1}^n (X_k - \bar{X})^2 \quad (3.18)$$

heißt die **Stichprobenvarianz** (sample variance). Der Wert der Stichprobenvarianz für ein Ergebnis x wird üblicherweise mit $S^2(x) = s^2$ bezeichnet. Wie gerade berechnet, ist die Stichprobenvarianz eine erwartungstreue Schätzfunktion für die Parameterfunktion $\tau_2(\mu, \sigma^2) = \sigma^2$, also die Varianz der Normalverteilung.

3.2.3. Die Wirksamkeit

Gibt es zu einer Parameterfunktion $\tau(\theta)$ mehrere erwartungstreue Schätzfunktionen, so wird man sich für diejenigen entscheiden, deren mittlere Abweichung vom zu schätzenden Wert am geringsten ist. Da man den Zustand θ aber nicht kennt, muss diese Eigenschaft für alle Zustände erfüllt sein. Daraus ergibt sich die

Definition 3.2 Ist \mathcal{D}_τ eine Klasse von erwartungstreuen Schätzfunktionen für eine Parameterfunktion τ , so heißt ein $T_0 \in \mathcal{D}_\tau$ **wirksamste** Schätzfunktion in \mathcal{D}_τ , wenn für **alle** $T \in \mathcal{D}_\tau$ und **alle** $\theta \in \Theta$ gilt

$$\text{var}_\theta(T_0) \leq \text{var}_\theta(T)$$

Satz 3.2 Ist $D : \mathcal{X} \rightarrow \mathbb{R}$ eine für $\tau(\theta)$ erwartungstreue Schätzfunktion und $T : \mathcal{X} \rightarrow \mathcal{T}$ eine suffiziente Statistik, so gibt es eine Statistik $D_0 : \mathcal{T} \rightarrow \mathbb{R}$ so, dass für $\tilde{D}(x) := D_0(T(x))$ und alle $\theta \in \Theta$ gilt:

$$\begin{aligned} \mathcal{E}_\theta \tilde{D} &= \mathcal{E}_\theta D = \tau(\theta) \\ \text{var}_\theta(\tilde{D}) &\leq \text{var}_\theta(D) \end{aligned}$$

Bei der Suche nach wirksamsten Schätzfunktionen kann man sich daher auf Funktionen der Form $\tilde{D}(x) = D_0(T(x))$ beschränken, wobei die Statistik $D_0(t)$ erwartungstreu für $\tau(\theta)$ bezüglich der Verteilungsannahme der suffizienten Statistik T ist:

$$\mathcal{E}_\theta^T D_0 = \int D_0(t) P_\theta^T(dt) = \tau(\theta) \text{ für alle } \theta \quad (3.19)$$

Bei den für die Anwendungen wichtigsten Verteilungsannahmen wie Normal-, Exponential-, Poisson- und Binomialverteilung sind außerdem die Verteilungsannahmen der minimalen suffizienten Statistiken **vollständig**, d.h. zu einer Parameterfunktion $\tau(\theta)$ gibt es bis auf Abweichungen auf Nullmengen **höchstens eine** Schätzfunktion D_0 mit der Eigenschaft (3.19), so dass die Schätzfunktionen $\tilde{D}(x) = D_0(T(x))$ wirksamste Schätzfunktionen sind.

Insbesondere sind das Stichprobenmittel bzw. die Stichprobenvarianz wirksamste Schätzfunktion ist der Klasse aller erwartungstreuen Schätzfunktionen für den Mittelwert bzw. die Varianz der Normalverteilung.

4. Konfidenzbereiche

Die Aussage, dass nach der Durchführung eines Experiments mit Normalverteilungsannahme das arithmetische Mittel

$$\bar{x} = \frac{1}{n}(x_1 + x_2 + \dots + x_n) = \bar{X}(x)$$

der Messwerte ein *guter Schätzwert für den Mittelwert μ* ist, weil er von einer wirksamsten erwartungstreuen Schätzfunktion stammt, ist nicht sehr befriedigend. Man möchte wissen, wie weit der Schätzwert vom zu schätzenden Wert entfernt liegt, also eine Fehlerabschätzung. Da wir es aber mit Zufallsexperimenten zu tun haben, ist keine exakte Fehlerabschätzung möglich. Man kann höchstens Bereiche angeben, in denen der gesuchte Parameter mit *sehr großer Wahrscheinlichkeit* liegt. Derartige Bereiche heißen in der Statistik Konfidenzbereiche. Bei eindimensionalen Parametern bzw. Parameterfunktionen ist der gesuchte Bereich im allgemeinen ein Intervall, so dass man von **Konfidenzintervallen** spricht.

Definition 4.1 Ist $(\Theta, \mathcal{X}, \mathcal{P})$ ein statistischer Raum und $\tau : \Theta \rightarrow \mathbb{R}$ eine Parameterfunktion, so heißt ein Paar $T_1(x), T_2(x)$ von Statistiken $T_i : \mathcal{X} \rightarrow \mathbb{R}$ mit der Eigenschaft

$$P_\theta(T_1 \leq \tau(\theta) \leq T_2) \geq \gamma \quad \text{für alle } \theta \in \Theta \quad (4.20)$$

ein **Konfidenzintervall für τ zur Konfidenzzahl γ** .

Die Konfidenzzahl ist ein Wahrscheinlichkeitswert, der die Formulierungen „höchstwahrscheinlich“ oder „so gut wie sicher“ zahlenmäßig präzisiert. In der statistischen Praxis sind die Werte 0.95 oder 0.99 üblich.

Hat man das Experiment durchgeführt und die Stichprobe $x \in \mathcal{X}$ erhalten, so ist gemäß dieser Definition die Aussage, dass der unbekannte Wert $\tau(\theta)$ im Intervall $[T_1(x), T_2(x)]$ liegt, mit einer Wahrscheinlichkeit von mindestens γ richtig.

Wir entwickeln zunächst schrittweise ein Konfidenzintervall für den Mittelwert bei der n -fachen Wiederholung eines Experiments mit $\mathcal{N}(\mu, \sigma^2)$ -Verteilung. Aus dieser Herleitung wird sich ein Rezept zur Konstruktion von Konfidenzintervallen bei anderen Verteilungsannahmen herauskristallisieren, das am Ende dieses Abschnitts vorgestellt wird.

4.1. Ein Konfidenzintervall für den Mittelwert

Der Ansatzpunkt für die Konstruktion eines Konfidenzintervalls für den Mittelwert ist das Stichprobenmittel

$$\bar{X} = \frac{1}{n}(X_1 + \dots + X_n)$$

als wirksamste erwartungstreue Schätzfunktion.

Auf dem Wahrscheinlichkeitsraum $(\mathbb{R}^n, \mathcal{B}_n, P_{(\mu, \sigma^2)})$ sind die Koordinatenvariablen X_k stochastisch unabhängig und $\mathcal{N}(\mu, \sigma^2)$ -verteilt und \bar{X} als Linearkombination dieser Variablen ebenfalls normalverteilt.

Aus dem vorhergehenden Abschnitt über Gütekriterien wissen wir, dass $\mathcal{E}_{(\mu, \sigma^2)} \bar{X} = \mu$ und $\text{var}_{(\mu, \sigma^2)}(\bar{X}) = \frac{1}{n} \sigma^2$.

\bar{X} ist damit auf $(\mathbb{R}^n, \mathcal{B}_n, P_{(\mu, \sigma^2)})$ $\mathcal{N}(\mu, \frac{\sigma^2}{n})$ -verteilt und die normierte Zufallsvariable

$$Y = \sqrt{\frac{n}{\sigma^2}}(\bar{X} - \mu)$$

$\mathcal{N}(0, 1)$ -verteilt. Es gilt also

$$P_{(\mu, \sigma^2)}(c_1 \leq Y \leq c_2) = P_{(\mu, \sigma^2)}(Y \in [c_1, c_2]) = \int_{c_1}^{c_2} \varphi(t) dt = \Phi(c_2) - \Phi(c_1)$$

mit der Dichte $\varphi(t)$ bzw der Verteilungsfunktion $\Phi(t)$ der standardisierten Normalverteilung.

Die beiden Zahlen c_1 und c_2 können so bestimmt werden, dass

$$\int_{c_1}^{c_2} \varphi(t) dt = \gamma$$

Wegen der Symmetrie der Dichte bietet sich wie in Abbildung 2 dargestellt an, $c_2 = c$ und $c_1 = -c$ mit

$$\int_{-\infty}^c \varphi(t) dt = \Phi(c) = \gamma + \frac{1-\gamma}{2} = \frac{1+\gamma}{2}$$

zu wählen.

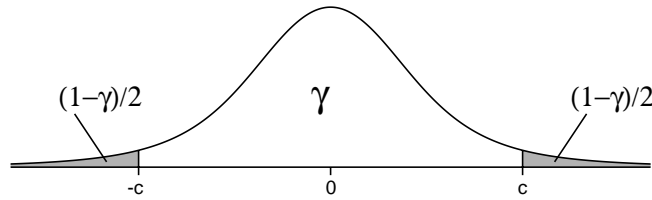


Abbildung 2: Wahl von c_1, c_2

Die Umformung der Ungleichungen

$$c_1 \leq Y(x) = \sqrt{\frac{n}{\sigma^2}}(\bar{x} - \mu) \Leftrightarrow \mu \leq \bar{x} - c_1 \sqrt{\frac{\sigma^2}{n}}$$

$$Y(x) = \sqrt{\frac{n}{\sigma^2}}(\bar{x} - \mu) \leq c_2 \Leftrightarrow \bar{x} - c_2 \sqrt{\frac{\sigma^2}{n}} \leq \mu$$

ergibt

$$(c_1 \leq Y \leq c_2) = (D_1(\cdot, \sigma^2) \leq \mu \leq D_2(\cdot, \sigma^2))$$

mit den Funktionen

$$D_1(x, \sigma^2) = \bar{x} - c_2 \sqrt{\frac{\sigma^2}{n}}$$

und

$$D_2(x, \sigma^2) = \bar{x} - c_1 \sqrt{\frac{\sigma^2}{n}}$$

und damit

$$P_{(\mu, \sigma^2)}(D_1(\cdot, \sigma^2) \leq \mu \leq D_2(\cdot, \sigma^2)) = \gamma$$

Falls die Varianz σ^2 **bekannt** wäre, hätte man damit ein Konfidenzintervall. Da die Varianz aber im allgemeinen nicht bekannt ist, kann man die Funktionswerte $D_1(x, \sigma^2)$, $D_2(x, \sigma^2)$ nicht berechnen. D_1 und D_2 sind **keine Statistiken!**

Als Ausweg bietet sich an, den unbekanntem Parameter σ^2 durch seinen Schätzwert $S^2(x) = s^2$ zu ersetzen, d.h. mit den Statistiken

$$T_1(x) = D_1(x, s^2) = \bar{x} - c_2 \sqrt{\frac{s^2}{n}}$$

und

$$T_2(x) = D_2(x, s^2) = \bar{x} - c_1 \sqrt{\frac{s^2}{n}}$$

zu arbeiten bzw. — wie man durch Zurückrechnen der Ungleichungen sieht — die Zufallsvariable Y durch $\tilde{Y} = \sqrt{\frac{n}{S^2}}(\bar{X} - \mu)$ zu ersetzen.

Diese Zufallsvariable ist aber vermutlich nicht mehr normalverteilt. Zur Berechnung der Wahrscheinlichkeit $P_{(\mu, \sigma^2)}(c_1 \leq \tilde{Y} \leq c_2)$ muss die Verteilung dieser Zufallsvariablen bestimmt werden.

4.2. Die Studentsche t-Verteilung

Bezüglich der Wahrscheinlichkeit $P_{(\mu, \sigma^2)}$ sind die Zufallsvariablen $G_k = \frac{1}{\sqrt{\sigma^2}}(X_k - \mu)$ stochastisch unabhängig und $\mathcal{N}(0, 1)$ -verteilt. Wendet man auf den Zufallsvektor mit diesen G_k als Komponenten eine Orthogonalmatrix an, so ist nach einem Satz über die Normalverteilung (s. [6]) das Resultat wieder ein gaussischer Einheitsvektor.

Wir wählen als Orthogonalmatrix eine, deren erste Zeile aus lauter gleichen Komponenten besteht:

$$\begin{pmatrix} H_1 \\ H_2 \\ \vdots \\ H_n \end{pmatrix} = \begin{pmatrix} \frac{1}{\sqrt{n}} & \frac{1}{\sqrt{n}} & \cdots & \frac{1}{\sqrt{n}} \\ u_{21} & u_{22} & \cdots & u_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ u_{n1} & u_{n2} & \cdots & u_{nn} \end{pmatrix} \begin{pmatrix} G_1 \\ G_2 \\ \vdots \\ G_n \end{pmatrix}$$

Die H_k sind $\mathcal{N}(0, 1)$ -verteilt. Speziell ist

$$\begin{aligned}
 H_1 &= \frac{1}{\sqrt{n}} \sum_{k=1}^n G_k = \frac{1}{\sqrt{n}} \frac{1}{\sqrt{\sigma^2}} \sum_{k=1}^n (X_k - \mu) \\
 &= \frac{1}{\sqrt{n}} \frac{1}{\sqrt{\sigma^2}} \left(\sum_{k=1}^n X_k - n\mu \right) \\
 &= \frac{1}{\sqrt{\sigma^2}} \frac{n}{\sqrt{n}} (\bar{X} - \mu) \\
 &= \sqrt{\frac{n}{\sigma^2}} (\bar{X} - \mu)
 \end{aligned}$$

$\mathcal{N}(0, 1)$ -verteilt, was wir bereits wissen.

Bei der Transformation von Vektoren mit einer Orthogonalmatrix U bleibt die euklidische Norm erhalten, d.h. für $y = Ux$ ist $\|y\| = \|x\|$ oder

$$y_1^2 + y_2^2 + \dots + y_n^2 = x_1^2 + x_2^2 + \dots + x_n^2$$

Für unsere Zufallsvariablen bedeutet das

$$H_1^2 + H_2^2 + \dots + H_n^2 = G_1^2 + G_2^2 + \dots + G_n^2$$

H_1^2 ist schon bekannt. Deshalb wird dieser Term auf die rechte Seite gesetzt:

$$\begin{aligned}
 &H_2^2 + H_3^2 + \dots + H_n^2 = G_1^2 + \dots + G_n^2 - H_1^2 \\
 &= \sum_{k=1}^n \frac{1}{\sigma^2} (X_k - \mu)^2 - \frac{n}{\sigma^2} (\bar{X} - \mu)^2 \\
 &= \frac{1}{\sigma^2} \left\{ \sum_{k=1}^n (X_k^2 - 2\mu X_k + \mu^2) - (n\bar{X}^2 + 2n\mu\bar{X} - n\mu^2) \right\} \\
 &= \frac{1}{\sigma^2} \left\{ \sum_{k=1}^n X_k^2 - 2\mu \sum_{k=1}^n X_k + n\mu^2 - n\bar{X}^2 + 2n\mu\bar{X} - n\mu^2 \right\}
 \end{aligned}$$

Wegen $\sum_{k=1}^n X_k = n\bar{X}$ vereinfacht sich der Term in geschweiften Klammern. Man erhält

$$H_2^2 + H_3^2 + \dots + H_n^2 = \frac{1}{\sigma^2} \left\{ \sum_{k=1}^n X_k^2 - n\bar{X}^2 \right\}$$

und weiter nach der Umrechnungsformel (3.11) die Beziehung

$$H_2^2 + H_3^2 + \dots + H_n^2 = \frac{1}{\sigma^2} \sum_{k=1}^n (X_k - \bar{X})^2 = \frac{n-1}{\sigma^2} S^2 \quad (4.21)$$

Für die Zufallsvariable \tilde{Y} ergibt sich daraus die Darstellung

$$\sqrt{\frac{n}{S^2}}(\bar{X} - \mu) = \frac{\sqrt{\frac{n}{\sigma^2}}(\bar{X} - \mu)}{\sqrt{\frac{1}{\sigma^2}S^2}} = \frac{H_1}{\sqrt{\frac{1}{n-1}(H_2^2 + H_3^2 + \dots + H_n^2)}} \quad (4.22)$$

Die Verteilung einer derartigen Funktion von stochastisch unabhängigen $\mathcal{N}(0, 1)$ -verteilten Zufallsvariablen kann man mit Hilfe des Transformationsatzes für Dichten bestimmen. Wir beschränken uns hier auf die Bezeichnungen und die Formeln für die Dichten.

4.3. Testverteilungen

In diesem Abschnitt betrachten wir Zufallsvariablen auf einem allgemeinen Wahrscheinlichkeitsraum (Ω, \mathcal{A}, P) .

4.3.1. Die Chi-Quadrat-Verteilung

Sind G_1, G_2, \dots, G_m stochastisch unabhängige und $\mathcal{N}(0, 1)$ -verteilte Zufallsvariable, so heißt die Verteilung der Zufallsvariablen

$$Y = G_1^2 + G_2^2 + \dots + G_m^2 \quad (4.23)$$

die **Chi-Quadrat-Verteilung mit m Freiheitsgraden**, kurz die χ_m^2 -Verteilung. Sie besitzt die Dichte

$$f_{\chi^2, m}(t) = \begin{cases} \frac{1}{2^{m/2}\Gamma(m/2)} t^{m/2-1} e^{-t/2} & \text{für } t > 0 \\ 0 & \text{für } t \leq 0 \end{cases} \quad (4.24)$$

Γ bezeichnet dabei die Gamma-Funktion

$$\Gamma(x) = \int_0^\infty t^{x-1} e^{-t} dt$$

4.3.2. Die t-Verteilung

Sind $G_0, G_1, G_2, \dots, G_m$ stochastisch unabhängige und $\mathcal{N}(0, 1)$ -verteilte Zufallsvariable, so heißt die Verteilung der Zufallsvariablen

$$Z = \frac{G_0}{\sqrt{\frac{1}{m}(G_1^2 + G_2^2 + \dots + G_m^2)}} \quad (4.25)$$

die **t-Verteilung mit m Freiheitsgraden** oder kurz t_m -Verteilung. Sie besitzt die Dichte

$$f_{t, m}(x) = \frac{\Gamma\left(\frac{m+1}{2}\right)}{\Gamma\left(\frac{m}{2}\right)\sqrt{m\pi}} \cdot \frac{1}{\left(1 + \frac{x^2}{m}\right)^{\frac{(m+1)}{2}}} \quad (4.26)$$

4.4. Konfidenzintervalle für die Parameter der Normalverteilung

4.4.1. Konfidenzintervall für den Mittelwert

Gemäß (4.25) ist die Zufallsvariable (4.22) bezüglich der Wahrscheinlichkeit $P_{(\mu, \sigma^2)}$ t-verteilt mit $n-1$ Freiheitsgraden. Mit der Verteilungsfunktion $F_{t, n-1}$ dieser Verteilung gilt also

$$P_{(\mu, \sigma^2)} \left(c_1 \leq \sqrt{\frac{n}{S^2}} (\bar{X} - \mu) \leq c_2 \right) = F_{t, n-1}(c_2) - F_{t, n-1}(c_1)$$

Da die Ungleichungen

$$c_1 \leq \sqrt{\frac{n}{s^2}} (\bar{x} - \mu) \leq c_2$$

genau dann erfüllt sind wenn

$$T_1(x) \leq \mu \leq T_2(x)$$

mit $T_1(x) = \bar{x} - c_2 \sqrt{\frac{s^2}{n}}$ und $T_2(x) = \bar{x} - c_1 \sqrt{\frac{s^2}{n}}$, bildet das Intervall $[T_1(x), T_2(x)]$ ein Konfidenzintervall zur Konfidenzzahl γ für den Mittelwert μ , wenn

$$F_{t, n-1}(c_2) - F_{t, n-1}(c_1) = \gamma$$

Dies kann man dadurch erreichen, dass man c_1 und c_2 so wählt, dass

$$F_{t, n-1}(c_1) = \frac{1 - \gamma}{2} \quad \text{und} \quad F_{t, n-1}(c_2) = \frac{1 + \gamma}{2}$$

Da die Dichte(4.26) der t-Verteilung symmetrisch zum Nullpunkt ist, ergibt das — wie man sich wieder an der Abbildung 2 veranschaulicht — $c_2 = c$ und $c_1 = -c$ mit $F_{t, n-1}(c) = (1 + \gamma)/2$.

4.4.2. Konfidenzintervall für die Varianz

Aus den obigen Berechnungen ergibt sich gleichzeitig eine Möglichkeit, ein Konfidenzintervall für die Varianz σ^2 der Normalverteilung zu konstruieren. Aus (4.23) und (4.21) folgt nämlich, dass die Zufallsvariable $\frac{n-1}{\sigma^2} S^2$ bezüglich der Wahrscheinlichkeit $P_{(\mu, \sigma^2)}$ Chi-Quadrat-verteilt mit $n-1$ Freiheitsgraden ist:

$$P_{(\mu, \sigma^2)} \left(c_1 \leq \frac{n-1}{\sigma^2} S^2 \leq c_2 \right) = F_{\chi^2, n-1}(c_2) - F_{\chi^2, n-1}(c_1)$$

$F_{\chi^2, n-1}$ bezeichnet dabei die Verteilungsfunktion der χ_{n-1}^2 -Verteilung.

Da

$$c_1 \leq \frac{n-1}{\sigma^2} S^2 \leq c_2$$

genau dann gilt, wenn

$$\frac{(n-1)s^2}{c_2} \leq \sigma^2 \leq \frac{(n-1)s^2}{c_1}$$

ist dies ein Konfidenzintervall zur Konfidenzzahl γ , wenn man c_1 und c_2 so wählt, dass

$$F_{\chi^2, n-1}(c_1) = \frac{1 - \gamma}{2} \quad \text{und} \quad F_{\chi^2, n-1}(c_2) = \frac{1 + \gamma}{2}$$

4.5. Konfidenzbereiche

Wir verallgemeinern die Problemstellung etwas und betrachten nicht nur Intervalle $[T_1(x), T_2(x)]$ sondern Abbildungen K , die jedem Ergebnis x eine Teilmenge $K(x) \subset \mathbb{R}$ zuordnen. Mit den Ansätzen $K(x) = [T_1(x), \infty)$ bzw. $K(x) = (-\infty, T_2(x)]$ erfasst man z.B. die Aufgabe, Konfidenz-Unterschranken bzw. Konfidenz-Oberschranken für eine Parameterfunktion zu bestimmen.

Für eine mengenwertige Abbildung $x \mapsto K(x)$ benutzen wir wieder die Ereignisschreibweise

$$(t \in K) := \{x \in \mathcal{X}; t \in K(x)\}$$

Sei wieder $(\Theta, \mathcal{X}, \mathcal{P})$ ein statistischer Raum, $\tau : \Theta \rightarrow \mathbb{R}$ eine Parameterfunktion und γ eine reelle Zahl mit $0 < \gamma < 1$.

Definition 4.2 Eine Abbildung $x \mapsto K(x)$ die jedem $x \in \mathcal{X}$ eine Teilmenge $K(x) \subset \mathbb{R}$ zuordnet und die Eigenschaft

$$\mathcal{P}_\theta(\tau(\theta) \in K) \geq \gamma \quad \text{für alle } \theta \in \Theta$$

besitzt, heißt ein Konfidenzbereich für τ zur Konfidenzzahl γ .

4.5.1. Die statistische Methode

Ist der Parameter θ eindimensional und ist $T : \mathcal{X} \rightarrow \mathbb{R}$ eine Statistik, deren Verteilungsfunktion

$$F^T(t; \theta) = P_\theta^\top(-\infty, t] = P_\theta(T \leq t)$$

formelmäßig bekannt ist, so erhält man einen Konfidenzbereich für $\tau(\theta) = \theta$ auf die folgende Weise:

Zu $\gamma_1 = \frac{1}{2}(1 - \gamma)$ und $\gamma_2 = \frac{1}{2}(1 + \gamma)$ und alle θ bestimmt man Zahlen $c_1(\theta), c_2(\theta)$ mit

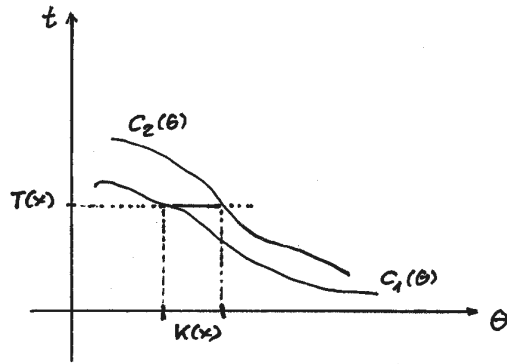
$$F^T(c_1(\theta) - 0; \theta) \leq \gamma_1 \quad \text{und} \quad F^T(c_2(\theta); \theta) \geq \gamma_2$$

dann ist

$$K(x) = \{\theta; c_1(\theta) \leq T(x) \leq c_2(\theta)\}$$

ein Konfidenzbereich für θ zur Konfidenzzahl γ .

Die Konstruktion kann man sich an der Skizze



veranschaulichen, weshalb diese Methode auch die graphische Methode heißt.
 Beweis: Es ist $\theta \in K(x)$ genau dann, wenn $c_1(\theta) \leq T(x) \leq c_2(\theta)$. Daher ist

$$\begin{aligned}
 P_\theta(\theta \in K) &= P_\theta\{x; \theta \in K(x)\} \\
 &= P_\theta\{x; c_1(\theta) \leq T(x) \leq c_2(\theta)\} \\
 &= P_\theta(c_1(\theta) \leq T \leq c_2(\theta)) \\
 &= P_\theta^\top[c_1(\theta), c_2(\theta)] \\
 &= F^\top(c_2(\theta); \theta) - F^\top(c_1(\theta) - 0; 0) \\
 &\geq \gamma_2 - \gamma_1 = \gamma
 \end{aligned}$$

Beispiel: Ein Konfidenzintervall für das Parameter $\theta > 0$ bei der n -fachen Wiederholung eines $\mathcal{U}[0, \theta]$ -Experiments.

Die Koordinatenvariablen X_k besitzen bezüglich P_θ die Verteilungsfunktion

$$F_0(t; \theta) = \begin{cases} 0 & \text{für } t < 0 \\ \frac{t}{\theta} & \text{für } 0 \leq t \leq \theta \\ 1 & \text{für } t > \theta \end{cases}$$

Die Verteilungsfunktion der Statistik $T = \max(X_1, X_2, \dots, X_n)$ ist daher

$$F^T(t; \theta) = \underbrace{F_0(t; \theta) \cdot F_0(t; \theta) \cdots F_0(t; \theta)}_{n\text{-mal}} = \begin{cases} 0 & \text{für } t < 0 \\ \left(\frac{t}{\theta}\right)^n & \text{für } 0 \leq t \leq \theta \\ 1 & \text{für } t > \theta \end{cases}$$

Die Lösungen der Gleichungen

$$F^\top(t; \theta) = \left(\frac{t}{\theta}\right)^n = \gamma_i$$

sind $c_1(\theta) = \theta \sqrt[n]{\gamma_1}$ und $c_2(\theta) = \theta \sqrt[n]{\gamma_2}$ und aus den Ungleichungen $c_1(\theta) \leq T(x) \leq c_2(\theta)$ mit $T(x) = \max(x_1, x_2, \dots, x_n)$ ergibt sich das Konfidenzintervall

$$\frac{\max(x_1 \dots x_n)}{\sqrt[n]{\gamma_2}} \leq \theta \leq \frac{\max(x_1 \dots x_n)}{\sqrt[n]{\gamma_1}}$$

4.5.2. Die Methode der Pivotvariablen

Eine weitere Methode zur Konstruktion von Konfidenzbereichen findet man, wenn man die Herleitung der Konfidenzintervalle für Mittelwert und Varianz der Normalverteilung auf ein allgemeines Prinzip hin untersucht. Dabei ergibt sich, dass als wesentlicher Schritt benutzt wurde, dass die beiden Zufallsvariablen

$$\sqrt{\frac{n}{s^2}}(\bar{X} - \mu) \quad \text{und} \quad \frac{n-1}{\sigma^2} S^2$$

bezüglich der Wahrscheinlichkeit $P_{(\mu, \sigma^2)}$ eine Verteilung besitzen, die **nicht** von den unbekanntenen Größen μ und σ^2 abhängt.

Für derartige Zufallsvariablen vergeben wir einen speziellen Namen:

Definition 4.3 Eine Funktion $Q : \mathcal{X} \times \mathbb{R} \rightarrow \mathbb{R}$ heißt eine **Pivotvariable** für die Parameterfunktion $\tau(\theta)$, wenn

$$P_\theta(Q(\cdot, \tau(\theta)) \in B) = P^Q(B) \quad (4.27)$$

für **alle** Parameterwerte θ mit einer **bekannt**en und **nicht vom Parameter θ abhängigen** Wahrscheinlichkeitsverteilung P^Q gilt.

Dabei ist

$$(Q(\cdot, t) \in B) = \{x \in \mathcal{X} ; Q(x, t) \in B\} \quad (4.28)$$

und P^Q heißt die Verteilung der Pivotvariable Q .

Wie zuvor hergeleitet, sind die Funktionen

$$Q_1(x, t) = \sqrt{\frac{n}{s^2}}(\bar{x} - t) \quad (4.29)$$

und

$$Q_2(x, t) = \frac{n-1}{t} s^2 \quad (4.30)$$

Pivotvariable für die Parameterfunktionen $\tau_1(\mu, \sigma^2) = \mu$ bzw. $\tau_2(\mu, \sigma^2) = \sigma^2$ bei der n -fachen Wiederholung eines Normalverteilungsexperiments mit Parametervektor $\theta = (\mu, \sigma^2)$. P^{Q_1} ist die t_{n-1} -Verteilung und P^{Q_2} die χ_{n-1}^2 -Verteilung.

Hat man eine Pivotvariable zur Verfügung, so erhält man auf ganz einfache Weise einen Konfidenzbereich. Es gilt nämlich

Satz 4.1 Ist Q eine Pivotvariable für $\tau(\theta)$ und B_γ eine Borelsche Menge mit $P^Q(B_\gamma) \geq \gamma$, so ist

$$K(x) = \{t \in \mathbb{R} ; Q(x, t) \in B_\gamma\} \quad (4.31)$$

ein Konfidenzbereich für $\tau(\theta)$ zur Konfidenzzahl γ .

Zum Beweis muss man sich nur die beteiligten Mengen explizit aufschreiben. Es ist

$$\begin{aligned}(\tau(\theta) \in K) &= \{x ; \tau(\theta) \in K(x)\} \\ &= \{x ; Q(x, \tau(\theta)) \in B_\gamma\} \\ &= (Q(\cdot, \tau(\theta)) \in B_\gamma)\end{aligned}$$

und daher

$$P_\theta(\tau(\theta) \in K) = P_\theta(Q(\cdot, \tau(\theta)) \in B_\gamma) = P^Q(B_\gamma) \geq \gamma$$

Wahl der Menge B_γ

Üblicherweise wählt man B_γ als Intervall: $B_\gamma = [c_1, c_2]$. Dann ist

$$P^\theta(B_\gamma) = F^Q(c_2) - F^Q(c_1)$$

mit der Verteilungsfunktion $F^Q(t)$ der Verteilung P^Q . c_1 und c_2 wählt man dann im Allgemeinen so, dass

$$F^Q(c_1) = \frac{1-\gamma}{2} \quad \text{und} \quad F^Q(c_2) = \frac{1+\gamma}{2}$$

Die Umrechnung ergibt dann normalerweise ein Konfidenzintervall $K(x) = [T_1(x), T_2(x)]$. Ist man aber beispielsweise nur an Ober- bzw. Unterschranken für den Wert $\tau(\theta)$ interessiert, so liefert ein Ansatz $B_\gamma = (-\infty, c]$ oder $B_\gamma = [c, \infty)$ je nach Form der Pivotvariable Konfidenzbereiche der Form $K(x) = [T(x), \infty)$ oder $K(x) = (-\infty, T(x)]$.

4.5.3. Pivotvariable für die Exponentialverteilung

Die (eindimensionale) Exponentialverteilung ist charakterisiert durch einen Parameter $\theta \in (0, \infty)$ und besitzt die Dichte $f_0(t; \theta)$ und Verteilungsfunktion $F_0(t; \theta)$ definiert durch

$$f_0(t; \theta) = \begin{cases} 0 & \text{für } t \leq 0 \\ \theta e^{-\theta t} & \text{für } t > 0 \end{cases} \quad \text{und} \quad F_0(t; \theta) = \begin{cases} 0 & \text{für } t \leq 0 \\ 1 - e^{-\theta t} & \text{für } t > 0 \end{cases}$$

Die n -fache Wiederholung eines solchen Experiments besitzt den Stichprobenraum \mathbb{R}^n mit Ergebnissen $x = (x_1, x_2, \dots, x_n)$ und die Wahrscheinlichkeit P_θ ist dadurch charakteristisch, dass die Koordinatenvariablen X_k stochastisch unabhängig und exponentiell verteilt mit Parameter θ sind, d.h. die Verteilungsfunktion $F_0(t, \theta)$ besitzen. Daher gilt

$$P_\theta(2\theta X_k \leq t) = P_\theta(X_k \leq \frac{t}{2\theta}) = F_0(\frac{t}{2\theta}; \theta) = \begin{cases} 0 & t \leq 0 \\ 1 - e^{-\frac{1}{2}t} & t > 0 \end{cases}$$

Die Zufallsvariable $2\theta X_k$ ist also bezüglich der Wahrscheinlichkeit P_θ exponentiell verteilt mit Parameter $\frac{1}{2}$ und damit gemäß Definition eine Pivotvariable.

Um die statistische Information voll auszunutzen, verwenden wir die Zufallsvariable $2\theta(X_1 + X_2 + \dots + X_n)$.

Zur Bestimmung der Verteilung dieser Zufallsvariablen verwenden wir das Resultat aus der Vorlesung Wahrscheinlichkeitsrechnung I (s. [6]), dass eine Summe $G_1^2 + G_2^2$ zweier

stochastisch unabhängiger $\mathcal{N}(0, 1)$ -verteilter Zufallsvariablen ebenfalls exponentiell verteilt mit Parameter $\frac{1}{2}$ ist. Die Summe

$$2\theta(X_1 + \dots + X_n) = 2\theta X_1 + 2\theta X_2 + \dots + 2\theta X_n$$

besitzt also die gleiche Verteilung wie eine Zufallsvariable der Form

$$G_1^2 + G_2^2 + G_3^2 + G_4^2 + \dots + G_{2n-1}^2 + G_{2n}^2$$

mit stochastisch unabhängigen $\mathcal{N}(0, 1)$ -verteilten Variablen G_k , die bekanntlich Chi-Quadrat-verteilt mit $2n$ Freiheitsgraden ist. Daraus erhalten wir den

Satz 4.2 Die Funktion $Q(x, t) = 2t(x_1 + \dots + x_n)$ ist eine χ_{2n}^2 -verteilte Pivotvariable für den Parameter θ der Exponentialverteilung.

Der Mittelwert der Exponentialverteilung ist die Parameterfunktion $\tau(\theta) = \frac{1}{\theta}$.
Wegen

$$2\theta(X_1 + X_2 + \dots + X_n) = \frac{2}{(\frac{1}{\theta})}(X_1 + X_2 + \dots + X_n) = \frac{2}{\tau(\theta)}(X_1 + X_2 + \dots + X_n)$$

folgt daraus

Satz 4.3 Die Funktion

$$Q_1(x, t) = \frac{2}{t}(x_1 + x_2 + \dots + x_n) \tag{4.32}$$

ist eine χ_{2n}^2 -verteilte Pivotvariable für den Mittelwert der Exponentialverteilung.

4.5.4. Berechnung mit R

Die Berechnung von Konfidenzintervallen für den Mittelwert der Normalverteilung ist in der Prozedur `t.test` enthalten, die wir im nächsten Kapitel ausführlich beschreiben. Um ein zweiseitiges Konfidenzintervall für den Mittelwert des Wachsschmelzpunkts aus den Messwerten aus Tabelle 1 (Seite 5) zur Konfidenzzahl $\gamma = 0.99$ zu erhalten, wenn diese in der Datei `bee.txt` abgelegt sind, gibt man

```
> x <- scan("bee.txt")
> ci <- t.test(x, alt="two.sided", conf.level=0.99)
```

ein. Das Konfidenzintervall ist dann dann in der Variable `conf.int` des R-Objekts `ci` enthalten:

```
> ci$conf.int
[1] 63.46842 63.70921
attr(,"conf.level")
[1] 0.99
```

Einseitige Konfidenzintervalle erhält man durch Variation des Arguments `alt`:

```
> cless <- t.test(x,alt="less",conf.level=0.99)
> cless$conf.int
[1] -Inf 63.69696
attr("conf.level")
[1] 0.99
> cgreater <- t.test(x,alt="greater",conf.level=0.99)
> cgreater$conf.int
[1] 63.48067 Inf
attr("conf.level")
[1] 0.99
```

In allen anderen Fällen muss man auf die entsprechenden Formeln zurückgreifen. Zur Berechnung eines Konfidenzintervalls für die Varianz der Normalverteilung etwa zur Konfidenzzahl $\gamma = 0.95$ benötigt man den Wert $z := (n - 1)s^2$ und die Quantile c_1 und c_2 der Chi-Quadrat-Verteilung mit $n - 1$ Freiheitsgraden:

```
> n <- length(x)
> gamma <- 0.95
> z <- (n-1)*var(x)
> c1 <- qchisq((1-gamma)/2,n-1)
> c2 <- qchisq((1+gamma)/2,n-1)
> # Konfidenzintervall fuer die Varianz:
> c(z/c2,z/c1)
[1] 0.0863973 0.1800202
```

Wenn wir annehmen, dass unsere Messwerte exponentiell verteilt mit Parameter θ sind, erhält man ein Konfidenzintervall für den Mittelwert $\tau(\theta)$ mit Hilfe der Pivotvariablen (4.32):

$$c_1 \leq Q_1(x, t) \leq c_2 \quad \Leftrightarrow \quad \frac{2 \sum_i x_i}{c_2} \leq \tau(\theta) \leq \frac{2 \sum_i x_i}{c_1}$$

wobei c_1 und c_2 die $(1 - \gamma)/2$ - und $(1 + \gamma)/2$ -Quantile der χ_{2n}^2 -Verteilung sind:

```
> n <- length(x)
> gamma <- 0.95
> z <- 2 * sum(x)
> c1 <- qchisq((1-gamma)/2,2*n)
> c2 <- qchisq((1+gamma)/2,2*n)
> c(z/c2,z/c1)
[1] 50.03758 83.53251
```

4.6. Asymptotische Pivotvariable

In vielen Fällen, in denen sich keine exakte Pivotvariable finden lässt, kann man auf der Basis des zentralen Grenzwertsatzes eine Näherungslösung finden. Als Beispiel betrachten wir das Problem, ein Konfidenzintervall für den Ausschussanteil einer Warenlieferung zu bestimmen.

4.6.1. Konfidenzintervalle für die Binominalverteilung

Für die Analyse der Entnahme einer Stichprobe vom Umfang n aus einer Warenlieferung kann man, wie die Suffizienzüberlegungen ergeben haben, den statistischen Raum $(\theta, \mathcal{X}_n, \mathcal{P}_n)$ mit $\Theta = [0, 1]$, $\mathcal{X}_n = \{0, 1, 2, \dots, n\}$ und der Verteilungsannahme $\mathcal{P}_n = \{P_\theta^{(n)}; 0 \leq \theta \leq 1\}$ mit den Wahrscheinlichkeitsfunktionen

$$f_n(x; \theta) = \binom{n}{x} \theta^x (1 - \theta)^{n-x}$$

der Binominalverteilung benutzen. θ ist dabei der Ausschussanteil und x die Anzahl der defekten Stücke in der Stichprobe.

Sei $0 < \theta < 1$ ein fest gewählter Parameterwert. Bezüglich der Wahrscheinlichkeit $P_\theta^{(n)}$ ist die Koordinatenvariable $X^{(n)}: \mathcal{X}_n \rightarrow \mathcal{X}_n$, $X^{(n)}(x) := x$, binomialverteilt mit dem Erwartungswert $\mathcal{E}_\theta X^{(n)} = n\theta$ und der Varianz $\text{var}_\theta(X^{(n)}) = n\theta(1 - \theta)$.

Nach dem Grenzwertsatz von Moivre und Laplace gilt für die Zufallsvariablen

$$x \mapsto Q_n(x, \theta) = \frac{X^{(n)}(x) - n\theta}{\sqrt{n\theta(1 - \theta)}} = \frac{x - n\theta}{\sqrt{n\theta(1 - \theta)}}$$

dass

$$\lim_{n \rightarrow \infty} P_\theta^{(n)}(a \leq Q_n(\cdot, \theta) \leq b) = \int_a^b \varphi(t) dt = P^Q[a, b]$$

mit der $\mathcal{N}(0, 1)$ -Verteilung P^Q .

Für „große“ Stichprobenumfänge n gilt also näherungsweise

$$P_\theta^{(n)}(a \leq Q_n(\cdot, \theta) \leq b) \approx P^Q[a, b]$$

d.h. $Q_n(x, \theta)$ ist näherungsweise eine $\mathcal{N}(0, 1)$ -verteilte Pivotvariable für θ .

Die Konvergenzgeschwindigkeit bzw. die Güte der Näherung hängt vom Wert des Parameters θ ab. Je näher θ bei Null oder Eins liegt, desto schlechter ist sie. In der statistischen Praxis geht man von der Faustregel aus, dass man Q_n als $\mathcal{N}(0, 1)$ -verteilt annehmen kann, wenn $n\theta(1 - \theta) \geq 9$.

Unter dieser Annahme ist (approximativer) Konfidenzbereich für θ zur Konfidenzzahl γ gegeben durch

$$K(x) = \{\theta; -c \leq Q_n(x, \theta) \leq c\}$$

wobei c das $\frac{1+\gamma}{2}$ -Quantil der $\mathcal{N}(0, 1)$ -Verteilung, d.h. die Lösung der Gleichung $\Phi(c) = \frac{1+\gamma}{2}$ ist.

Es gilt $-c \leq Q_n(x, \theta) \leq c$ bzw.

$$-c \leq \frac{x - n\theta}{\sqrt{n\theta(1-\theta)}} \leq c$$

genau dann, wenn

$$\left(\frac{x - n\theta}{\sqrt{n\theta(1-\theta)}} \right)^2 = \frac{(x - n\theta)^2}{n\theta(1-\theta)} \leq c^2$$

Multiplikation der beiden Seiten dieser Ungleichung mit $n\theta(1-\theta)$ ergibt

$$(x - n\theta)^2 \leq c^2 n\theta(1-\theta)$$

und anschließendes Ausmultiplizieren der Terme und Ordnen nach Potenzen von θ :

$$(n^2 + c^2 n)\theta^2 - (2nx + c^2 n)\theta + x^2 \leq 0$$

Dividiert man noch durch n , so ist der Konfidenzbereich $K(x)$ die Menge aller Parameterwerte θ , die die Ungleichung

$$h(\theta) := (n + c^2)\theta^2 - (2x + c^2)\theta + \frac{x^2}{n} \leq 0$$

erfüllen. Das quadratische Polynom $h(\theta)$ geht für $\theta \rightarrow \pm \infty$ gegen Unendlich und für $\theta = \frac{x}{n}$ gilt

$$h\left(\frac{x}{n}\right) = c^2 \frac{x}{n} \left(\frac{x}{n} - 1\right) \leq 0$$

Es gibt also zwei (evtl. gleiche) Nullstellen θ_1, θ_2 von h , so dass

$$K(x) = \{\theta ; h(\theta) \leq 0\} = [\theta_1, \theta_2]$$

Nach der Formel für die Nullstellen quadratischer Gleichungen sind θ_1 und θ_2 gegeben durch

$$\theta_{1,2} = \frac{2x + c^2 \pm \sqrt{(2x + c^2)^2 - 4(n + c^2)\frac{x^2}{n}}}{2(n + c^2)}$$

5. Signifikanztests

5.1. Tests

Ein Test ist ein statistisches Verfahren, mit dem auf der Basis einer Stichprobe entschieden wird, ob eine Aussage über den Zustand der Grundgesamtheit zutrifft oder nicht.

Sind beispielsweise die in Tabelle 3 dargestellten Werte die Ozonkonzentrationen in $\mu\text{g}/\text{m}^3$, die innerhalb eines bestimmten Zeitraums an verschiedenen Messstationen ermittelt wurden, so kann man sich die Frage stellen, ob daraus geschlossen werden kann, dass der EU-Grenzwert von $120 \mu\text{g}/\text{m}^3$ für die mittlere Ozonkonzentration in der betreffenden Region überschritten ist oder nicht.

126.1	119.8	121.8	109.3	112.5	123.0	118.3	121.2	118.0	109.2
121.9	124.4	122.2	123.0	123.6	124.2	113.9	123.8	127.9	121.2
122.4	118.8	124.8	110.1	122.5	116.4	120.9	119.4	118.7	118.0
117.0	119.8	113.1	117.8	115.1	119.1	116.3	123.3	120.0	116.1
110.1	118.1	120.6	122.4	116.7	121.8	120.7	113.9	113.6	123.2
120.7	117.1	117.1	117.8	125.2	122.5	120.5	125.8	126.1	120.1
115.2	130.4	117.9	123.7	117.4	120.0	116.1	130.0	113.8	120.0
120.4	131.0	123.2	119.3	120.9	121.4	118.6	123.4	111.2	121.9
133.1	114.5	118.2	117.7	124.4	128.0	126.3	118.7	126.5	121.8

Tabelle 3: Messungen der Ozonkonzentration

Wenn man annimmt, dass auf diese Messwerte die Normalverteilungsannahme mit der mittleren Ozonkonzentration als Mittelwert μ zutrifft, so ist in diesem Beispiel zu entscheiden, ob die Aussage

$$H_0 : „\mu \leq 120“$$

oder ihr Gegenteil

$$H_1 : „\mu > 120“$$

zutrifft.

Eine Aussage über den Parameter θ einer statistischen Grundgesamtheit heißt eine **Hypothese**. Einer solchen Hypothese H ist die Menge

$$\Theta_H = \{\theta \in \Theta ; H \text{ trifft auf } \theta \text{ zu}\}$$

zugeordnet. Wie schon in der Wahrscheinlichkeitsrechnung beim Begriff des Ereignisses, unterscheiden wir im folgenden nicht zwischen einer Aussage und der zugehörigen Menge und legen für mathematische Definitionen fest:

Sei $(\Theta, \mathcal{X}, \mathcal{P})$ ein statistischer Raum. Dann ist eine **Hypothese** eine **Teilmenge H des Parameterraums**. Die entsprechende Aussage lautet dann „Der Zustand θ liegt in H “

und die gegenteilige Aussage, dass θ nicht in H liegt, wird durch die Komplementärmenge $\overline{H} = \{\theta \in \Theta ; \theta \notin H\}$ charakterisiert.

Ein **Test für eine Hypothese** ist dann einfach eine Statistik $D : \mathcal{X} \rightarrow \{H, \overline{H}\}$. Ist x das Ergebnis des statistischen Experiments und ist $D(x) \in H$, so wurde durch den Test entschieden, dass die Hypothese H zutrifft.

5.2. Signifikanztests

Da das Ergebnis eines statistischen Experiments vom Zufall abhängt, sind Fehlentscheidungen

$$D(x) = H \quad \text{obwohl} \quad \theta \in \overline{H}$$

bzw

$$D(x) = \overline{H} \quad \text{obwohl} \quad \theta \in H$$

unvermeidlich. Mit statistischen Mitteln, d.h. bei vorgegebenem statistischem Raum $(\Theta, \mathcal{X}, \mathcal{P})$ kann man lediglich durch geeignete Wahl der Statistik D versuchen, die Wahrscheinlichkeit für das Auftreten solcher Fehlentscheidungen möglichst klein bzw unterhalb einer vorgegebenen Schranke zu halten.

Im allgemeinen ist es aber nicht möglich, die Wahrscheinlichkeiten beider Fehlerarten gleichzeitig unter ein bestimmtes Niveau zu drücken. Da die Menge $(D = \overline{H}) = \{x \in \mathcal{X}; D(x) = \overline{H}\}$ das Komplement der Menge $(D = H) = \{x \in \mathcal{X}; D(x) = H\}$ ist, gilt stets

$$P_\theta(D = \overline{H}) = 1 - P_\theta(D = H)$$

Da außerdem im allgemeinen die Wahrscheinlichkeit P_θ stetig von θ abhängt, haben kleine Werte von $P_\theta(D = \overline{H})$ für $\theta \in H$ zur Folge, dass die Fehlerwahrscheinlichkeit $P_\theta(D = H)$ für solche Parameter $\theta \in \overline{H}$, die in der Nähe der Grenzlinie zwischen H und \overline{H} liegen, sehr groß sind, und umgekehrt.

Man kann also im Prinzip nur eine der beiden Fehlerarten unter Kontrolle halten. Das bedeutet, dass eine der beiden Hypothesen H und \overline{H} in dieser Hinsicht bevorzugt behandelt wird.

In den meisten Anwendungsfällen ist diese Situation von vorneherein gegeben. Beim Beispiel der Ozonwertmessungen wird die entsprechende Verwaltungsstelle einen „falschen Alarm“ vermeiden wollen. D.h. es muss „so gut wie sicher“ sein, dass der Grenzwert überschritten ist, bevor Maßnahmen wie Verkehrsbeschränkungen u.ä. angeordnet werden.

Der entsprechende Test sollte also so konstruiert sein, dass die Entscheidung für die Hypothese „ $\mu > 120$ “ mit einer Fehlerwahrscheinlichkeit getroffen wird, die „so gut wie Null“ ist und im Zweifelsfall angenommen wird, dass die Hypothese „ $\mu \leq 120$ “ zutrifft. Wenn ein solcher Test auf die Aussage „ $\mu > 120$ “ entscheidet, dann ist damit „so gut wie bewiesen“, dass diese Aussage richtig ist.

Ein Test mit einer derartigen ungleichgewichtigen Behandlung der beiden alternativen Hypothesen heißt ein **Signifikanztest**.

Die Hypothese, die man beweisen möchte, heißt die **Alternative** und wird im allgemeinen mit dem Buchstaben A bezeichnet. Ihr Komplement, also die Aussage, auf die der Test entscheidet, wenn die Alternative nicht nachweisbar ist, erhält den Buchstaben H_0 und heißt die **Nullhypothese**.

Seien also $(\Theta, \mathcal{X}, \mathcal{P})$ wieder ein statistischer Raum, H_0 und A Teilmengen von Θ mit $H_0 + A = \Theta$ und α eine reelle Zahl mit $0 < \alpha < 1$.

Definition 5.1 Eine Statistik $D : \mathcal{X} \rightarrow \{H_0, A\}$ mit der Eigenschaft

$$P_\theta(D = A) \leq \alpha \quad \text{für alle } \theta \in H_0 \quad (5.33)$$

heißt ein **Signifikanztest** zum Signifikanzniveau α für die Nullhypothese H_0 (bzw. zum Nachweis der Alternative A).

Bezeichnungen:

- Die Fehlentscheidung $D(x) = A$ bei $\theta \in H_0$ heißt in diesem Zusammenhang der **Fehler 1. Art** und dementsprechend die fälschliche Entscheidung $D(x) = H_0$ bei $\theta \in A$ der **Fehler 2. Art**.
- Die Zahl α heißt die **Signifikanzzahl** oder das Signifikanzniveau des Tests. Sie ist die maximale Wahrscheinlichkeit, mit der ein Fehler 1. Art auftreten darf. Die üblichen Werte für die Signifikanzzahl sind 0.05 oder 0.01.
- Die Menge $\mathcal{C} = \{x \in \mathcal{X}; D(x) = A\}$ heißt der **kritische Bereich** des Signifikanztests, ihr Komplement $\bar{\mathcal{C}} = \{x \in \mathcal{X}; D(x) = H_0\}$ der **Annahmebereich**.

Wie die letzte Anmerkung verdeutlicht, besteht die Konstruktion eines Signifikanztests einfach aus der Wahl einer Teilmenge $\mathcal{C} \subset \mathcal{X}$ als kritischem Bereich, d.h. einer Menge mit der Eigenschaft

$$P_\theta(\mathcal{C}) \leq \alpha \quad \text{für alle } \theta \in H_0 \quad (5.34)$$

Ist nach Durchführung des zugehörigen statistischen Experiments das Ergebnis x ein Element des kritischen Bereichs, so gilt die **Alternative als signifikant nachgewiesen**, andernfalls sagt man, dass **die Nullhypothese angenommen** wurde.

5.3. Tests für den Mittelwert der Normalverteilung

Wie bei den Konfidenzintervallen beginnen wir wieder mit den klassischen Tests unter der Normalverteilungsannahme. Das Ergebnis ist dabei ein Vektor $x = (x_1, x_2, \dots, x_n)$ von reellen Zahlen und der Parameter ist $\theta = (\mu, \sigma^2)$, wobei die einzelnen Messwerte x_k aus unabhängigen Zufallsexperimenten mit einer $\mathcal{N}(\mu, \sigma^2)$ -Verteilung entstehen.

5.3.1. Einseitige Hypothesen

In Verallgemeinerung des einführenden Beispiels sei eine Zahl μ_0 als Grenzwert für den (unbekannten) Mittelwert μ vorgegeben und ein kritischer Bereich \mathcal{C} für die Nullhypothese

$$H_0 = \{(\mu, \sigma^2) ; \mu \leq \mu_0\}$$

und die Alternative

$$A = \{(\mu, \sigma^2) ; \mu > \mu_0\}$$

zur Signifikanzzahl α zu bestimmen.

Als Schätzwert für den Abstand des unbekanntes Mittelwertes μ vom Grenzwert μ_0 bietet sich die Differenz $\bar{x} - \mu_0$ des arithmetischen Mittels der Messergebnisse und diesem Wert an. Da aber — wie sich aus der Formel für das entsprechende Konfidenzintervall zeigt — dieser Mittelwert als Schätzwert für μ umso unzuverlässiger ist, je mehr die Messwerte um den gemeinsamen Mittelwert streuen, sollte man die Differenz $\bar{x} - \mu_0$ relativ zur Größe der Stichprobenvarianz s^2 betrachten. Das führt uns zur Pivotvariable

$$Q(x, \mu_0) = \sqrt{\frac{n}{s^2}}(\bar{x} - \mu_0)$$

für den Mittelwert an der Stelle μ_0 als *vernünftigen* Abstandsmaß und zu

$$\mathcal{C} = \{x ; Q(x, \mu_0) > c\}$$

mit einem noch zu bestimmenden Schwellenwert c als Ansatz für einen kritischen Bereich. Für $\theta = (\mu_0, \sigma^2)$ besagen die Eigenschaften einer Pivotvariablen, dass

$$P_{(\mu_0, \sigma^2)}(\mathcal{C}) = P_{(\mu_0, \sigma^2)}(Q(\cdot, \mu_0) > c) = P^Q(c, \infty) = 1 - F^Q(c)$$

wo P^Q die Verteilung der Pivotvariablen und F^Q die zugehörige Verteilungsfunktion ist. Im vorliegenden Fall handelt es sich um die t-Verteilung mit $n - 1$ Freiheitsgraden. Ist $\theta = (\mu, \sigma^2) \in H_0$, d.h. $\mu \leq \mu_0$, so folgt aus

$$Q(x, \mu_0) = \sqrt{\frac{n}{s^2}}(\bar{x} - \mu_0) \leq \sqrt{\frac{n}{s^2}}(\bar{x} - \mu) = Q(x, \mu)$$

dass

$$\mathcal{C} = \{x ; Q(x, \mu_0) > c\} \subset \{x ; Q(x, \mu) > c\}$$

und damit

$$P_{(\mu, \sigma^2)}(\mathcal{C}) \leq P_{(\mu, \sigma^2)}(Q(\cdot, \mu) > c) = P^Q(c, \infty) = 1 - F^Q(c)$$

Wählt man c so, dass $1 - F^Q(c) = \alpha$, d.h. $F^Q(c) = 1 - \alpha$, so folgt

$$P_{(\mu, \sigma^2)}(\mathcal{C}) \leq \alpha$$

für alle $\theta = (\mu, \sigma^2) \in H_0$. Die Menge \mathcal{C} ist somit ein kritischer Bereich zur Signifikanzzahl α .

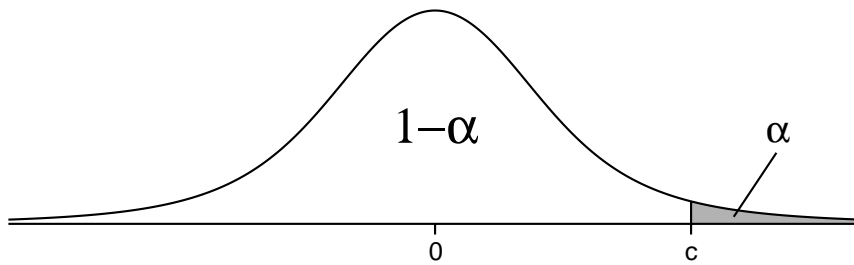


Abbildung 3: Einseitige Hypothesen

Mit dem Graphen der Dichte $f^Q(t)$ der Verteilung P^Q , s. Abbildung 3, kann man sich die Durchführung des Tests wie folgt veranschaulichen:

Zur Signifikanzzahl α bestimme man eine Zahl c so, dass die Fläche unter dem Graphen von f^Q ab der Stelle c (das ist gerade $1 - F^Q(c)$) gleich α ist.

Liegt der Wert

$$T(x) = \sqrt{\frac{n}{s^2}}(\bar{x} - \mu_0)$$

oberhalb von c , so gilt die Alternative " $\mu > \mu_0$ " als signifikant nachgewiesen, andernfalls wird die Nullhypothese angenommen.

5.3.2. Der p-Wert

Das Ergebnis eines Signifikanztests hängt natürlich auch vom Signifikanzniveau α ab. Für zwei Signifikanzzahlen $\alpha' < \alpha$ gilt wegen der Monotonie der Verteilungsfunktion F^Q für die zugehörigen Lösungen der Gleichungen

$$F^Q(c') = 1 - \alpha' \quad \text{und} \quad F^Q(c) = 1 - \alpha$$

dass $c' > c$. Es kann also der Fall eintreten, dass $c < T(x) < c'$. D.h., bezüglich der Signifikanzzahl α gilt die Alternative als signifikant nachgewiesen, während bei α' die Nullhypothese angenommen wird.

In der Statistik ist es daher üblich, das Ergebnis eines Signifikanztests so zu formulieren, dass der Leser des Ergebnisberichtes die Konsequenz gemäß seinen persönlichen Vorstellungen bezüglich des Signifikanzniveaus ziehen kann.

Statistik-Programmpakete verlangen bei der Datenangabe daher keine Signifikanzzahl und geben als Rechenergebnis keine Entscheidung sondern den so genannten p-Wert aus.

Der **p-Wert** $p(x)$ eines Signifikanztests ist definiert als **die kleinste Signifikanzzahl**, bezüglich der beim Ergebnis x die Alternative als signifikant nachgewiesen gilt.

Wie man anhand der Abbildung 4 nachvollziehen kann, ist bei dem im vorhergehenden Abschnitt beschriebenen Test mit der Alternative $\mu > \mu_0$ für den Mittelwert der Normalverteilung der p-Wert gegeben durch

$$p(x) = 1 - F^Q(T(x))$$

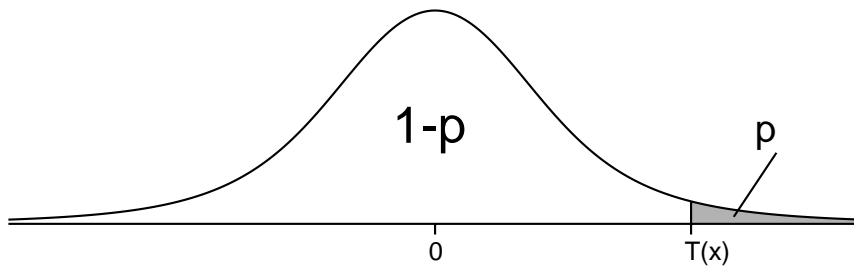


Abbildung 4: Der p-Wert bei einseitigen Hypothesen

Bei einer Signifikanzzahl $\alpha' < p(x)$ würde mit dem zugehörigen Schwellenwert c' gelten, dass $T(x) < c'$, d.h. die Nullhypothese wird angenommen. Ist c der Schwellenwert zu $\alpha > p(x)$, so ist $T(x) > c$ und die Alternative gilt als nachgewiesen.

$T(x)$ ist der Schwellenwert zur Signifikanzzahl $p(x)$. Nach der Formulierung des Tests im letzten Abschnitt läge damit streng genommen das Ergebnis x nicht mehr im kritischen Bereich \mathcal{C} , sondern auf dem Rand. Wegen der Stetigkeit der Verteilungsannahme ändert sich aber nichts an den Wahrscheinlichkeiten, wenn wir den Rand noch zur Menge \mathcal{C} dazunehmen.

5.3.3. Einfache Nullhypothese

Zur Beantwortung der Frage, ob der Mittelwert μ der Normalverteilung gleich einem gegebenen Wert μ_0 ist oder nicht, kann man die Hypothese „ $\mu = \mu_0$ “ nicht als Alternative wählen, denn Messwerte, die aus Zufallsexperimenten herrühren, lassen keine exakten Rückschlüsse auf die Parameter zu. Man kann nur die Nullhypothese $H_0 : \mu = \mu_0$ gegen die Alternative $A : \mu \neq \mu_0$ testen. Eine Nullhypothese dieser Form nennt man eine **einfache Nullhypothese**. Sieht man die Nullhypothese als widerlegt an, wenn der Mittelwert \bar{x} als Schätzwert für μ deutlich kleiner oder deutlich größer als μ_0 ist, erhält man mit den gleichen Argumenten wie beim Fall einseitiger Hypothesen als Ansatz für einen kritischen Bereich die Menge

$$\mathcal{C} = \left\{ x \in \mathbb{R}^n ; \sqrt{\frac{n}{s^2}}(\bar{x} - \mu_0) < c_1 \text{ oder } \sqrt{\frac{n}{s^2}}(\bar{x} - \mu_0) > c_2 \right\}$$

mit noch festzulegenden Grenzen c_1 und c_2 .

Ihr Komplement, die Menge

$$\bar{\mathcal{C}} = \{x \in \mathbb{R}^n; c_1 \leq \sqrt{\frac{n}{s^2}}(\bar{x} - \mu_0) \leq c_2\}$$

heißt auch der **Annahmebereich**. Wegen $\mathcal{P}_\theta(\bar{\mathcal{C}}) = 1 - P_\theta(\mathcal{C})$ ist $\bar{\mathcal{C}}$ dann ein kritischer Bereich zur Signifikanzzahl α , wenn für alle Parameter $\theta \in H_0$ die Ungleichung

$$P_\theta(\bar{\mathcal{C}}) \geq 1 - \alpha$$

erfüllt ist. Wegen $H_0 = \{(\mu_0, \sigma^2) ; \sigma^2 > 0\}$ und der Tatsache, dass

$$Q(x, \mu) = \sqrt{\frac{n}{s^2}}(\bar{x} - \mu)$$

Pivotvariable für μ ist, gilt

$$\begin{aligned} P_{(\mu_0, \sigma^2)}(\bar{\mathcal{C}}) &= P_{(\mu_0, \sigma^2)}(c_1 \leq Q(\cdot, \mu_0) \leq c_2) \\ &= P^Q[c_1, c_2] \\ &= F^Q(c_2) - F^Q(c_1) \end{aligned}$$

mit der Verteilungsfunktion F^Q der Pivotvariablen, d.h. in diesem Fall der Verteilungsfunktion der t_{n-1} -Verteilung.

Wählt man c_1 und c_2 so, dass $F^Q(c_2) - F^Q(c_1) = 1 - \alpha$, erhält man einen Annahmehereich bzw kritischen Bereich zur Signifikanzzahl α .

Wie in Abbildung 5 dargestellt, wählt man den Annahmehereich symmetrisch bezüglich der Fehlerwahrscheinlichkeiten, d.h.

$$F^Q(c_1) = \frac{\alpha}{2} \text{ und } F^Q(c_2) = 1 - \frac{\alpha}{2},$$

was wegen der Symmetrie der t -Verteilung gleichzeitig bedeutet, dass $c_1 = -c_2$.

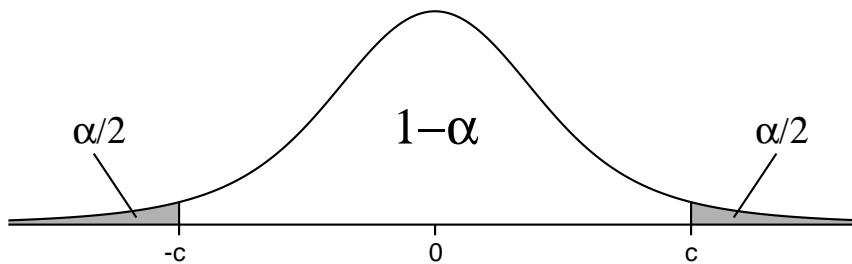


Abbildung 5: Einfache Nullhypothese

Wie man der Abbildung 6 entnimmt, erhält man den p-Wert $p(x)$ für diesen Test aus der Gleichung

$$F^Q(T(x)) = 1 - \frac{p(x)}{2}$$

bzw

$$F^Q(T(x)) = \frac{p(x)}{2}$$

je nachdem, ob der Wert $T(x) = Q(x, \mu_0)$ positiv oder negativ ist. Unter Verwendung des Absolutbetrags kann man dies zu

$$p(x) = 2(1 - F^Q(|T(x)|))$$

zusammenfassen.

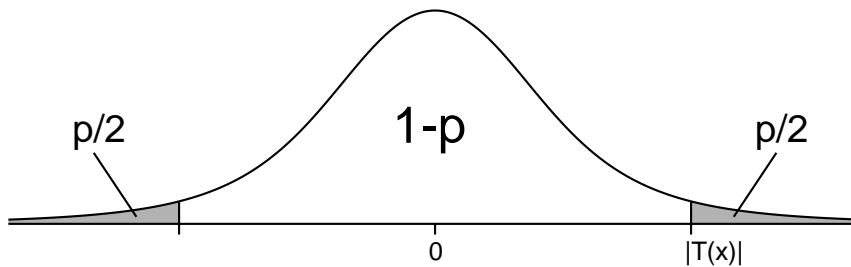


Abbildung 6: p-Wert bei einfacher Nullhypothese

5.3.4. Berechnung mit R

Tests für den Mittelwert der Normalverteilung werden in R mit dem Kommando `t.test` berechnet. In den gerade behandelten Fällen der unabhängigen Wiederholung eines $\mathcal{N}(\mu, \sigma^2)$ -Experiments lautet der Aufruf

```
t.test(x, alternative = art, mu =  $\mu_0$  (,conf.level =  $\gamma$ ))
```

Dabei ist `x` der Vektor der Messwerte, `art = "two.sided", "greater" oder "less"` die Art der Alternative und μ_0 die Grenze zwischen Nullhypothese und Alternative. Dabei wird, wie im vorhergehenden Kapitel erwähnt, gleichzeitig ein Konfidenzintervall für den Mittelwert berechnet, zu dem optional die Konfidenzzahl γ angegeben werden kann. Voreingestellt ist $\gamma = 0.95$.

Für die Werte aus Tabelle 3 und die einseitige Alternative $A : \mu > 120$ erhält man

```
> x <- scan("a08.txt")
Read 90 items
> t.test(x,alternative="greater",mu=120)

One Sample t-test

data:  x
t = 0.3043, df = 89, p-value = 0.3808
alternative hypothesis: true mean is greater than 120
95 percent confidence interval:
 119.3109      Inf
sample estimates:
mean of x
 120.1544
```

`t` ist der Wert $T(x)$, `df` die Anzahl der Freiheitsgrade der t-Verteilung der Pivotvariable und `p-value` der p-Wert. Letzterer liegt hier weit über den üblichen Signifikanzzahlen, so dass die Nullhypothese $H_0 : \mu \leq 120$ nicht widerlegt werden konnte.

In diesem Beispiel war aufgrund der Problemstellung klar, welche Alternative gewählt werden musste. Wenn man bei den Schmelzpunktwerten der Tabelle 1 untersuchen will,

ob der Schmelzpunkt gleich 63.5 Grad Celsius ist, so darf man sich aus der Tatsache, dass das Stichprobenmittel $\bar{x} = 63.58881$ größer als 63.5 ist, nicht dazu verleiten lassen, eine einseitige Alternative $\mu > 63.5$ anzusetzen! **Die Hypothesen müssen stets vor der Durchführung des Experiments festgelegt werden!** Falls zu diesem Zeitpunkt nichts darüber bekannt ist, dass der Schmelzpunkt höchstens höher liegen kann, muss eine einfache Nullhypothese $\mu = 63.5$ und eine zweiseitige Alternative $\mu \neq 63.5$ gewählt werden.

```
> x <- scan("bees.txt")
Read 59 items
> t.test(x,alt="two.sided",mu=63.5)

One Sample t-test

data:  x
t = 1.9647, df = 58, p-value = 0.05424
alternative hypothesis: true mean is not equal to 63.5
95 percent confidence interval:
 63.49833 63.67930
sample estimates:
mean of x
 63.58881
```

Der p-Wert liegt knapp über einer Signifikanzzahl von $\alpha = 0.05$, die Alternative ist damit auch hier nicht nachweisbar. In der Praxis sollte man aber in einer solchen Grenzfall wenn möglich noch weitere Messungen durchführen.

5.4. Konstruktionsverfahren für kritische Bereiche

Aus der Herleitung der kritischen Bereiche zu Tests für den Mittelwert der Normalverteilung ergibt sich ein allgemeines Rezept für den Fall, dass die Hypothesen von der Form

$$H_0 : \tau(\theta) \in I \text{ und } A : \tau(\theta) \notin I$$

sind, wobei $\tau(\theta)$ eine Parameterfunktion auf einem statistischen Raum, I eine Teilmenge von reellen Zahlen und außerdem eine Pivotvariable $Q(x, t)$ für diese Parameterfunktion bekannt ist.

Bei den beschriebenen Tests ist $I = (-\infty, \mu_0]$ bzw $I = \{\mu_0\}$.

Satz 5.1 *Ist P^Q die Verteilung der Pivotvariable $Q(x, t)$ und $B \subset \mathbb{R}$ eine Borelsche Menge mit $P^Q(B) = \alpha$, so ist*

$$C = \{x \in \mathcal{X} ; Q(x, t) \in B \text{ für alle } t \in I\}$$

ein kritischer Bereich zur Signifikanzzahl α

Beweis: Es ist

$$\mathcal{C} = \bigcap_{t \in I} \{x \in \mathcal{X} ; Q(x, t) \in B\}$$

Die Mengen $\mathcal{C}_t = \{x \in \mathcal{X} ; Q(x, t) \in B\}$ sind Obermengen der Menge \mathcal{C} : $\mathcal{C} \subset \mathcal{C}_t$ für alle $t \in I$.

Ist $\theta \in H_0$, d.h. $\tau(\theta) \in I$, so gilt daher $\mathcal{C} \subset \mathcal{C}_{\tau(\theta)}$ und

$$P_\theta(\mathcal{C}) \leq P_\theta(\mathcal{C}_{\tau(\theta)}) = P_\theta(Q(\cdot, \tau(\theta)) \in B) = P^Q(B) = \alpha$$

Wie die Menge B zu wählen ist, hängt natürlich von der jeweiligen Problemstellung ab. Auf die Theorie und Konstruktion optimaler Tests wollen wir in dieser Vorlesung nicht weiter eingehen.

5.5. Vergleich zweier Mittelwerte

Der Vergleich der Mittelwerte der Verteilungen zweier Messreihen ist eine der häufigsten Anwendungsfälle der Statistik.

Dabei liegen Messreihen

$$x_1, x_2, \dots, x_{n_1} \text{ und } y_1, y_2, \dots, y_{n_2}$$

mit reellen Zahlen vor, von denen man annimmt, dass sie aus unabhängigen Zufallsexperimenten mit $\mathcal{N}(\mu_1, \sigma_1^2)$ - bzw. $\mathcal{N}(\mu_2, \sigma_2^2)$ -Verteilung herrühren. Zu entscheiden ist, ob sich die Mittelwerte μ_1 und μ_2 unterscheiden oder nicht.

Mit $n = n_1 + n_2$ hat man es mit dem Stichprobenraum $\mathcal{X} = \mathbb{R}^n$ zu tun, dessen Elemente wir je nach Situation mit

$$z = (x, y) = (x_1, x_2, \dots, x_{n_1}, y_1, y_2, \dots, y_{n_2})$$

bezeichnen, und der Parameter ist

$$\theta = (\mu_1, \mu_2, \sigma_1^2, \sigma_2^2)$$

Die Verteilungsannahme lässt sich wie im Rahmen der Normalverteilung bisher schon üblich dahingehend formulieren, dass die Koordinatenvariablen X_1, X_2, \dots, X_{n_1} und Y_1, Y_2, \dots, Y_{n_2} bezüglich der Wahrscheinlichkeit P_θ stochastisch unabhängig sind, die X_i für alle $i = 1, \dots, n_1$ eine $\mathcal{N}(\mu_1, \sigma_1^2)$ -Verteilung und die Y_k für $k = 1, \dots, n_2$ eine $\mathcal{N}(\mu_2, \sigma_2^2)$ -Verteilung besitzen.

Die Hypothesen lassen sich mit Hilfe der Parameterfunktion

$$\tau(\theta) = \tau(\mu_1, \mu_2, \sigma_1^2, \sigma_2^2) = \mu_2 - \mu_1$$

formulieren, so dass wir *nur noch* eine Pivotvariable für diese Parameterfunktion finden müssen.

Eine für diese Parameterfunktion erwartungstreue Schätzfunktion ist

$$\bar{Y} - \bar{X} = \frac{1}{n_2}(Y_1 + \dots + Y_{n_2}) - \frac{1}{n_1}(X_1 + \dots + X_{n_1})$$

Sie ist als Linearkombination von normalverteilten Zufallsvariablen ebenfalls normalverteilt mit dem Erwartungswert

$$\mathcal{E}_\theta(\bar{Y} - \bar{X}) = \mathcal{E}_\theta \bar{Y} - \mathcal{E}_\theta \bar{X} = \mu_2 - \mu_1 = \tau(\theta)$$

und der Varianz

$$\text{var}_\theta(\bar{Y} - \bar{X}) = \text{var}_\theta(\bar{X}) + \text{var}_\theta(\bar{Y}) = \frac{1}{n_1}\sigma_1^2 + \frac{1}{n_2}\sigma_2^2$$

Entsprechend der Vorgehensweise bei der Konstruktion einer Pivotvariable für den Mittelwert würde man annehmen, dass man durch Division dieser Schätzfunktion durch eine passende Funktion der jeweiligen Stichprobenvarianzen ebenfalls zu einer Pivotvariablen gelangt. Für vollständig beliebige unbekannte Varianzen σ_1^2 und σ_2^2 erwies sich dies allerdings als eines der schwierigsten Probleme der mathematischen Statistik, das sog. Behrens-Fisher-Problem, auf das wir am Ende dieses Kapitels noch kurz eingehen werden.

Unter der Voraussetzung, dass die Varianzen zwar als unbekannt aber gleich angenommen werden können:

$$\sigma_1^2 = \sigma_2^2 = \sigma^2$$

wurde das Problem von STUDENT (Paul Gosset) gelöst.

5.5.1. Der Studentsche t-Test

Unter der Voraussetzung $\sigma_1^2 = \sigma_2^2 = \sigma^2$, d.h. bezüglich des verkürzten Parametervektors $\theta = (\mu_1, \mu_2, \sigma^2)$ sind die Koordinatenvariablen X_j und Y_k normalverteilt mit Mittelwerten μ_1 bzw. μ_2 und stets der gleichen Varianz σ^2 .

Dementsprechend sind

$$G_j = \frac{1}{\sqrt{\sigma^2}}(X_j - \mu_1) \quad \text{und} \quad H_k = \frac{1}{\sqrt{\sigma^2}}(Y_k - \mu_2)$$

stochastisch unabhängige $\mathcal{N}(0, 1)$ -verteilte Zufallsvariable, die wir ähnlich wie in Abschnitt 4.2 zu einem gaussischen Einheitsvektor zusammensetzen und darauf eine Orthogonalmatrix der folgenden Gestalt anwenden.

$$\begin{pmatrix} \tilde{G}_1 \\ \tilde{G}_2 \\ \vdots \\ \tilde{G}_{n_1} \\ \hline \tilde{H}_1 \\ \tilde{H}_2 \\ \vdots \\ \tilde{H}_{n_2} \end{pmatrix} = \begin{pmatrix} \frac{1}{\sqrt{n_1}} & \frac{1}{\sqrt{n_1}} & \cdots & \frac{1}{\sqrt{n_1}} & & & \\ u_{21} & u_{22} & \cdots & u_{2n_1} & & & \\ \vdots & \vdots & & \vdots & & & \\ u_{n_1 1} & u_{n_1 2} & \cdots & u_{n_1 n_1} & & & \\ \hline & & & & \frac{1}{\sqrt{n_2}} & \frac{1}{\sqrt{n_2}} & \cdots & \frac{1}{\sqrt{n_2}} \\ & & & & v_{21} & v_{22} & \cdots & v_{2n_2} \\ & & & & \vdots & \vdots & & \vdots \\ & & & & v_{n_2 1} & v_{n_2 2} & \cdots & v_{n_2 n_2} \end{pmatrix} \begin{pmatrix} G_1 \\ G_2 \\ \vdots \\ G_{n_1} \\ \hline H_1 \\ H_2 \\ \vdots \\ H_{n_2} \end{pmatrix} \quad (5.35)$$

Für die beiden $\mathcal{N}(0, 1)$ -verteilten Zufallsvariablen \tilde{G}_1 und \tilde{H}_1 gilt

$$\tilde{G}_1 = \sqrt{\frac{n_1}{\sigma^2}}(\bar{X} - \mu_1) \quad \text{und} \quad \tilde{H}_1 = \sqrt{\frac{n_2}{\sigma^2}}(\bar{Y} - \mu_2)$$

Die Stichprobenmittelwerte $\bar{X} = \mu_1 + \sqrt{\frac{\sigma^2}{n_1}}\tilde{G}_1$ und $\bar{Y} = \mu_2 + \sqrt{\frac{\sigma^2}{n_2}}\tilde{H}_1$ sind daher $\mathcal{N}\left(\mu_1, \frac{\sigma^2}{n_1}\right)$ - bzw. $\mathcal{N}\left(\mu_2, \frac{\sigma^2}{n_2}\right)$ -verteilt.

Die Differenz $\bar{Y} - \bar{X}$ ist dann ebenfalls normalverteilt mit dem Mittelwert

$$\begin{aligned}
\mathcal{E}_{(\mu_1, \mu_2, \sigma^2)}(\bar{Y} - \bar{X}) &= \mathcal{E}_{(\mu_1, \mu_2, \sigma^2)}\bar{Y} - \\
&- \mathcal{E}_{(\mu_1, \mu_2, \sigma^2)}\bar{X} = \mu_2 - \mu_1 = \tau(\mu_1, \mu_2, \sigma^2)
\end{aligned}$$

und der Varianz

$$\begin{aligned}
\text{var}_{(\mu_1, \mu_2, \sigma^2)}(\bar{Y} - \bar{X}) &= \text{var}_{(\mu_1, \mu_2, \sigma^2)}\bar{Y} + \\
(-1)^2 \text{var}_{(\mu_1, \mu_2, \sigma^2)}(\bar{X}) &= \frac{\sigma^2}{n_2} + \frac{\sigma^2}{n_1} \\
&= \frac{n_1 + n_2}{n_1 n_2} \sigma^2
\end{aligned}$$

Durch die übliche Normierung erhält man eine gaussische Einheitsvariable

$$H_0 = \sqrt{\frac{n_1 n_2}{(n_1 + n_2) \sigma^2}} (\bar{Y} - \bar{X} - (\mu_2 - \mu_1))$$

die nur von \tilde{G}_1 und \tilde{H}_1 abhängt und somit stochastisch unabhängig von $\tilde{G}_2, \dots, \tilde{G}_{n_1}, \tilde{H}_2, \dots, \tilde{H}_{n_2}$ ist.

Wie in Abschnitt 4.2 erhält man weiter

$$\tilde{G}_2^2 + \dots + \tilde{G}_{n_1}^2 = \frac{1}{\sigma^2} \sum_{j=1}^{n_1} (X_j - \bar{X})^2 = \frac{n_1 - 1}{\sigma^2} S_x^2$$

$$\tilde{H}_2^2 + \dots + \tilde{H}_{n_2}^2 = \frac{1}{\sigma^2} \sum_{k=1}^{n_2} (Y_k - \bar{Y})^2 = \frac{n_2 - 1}{\sigma^2} S_y^2$$

und daraus die mit $n_1 + n_2 - 2$ Freiheitsgraden Chi-Quadrat-verteilte Zufallsvariable

$$\tilde{G}_2^2 + \dots + \tilde{G}_{n_1}^2 + \tilde{H}_2^2 + \dots + \tilde{H}_{n_2}^2 = \frac{1}{\sigma^2} \left((n_1 - 1) S_x^2 + (n_2 - 1) S_y^2 \right)$$

Gemäß (4.25) erhält man durch die Konstruktion

$$\begin{aligned} & \frac{H_0}{\sqrt{\frac{1}{n_1+n_2-2} (\tilde{G}_2^2 + \dots + \tilde{G}_{n_1}^2 + \tilde{H}_2^2 + \dots + \tilde{H}_{n_2}^2)}} \\ &= \frac{\sqrt{\frac{n_1 n_2}{(n_1+n_2)\sigma^2}} (\bar{Y} - \bar{X} - (\mu_2 - \mu_1))}{\sqrt{\frac{1}{n_1+n_2-2} \frac{1}{\sigma^2} \left((n_1 - 1) S_x^2 + (n_2 - 1) S_y^2 \right)}} \\ &= \sqrt{\frac{n_1 n_2 (n_1 + n_2 - 2)}{n_1 + n_2}} \frac{\bar{Y} - \bar{X} - (\mu_2 - \mu_1)}{\sqrt{(n_1 - 1) S_x^2 + (n_2 - 1) S_y^2}} \end{aligned}$$

eine t-verteilte Zufallsvariable, die nicht mehr den Parameter σ^2 enthält, sondern vom Zustand nur über den Wert $\tau(\mu_1, \mu_2, \sigma^2) = \mu_2 - \mu_1$ der Parameterfunktion τ abhängt. Das bedeutet, dass die Funktion

$$\begin{aligned} Q(x, y, t) &= \sqrt{\frac{n_1 n_2 (n_1 + n_2 - 2)}{n_1 + n_2}} \frac{\bar{Y}(x, y) - \bar{X}(x, y) - t}{\sqrt{(n_1 - 1) S_x^2(x, y) + (n_2 - 1) S_y^2(x, y)}} \\ &= \sqrt{\frac{n_1 n_2 (n_1 + n_2 - 2)}{n_1 + n_1}} \frac{\bar{x} - \bar{y} - t}{\sqrt{(n_1 - 1) s_x^2 + (n_2 - 1) s_y^2}} \end{aligned}$$

eine mit $n_1 + n_2 - 2$ Freiheitsgraden t -verteilte Pivotvariable für die Parameterfunktion τ ist.

Signifikanztest: Der Test mit der einfachen Nullhypothese „ $\tau(\theta) = 0$ “ bzw. „ $\mu_1 = \mu_2$ “ gegen die Alternative „ $\tau(\theta) \neq 0$ “ besitzt einen Annahmebereich mit der gleichen Struktur wie der zu der einfachen Hypothese „ $\mu = \mu_0$ “ über den Mittelwert der Normalverteilung:

$$\bar{C} = \{(x, y) ; -c \leq T(x, y) \leq c\}$$

mit

$$T(x, y) = Q(x, y, 0) = \sqrt{\frac{n_1 n_2 (n_1 + n_2 - 2)}{n_1 + n_2}} \frac{\bar{y} - \bar{x}}{\sqrt{(n_1 - 1)s_x^2 + (n_2 - 1)s_y^2}} \quad (5.36)$$

Der einzige Unterschied besteht darin, dass zur Festlegung der kritischen Grenze c mit der Gleichung $F(c) = 1 - \frac{\alpha}{2}$ bzw. zur Berechnung des p -Wertes

$$p = 2 \left(1 - F(|T(x, y)|) \right)$$

die Verteilungsfunktion der t -Verteilung mit der Anzahl $n_1 + n_2 - 2$ von Freiheitsgraden verwendet werden muss.

5.5.2. Vergleich zweier Varianzen

Der STUDENTSche t -Test liefert nur korrekte Resultate, wenn die Varianz bei den beiden verglichenen Messreihen gleich ist: $\sigma_1^2 = \sigma_2^2 = \sigma^2$.

Falls man nicht weiß, ob diese Voraussetzung gegeben ist, kann man versuchen, anhand der Messwerte durch einen Test zu entscheiden, ob man die Nullhypothese „ $\sigma_1^2 = \sigma_2^2$ “ annehmen kann.

Bezüglich des Parameters $\theta = (\mu_1, \mu_2, \sigma^2, \sigma^2)$ für den allgemeinen Fall erhält man aus der Orthogonaltransformation (5.35) die Gleichungen

$$\tilde{G}_2^2 + \tilde{G}_3^2 + \dots + \tilde{G}_{n_1}^2 = \frac{n_1 - 1}{\sigma_1^2} S_x^2$$

und

$$\tilde{H}_2^2 + \tilde{H}_3^2 + \dots + \tilde{H}_{n_2}^2 = \frac{n_2 - 1}{\sigma_2^2} S_y^2$$

bzw.

$$\left(\frac{\sigma_2^2}{\sigma_1^2} \right) \cdot \frac{S_x^2}{S_y^2} = \frac{\frac{1}{n_1 - 1} (\tilde{G}_2^2 + \dots + \tilde{G}_{n_1}^2)}{\frac{1}{n_2 - 1} (\tilde{H}_2^2 + \dots + \tilde{H}_{n_2}^2)} \quad (5.37)$$

Die F-Verteilung

Definition 5.2 sind $G_1, G_2, \dots, G_m, H_1, H_2, \dots, H_n$ stochastisch unabhängige $\mathcal{N}(0, 1)$ -verteilte Zufallsvariable auf einem Wahrscheinlichkeitsraum (Ω, A, P) , so heißt die Verteilung der Zufallsvariablen

$$Y = \frac{\frac{1}{m} (G_1^2 + G_2^2 + \dots + G_m^2)}{\frac{1}{n} (H_1^2 + H_2^2 + \dots + H_n^2)}$$

die **F-Verteilung** mit (m, n) Freiheitsgraden oder kurz $f_{(m, n)}$ -Verteilung.

Sie besitzt die Dichte

$$f_{(m,n)}(t) = \begin{cases} \frac{\left(\frac{m}{n}\right)^{m/2} t^{m/2-1}}{B\left(\frac{m}{2}, \frac{n}{2}\right)} \left(1 + \frac{m}{n}t\right)^{-\frac{m+n}{2}} & \text{für } t > 0 \\ 0 & \text{für } t \leq 0 \end{cases}$$

mit der Beta-Funktion

$$B(p, q) = \int_0^1 x^{p-1} (1-x)^{q-1} dx$$

Der F-Test

Aus (5.37) liest man wieder ab, dass die Funktion

$$Q(x, y, t) = t \cdot \frac{S_x^2(x, y)}{S_y^2(x, y)} = t \frac{s_x^2}{s_y^2} \quad (5.38)$$

eine Pivotvariable ist, dieses Mal für die Parameterfunktion

$$\tau(\mu_1, \mu_2, \sigma_1^2, \sigma_2^2) = \frac{\sigma_2^2}{\sigma_1^2}$$

und mit der $f_{(n_1-1, n_2-1)}$ -Verteilung als P^Q .

Bezeichnet $F(t)$ die zugehörige Verteilungsfunktion, so gilt

$$P_{(\mu_1, \mu_2, \sigma_1^2, \sigma_2^2)}\left(c_1 \leq Q(\cdot, \cdot, \frac{\sigma_2^2}{\sigma_1^2}) \leq c_2\right) = F(c_2) - F(c_1)$$

Um beim Vergleich zweier Mittelwerte zu prüfen, ob die Voraussetzung gleicher Varianzen angenommen werden kann, führt man einen Test mit der einfachen Nullhypothese „ $\tau(\theta) = 1$ “ durch.

Als Annahmebereich wird die Menge

$$\begin{aligned} \bar{\mathcal{C}} &= \{(x, y) ; c_1 \leq \frac{s_x^2}{s_y^2} \leq c_2\} \\ &= \{(x, y) ; c_1 \leq Q(x, y, 1) \leq c_2\} \end{aligned}$$

gewählt, wobei man zu einer vorgegebenen Signifikanzzahl α die Grenzen c_1 und c_2 aus den Gleichungen $F(c_1) = \frac{\alpha}{2}$ und $F(c_2) = 1 - \frac{\alpha}{2}$ bestimmt. Für die Parameter θ aus der Nullhypothese gilt dann

$$P_\theta(\bar{\mathcal{C}}) = F(c_2) - F(c_1) = 1 - \alpha$$

bzw. $P_\theta(\mathcal{C}) = \alpha$.

Da die F-Verteilung nicht symmetrisch ist, müssen bei der Berechnung des p -Werts zwei Fälle unterschieden werden. Mit $\hat{t} := \frac{s_x^2}{s_y^2}$ ist

$$p = \begin{cases} 2(1 - F(\hat{t})) & \text{falls } \hat{t} \geq 1 \\ 2F(\hat{t}) & \text{falls } \hat{t} < 1 \end{cases}$$

5.5.3. Der Welch-Test

Für das sogenannte Fischer-Behrens-Problem, einen Test zum Vergleich zweier Mittelwerte bei beliebigen unbekanntem Varianzen zu bestimmen, gibt es bisher nur zwei unbefriedigende Lösungen.

Einen exakten Test, d.h. einen kritischen Bereich, der auf einer Statistik mit bekannter Verteilung basiert, erhält man nur durch Verzicht auf einen Teil der statistischen Information, die in den Messreihen steckt. Bestehen die beiden Messreihen aus gleich vielen Messwerten, so betrachtet man die paarweisen Differenzen $z_i = y_i - x_i$, die wieder normalverteilt sind, wobei der Mittelwert μ der Differenz $\mu_2 - \mu_1$ entspricht.

Mit diesen Werten testet man auf die Nullhypothese „ $\mu = 0$ “.

Bei Messreihen unterschiedlicher Länge muss man bei dieser Vorgehensweise auf einige Messwerte verzichten.

Möchte man die statistische Information vollständig nutzen, so gibt es nur eine *asymptotische* Lösung, d.h. die Verteilung ist nur näherungsweise für große Stichprobenumfänge n_1 und n_2 bekannt.

Beim Test von Welch für die Nullhypothese „ $\mu_1 = \mu_2$ “ wird dabei die Statistik $T(x, y) = Q(x, y, 0)$ beim textscStudentschen t-Test (5.36) ersetzt durch

$$T(x, y) = \frac{\bar{y} - \bar{x}}{\sqrt{\frac{1}{n_1} s_x^2 + \frac{1}{n_2} s_y^2}}$$

die asymptotisch t-verteilt ist, wobei die Anzahl der Freiheitsgrade der ganzzahlige Anteil des Werts

$$k(x, y) = (n_1 - 1)(n_2 - 1) \frac{\left(\frac{S_x^2}{n_1} + \frac{S_y^2}{n_2}\right)^2}{(n_2 - 1)\left(\frac{S_x^2}{n_1}\right)^2 + (n_1 - 1)\left(\frac{S_y^2}{n_2}\right)^2}$$

ist.

5.5.4. Berechnung mit R

Nach diesem Abschnitt ist offensichtlich, warum der R-Befehl ausgerechnet `t.test` heißt. Bisher wurde er nur auf Spezialfälle angewandt. Für den Vergleich zweier Mittelwerte lautet die Kommandozeile

```
t.test(x, y, alternative = art, mu = mu_0, paired = [true/false], var.equal = [true/false], ...)
```


x und y sind die beiden Messreihen und μ_0 der Wert, mit dem $\tau(\theta) = \mu_1 - \mu_2$ verglichen wird. Voreingestellt ist $\mu = 0$. Mit *art* kann man wieder die Art der Alternative wählen. Mit `paired = TRUE` werden die paarweisen Differenzen der Messreihen getestet, wobei die Vektoren x und y gleich viele Komponenten besitzen müssen. Voreingestellt ist `paired = FALSE`.

Mit `var.equal = TRUE` bzw. `var.equal = FALSE` entscheidet man, ob der STUDENTSche t-Test bzw. der Welch-Test angewandt werden soll.

Vor der Anwendung des t-Tests sollte man prüfen, ob die Voraussetzung gleicher Varianzen angenommen werden kann. Dies geschieht mit

```
var.test(x, y, ratio =  $\tau$ , ...)
```

wobei τ der Wert ist, mit dem man den Quotienten σ_1^2/σ_2^2 vergleicht. Die Voreinstellung ist der in diesem Zusammenhang übliche Wert `ratio = 1`.

Um die Schätzungen der Hörsaalbreite von Beispiel 1.3 auf Seite 6 in den Maßeinheiten Meter und Fuss vergleichen zu können, muss man zuerst die Längenangaben auf gleiche Dimension bringen, denn sonst vergleicht man Äpfel mit Birnen. Wir rechnen Meter mit dem Konvertierungsfaktor 3.28 in Fuss um:

```
> rw <- read.table("rw.txt",header=TRUE)
> str(rw)
'data.frame': 113 obs. of 2 variables:
 $ unit : Factor w/ 2 levels "feet","metres": 2 2 2 2 2 2 2 2 2 2 ...
 $ width: int 8 9 10 10 10 10 10 10 11 11 ...
> convert <- 3.28
> x <- rw$width[rw$unit == "metres"] * convert
> y <- rw$width[rw$unit == "feet"]
> c(var(x),var(y))
[1] 549.1731 156.1854
>
```

Die Stichprobenvarianzen unterscheiden sich deutlich. Um zu entscheiden, ob trotzdem der STUDENTSche t-Test anwendbar ist, führen wir den F-Test durch:

```
> var.test(x,y,ratio=1)

F test to compare two variances

data: x and y
F = 3.5162, num df = 43, denom df = 68, p-value = 3.616e-06
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
 2.071286 6.174862
sample estimates:
ratio of variances
 3.516161
```

Das geht offensichtlich nicht, daher der Welch-Test. Die Stichprobenumfänge sind mit 44 bzw. 69 groß genug für ein zuverlässiges Resultat:

```
> t.test(x,y,alternative="two.sided",mu=0,var.equal=FALSE)
```

```
Welch Two Sample t-test
```

```
data: x and y
t = 2.3071, df = 58.788, p-value = 0.02459
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 1.17471 16.54308
sample estimates:
mean of x mean of y
 52.55455  43.69565
```

Der p-Wert ist $p(x, y) = 0.02459$. Bezüglich eines Signifikanzniveaus von $\alpha = 0.05$ ist damit signifikant nachgewiesen, dass sich die Einschätzungen mit verschiedenen Maßeinheiten unterscheiden. Bezüglich $\alpha = 0.01$ ist das nicht der Fall.

Nur zur Demonstration des Aufrufs hier noch das Resultat des STUDENTSchen t-Tests:

```
> t.test(x,y,alternative="two.sided",mu=0,var.equal=TRUE)
```

```
Two Sample t-test
```

```
data: x and y
t = 2.6147, df = 111, p-value = 0.01017
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 2.145052 15.572734
sample estimates:
mean of x mean of y
 52.55455  43.69565
```

6. Lineare Modelle

In diesem und den folgenden Kapiteln beschäftigen wir uns mit der Analyse von Messwerten x , die von einer Größe t abhängen, deren Wert vom Experimentator frei gewählt werden kann. Wenn wir das Experiment als System mit dem Input t und dem Output x interpretieren, können wir es schematisch in der Form



darstellen. Die Messwerte x sind dabei reelle Zahlen $x \in \mathbb{R}$ oder — was wir in dieser Vorlesung nicht behandeln — reelle Vektoren $x \in \mathbb{R}^n$.

t heißt in diesem Zusammenhang die **unabhängige Variable** und x die **abhängige Variable**.

Gesucht ist ein „Modell“ für diese Abhängigkeit, d.h. eine Funktion $g(t)$ aus einer vorgegebenen Funktionenklasse \mathcal{G} , die diese Abhängigkeit möglichst gut beschreibt.

Beispiel 6.1 *Ist $x(t)$ der Bremsweg bei einem Bremsversuch mit einem PKW, der bei Geschwindigkeit t gemessen wird, so ist die einfachste Erklärung für den Zusammenhang zwischen x und t die, dass sich der Bremsweg zusammensetzt aus der Strecke, die während der Schrecksekunde zurückgelegt wird und damit proportional zu t ist, und der Strecke die der eigentliche Bremsvorgang benötigt und damit proportional zur kinetischen Energie des Fahrzeugs bzw. dem Geschwindigkeitsquadrat t^2 ist.*

Man sucht daher nach einer passenden Funktion $g : \mathbb{R} \rightarrow \mathbb{R}$ aus der Menge

$$\mathcal{G} = \{g(t) = \beta_1 t + \beta_2 t^2; \beta_1, \beta_2, t \in \mathbb{R}\}$$

Allgemein gültige statistische Methoden kann man natürlich nur dann erwarten, wenn die Funktionenklasse \mathcal{G} nicht allzu komplex ist. Wir beschränken uns in dieser Vorlesung auf den Fall, dass die Funktion $g(t)$ Linearkombinationen

$$g(t) = \beta_1 g_1(t) + \beta_2 g_2(t) + \dots + \beta_p g_p(t) \quad (6.39)$$

mit bekannten Ansatzfunktionen $g_k(t)$ und beliebigen reellen Gewichtsfunktionen β_k sind.

6.1. Lineare Modelle

Zur Bestimmung der Gewichtsfaktoren β_k in (6.39) wird das System mit n nicht notwendig verschiedenen Werten t_i der unabhängigen Variablen t angestoßen und die zugehörigen Werte x_i der abhängigen Variablen werden gemessen.

Wenn man mit $e_i = x_i - g(t_i)$ die Abweichungen der gemessenen Werte x_i von den gemäß dem Modell (6.39) zu erwartenden Werten $g(t_i)$ bezeichnet, so gilt

$$x_i = g(t_i) + e_i = \sum_{k=1}^p \beta_k g_k(t_i) + e_i$$

Die Werte $g_{ik} := g_k(t_i)$ der Ansatzfunktionen an den Stellen t_i sind voraussetzungsgemäß bekannt. Mit dieser Bezeichnung ist

$$x_i = \sum_{k=1}^p g_{ik} \beta_k + e_i \quad \text{für } i = 1, \dots, n$$

ein lineares Gleichungssystem. Mit den Vektoren $x = (x_1, x_2, \dots, x_n)^\top$, $e = (e_1, e_2, \dots, e_n)^\top$, $b = (\beta_1, \beta_2, \dots, \beta_p)^\top$ und der Matrix

$$G = \begin{pmatrix} g_{11} & g_{12} & \cdots & g_{1p} \\ g_{21} & g_{22} & \cdots & g_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ g_{n1} & g_{n2} & \cdots & g_{np} \end{pmatrix} \quad (6.40)$$

kann man es kurz in der Form

$$x = Gb + e \quad (6.41)$$

schreiben.

Definition 6.1 Ein Ansatz der Form (6.41) mit einem bekannten Vektor $x \in \mathbb{R}^n$, einer bekannten $n \times p$ -Matrix G und einem zu bestimmenden Vektor $b \in \mathbb{R}^p$ heißt ein **lineares Modell** für den Messwertvektor x .

Zusätzliche Bedingungen

1. Mit genügend vielen Ansatzfunktionen $g_k(t)$ lässt es sich natürlich immer erreichen, dass das Gleichungssystem $x = Gb$ exakt gelöst wird, d.h. die Funktion $g(t)$ die Messwerte x_i vollständig erklärt.

Die grundlegende Ausgangsposition der Statistik ist aber, dass die Messergebnisse x_i sich aus den Werten $g(t_i)$ als theoretischen Mittelwerten und zufälligen Fehlern e_i zusammensetzen und eine Funktion $g(t)$ von möglichst geringer Komplexität gesucht wird, die die Messwerte ausreichend gut erklärt. Für lineare Modelle bedeutet das, dass die Anzahl p der Ansatzfunktionen bzw. freien Parameter β_k kleiner als die Anzahl n der Messwerte sein sollte:

$$p < n \quad (6.42)$$

2. Wenn die Spalten $g_{\cdot k} = (g_{1k}, g_{2k}, \dots, g_{nk})^\top$ der Matrix G linear abhängig sind, also der Rang r der Matrix G kleiner als p ist, gibt es r linear unabhängige Spalten $g_{\cdot k_1}, g_{\cdot k_2}, \dots, g_{\cdot k_r}$, so dass sich die restlichen $g_{\cdot k}$ als Linearkombination dieser Basisspalten darstellen lassen. In Matrixschreibweise bedeutet das mit der Matrix

$$\tilde{G} = (g_{\cdot k_1}, g_{\cdot k_2}, \dots, g_{\cdot k_r})$$

dass es zu jedem Vektor $b \in \mathbb{R}^p$ einen Vektor $\tilde{b} \in \mathbb{R}^r$ gibt, so dass

$$Gb = \tilde{G}\tilde{b}$$

Das bedeutet, dass wir im Hinblick auf möglichst geringe Komplexität anstelle des ursprünglichen Modells mit dem reduzierten Ansatz $x = \tilde{G}\tilde{b} + \tilde{e}$ arbeiten können.

Für das folgende nehmen wir daher an, dass der Ansatz (6.41) bereits einer solchen Reduzierung entspricht, d.h. die Spalten der Matrix G linear unabhängig sind, bzw.

$$\text{Rang}(G) = p < n \tag{6.43}$$

gilt.

6.2. Die Methode der kleinsten Quadrate

Unter den aufgeführten Voraussetzungen ist das Gleichungssystem $x = Gb$ im allgemeinen nicht lösbar. Für jeden Vektor $b \in \mathbb{R}^p$ bleibt ein Rest $e = e(b) = x - Gb$. Es ist daher ein Vektor b so zu bestimmen, dass der Abstand zwischen den Vektoren x und Gb möglichst klein wird.

Bei der Methode der kleinsten Quadrate wird dieser Abstand mit dem Quadrat der euklidischen Norm gemessen, d.h. mit der Größe

$$\|y\|^2 = y_1^2 + y_2^2 + \dots + y_m^2$$

für einen Vektor y mit den Komponenten y_1, y_2, \dots, y_m .

Zur Bestimmung des besten Parametervektors b nach dieser Methode ist somit die Zielfunktion

$$Z(b) = \|x - Gb\|^2 \tag{6.44}$$

zu minimieren.

Dass diese Zielfunktion überhaupt ein Minimum besitzt und die Minimalstellen dadurch charakterisiert sind, dass sie die sogenannten Normalgleichungen erfüllen, werden wir hier nicht in voller mathematischer Strenge beweisen, sondern die Herleitung anhand bekannter Resultate aus der linearen Algebra nur skizzieren.

Es sei

$$\mathcal{L} = \{z = Gb \in \mathbb{R}^n ; b \in \mathbb{R}^p\}$$

der Bildraum der Matrix G . Sind $g_{\cdot 1}, \dots, g_{\cdot p}$ die Spalten dieser Matrix, so kann man \mathcal{L} auch als den Untervektorraum des \mathbb{R}^n charakterisieren, der von diesen Spaltenvektoren aufgespannt wird:

$$\mathcal{L} = \{z = \beta_1 g_{\cdot 1} + \beta_2 g_{\cdot 2} + \dots + \beta_p g_{\cdot p} ; \beta_k \in \mathbb{R}\}$$

Gesucht ist ein Vektor $\hat{z} = G\hat{b} \in \mathcal{L}$, der von x minimalen (quadratischen) Abstand hat

$$\|x - \hat{z}\|^2 = \min_{z \in \mathcal{L}} \|x - z\|^2 = \min_{b \in \mathbb{R}^p} \|x - Gb\|^2$$

Aus der linearen Algebra ist bekannt, dass sich dieser Vektor \hat{z} dadurch auszeichnet, dass der Differenzvektor $x - \hat{z}$ orthogonal zu allen Vektoren des Untervektorraums \mathcal{L} ist. Insbesondere ist $x - \hat{z}$ orthogonal zu den Spalten $g_{\cdot,k}$, d.h.

$$\begin{aligned} g_{\cdot,1}^\top(x - \hat{z}) &= 0 \\ g_{\cdot,2}^\top(x - \hat{z}) &= 0 \\ &\vdots \\ g_{\cdot,p}^\top(x - \hat{z}) &= 0 \end{aligned}$$

Fasst man diese p Gleichungen zu einem Gleichungssystem zusammen, so lautet es $G^\top(x - \hat{z}) = 0$ oder $G^\top\hat{z} = G^\top x$.

Da der Vektor \hat{z} im Untervektorraum \mathcal{L} liegt, ist er von der Form $\hat{z} = G\hat{b}$. Setzt man dies in die obige Gleichung ein, so erhält man als Resultat, dass die Minimalstellen \hat{b} der Zielfunktion $Z(b)$ gerade die Lösungen des Gleichungssystems

$$G^\top Gb = G^\top x \tag{6.45}$$

sind.

Definition 6.2 Das Gleichungssystem (6.45) nennt man die *Gaußschen Normalgleichungen* oder kurz die **Normalgleichungen**.

Eindeutigkeit der Lösung

Der Vektor \hat{z} mit minimalem Abstand zu x ist eindeutig bestimmt. Er heißt die Projektion des Vektors x auf den Untervektorraum \mathcal{L} . Die Komponenten $\hat{\beta}_n$ des Vektors \hat{b} sind die Skalare, mit denen \hat{z} als Linearkombination der Spalten von G dargestellt wird:

$$\hat{z} = \hat{\beta}_1 g_{\cdot,1} + \hat{\beta}_2 g_{\cdot,2} + \dots + \hat{\beta}_p g_{\cdot,p}$$

Falls die Spalten $g_{\cdot,k}$ linear unabhängig sind, ist diese Linearkombination und damit der Vektor \hat{b} eindeutig bestimmt. Sind die Spalten linear abhängig, so gibt es mehrere Linearkombinationen zur Darstellung von \hat{z} und daher mehrere Lösungen \hat{b} der Normalgleichungen, wobei aber wegen $\hat{z} = G\hat{b}$ für alle diese Lösungen gilt, dass der Wert $Z(\hat{b})$ der Zielfunktion für alle gleich ist.

Unter den im vorhergehenden Abschnitt aufgeführten Voraussetzungen ist die Lösung eindeutig bestimmt:

Satz 6.1 *Besitzt die $n \times p$ -Matrix G den Rang p , so ist die $p \times p$ -Matrix $G^\top G$ nichtsingulär und die Normalgleichungen besitzen die eindeutig bestimmte Lösung*

$$\hat{b} = (G^\top G)^{-1} G^\top x$$

Definition 6.3 *Der Vektor*

$$\hat{b} = B(x) = (G^T G)^{-1} G^T x \quad (6.46)$$

heißt der **Regressionsvektor** und

$$\hat{q} = Q(x) = \|x - GB(x)\|^2 \quad (6.47)$$

heißt der **quadratische Restfehler**.

6.3. Lösung der Normalgleichungen mit R

Zur Lösung der Normalgleichungen (6.45) gibt es ausgefeilte numerische Verfahren, die auf dem sogenannten Q-R-Algorithmus beruhen, siehe z.B. [3].

Da wir uns in dieser Vorlesung nicht mit numerischer Mathematik, sondern mit der Anwendung auf statistische Fragestellungen beschäftigen und diese Verfahren in allen Statistik-Programmpaketen implementiert sind, befassen wir uns hier nur mit der Frage, wie man die Lösung der Normalgleichungen mit derartiger Software, speziell hier mit R berechnet.

Bezeichnet \mathbf{x} einen n -dimensionalen Vektor und \mathbf{G} eine $n \times p$ -Matrix, so wird das lineare Modell (6.41) durch die Formel

$$\mathbf{x} \sim \mathbf{G}$$

deklariert. Dabei muss man allerdings aufpassen. Da fast alle linearen Modelle in der Praxis von einem Ansatz der Form

$$g(t) = \beta_0 + \beta_1 g_1(t) + \dots + \beta_p g_p(t)$$

mit der Ansatzfunktion $g_0(t) = 1$ herrühren und damit die erste Spalte der Matrix \mathbf{G} aus lauter Einsen besteht, stellt der Formel-Interpreter automatisch eine solche Spalte vor die Matrix, wenn sie nicht bereits vorhanden ist.

Bei unserem Bremswegbeispiel wollen wir gerade das nicht, sondern wirklich den Ansatz $g(t) = \beta_1 t + \beta_2 t^2$ ohne $g_0(t) = 1$. Daher muss man R das Hinzufügen dieser Spalte verbieten. Das geht mit der Anweisung

$$\mathbf{x} \sim -1 + \mathbf{G}$$

Die Berechnung des Regressionsvektors und aller damit verbundenen statistischen Kerngrößen erfolgt dann mit der Anweisung

$$\mathbf{b} \leftarrow \text{lm}(\mathbf{x} \sim \mathbf{G}) \text{ bzw. } \mathbf{b} \leftarrow \text{lm}(\mathbf{x} \sim -1 + \mathbf{g})$$

Das Datenobjekt \mathbf{b} enthält dann alle berechneten Größen. Der Regressionsvektor ist unter $\mathbf{b}.\text{coefficients}$ gespeichert.

Ergänzend sei angemerkt, dass auch dann ein Regressionsvektor berechnet wird, wenn die Matrix G singulär ist. Welche der Lösungen der Normalgleichungen ausgewählt wird, wollen wir hier weiter verfolgen.

Deklaration mit Spalten

Sind anstelle der Matrix G die Spalten der Matrix als getrennte Größen (Vektoren) vorhanden, z.B. g_1 , g_2 und g_3 , so kann man das lineare Modell auch in der Form

$$x \sim g_1 + g_2 + g_3$$

deklarieren. Dabei ist wie oben zu beachten: Es wird automatisch der Vektor 1 hinzugefügt, der aus lauter Einsen besteht. Die obige Deklaration ist äquivalent zu

$$x \sim 1 + g_1 + g_2 + g_3$$

Will man das nicht, muss man den Vektor 3 explizit ausschließen.

$$x \sim -1 + g_1 + g_2 + g_3$$

Deklaration mit Funktionen

Beim Bremswegbeispiel lautet der Ansatz $g(t) = \beta_1 g_1(t) + \beta_2 g_2(t)$ mit $g_1(t) = t$ und $g_2(t) = t^2$. Ist \mathbf{t} der Vektor mit den Werten t_1, t_2, \dots, t_n der unabhängigen Variablen, so ist in $\mathbb{R}^{\mathbf{t}^2}$ der Vektor mit den Komponenten $t_1^2, t_2^2, \dots, t_n^2$. Dementsprechend sollte man das lineare Modell auch durch $x \sim -1 + \mathbf{t} + \mathbf{t}^2$ deklarieren können. Das stimmt leider nicht ganz, weil das \sim -Symbol bei der Deklaration von linearen Modellen eine spezielle Bedeutung besitzt. Um anzugeben, dass man den Vektor \mathbf{t}^2 meint, muss die Deklaration

$$x \sim -1 + \mathbf{t} + I(\mathbf{t}^2)$$

lauten. Diese Art der Formulierung ist für alle Operatoren möglich. Mit drei Vektoren x , u und v sind also die Deklarationen wie z.B.

$$x \sim u + v + I(1/u) + I(\exp(v)) + I(u * v)$$

möglich, was dem Ansatz

$$x = g(u, v) + e = \beta_0 + \beta_1 u + \beta_2 v + \beta_3 \frac{1}{u} + \beta_4 e^v + \beta_5 uv + e$$

entspricht.

7. Lineare Regression

In diesem Abschnitt behandeln wir das Kleinste-Quadrate-Problem aus statistischer Sicht. Wir nehmen an, dass die Messergebnisse x_1, x_2, \dots, x_n unseres statistischen Experiments *tatsächlich* einem linearen Ansatz der Form

$$x_i \approx \sum_{k=1}^p g_{ik} \beta_k$$

mit uns unbekanntem Parametern β_k folgen, dass aber stochastische Abweichungen überlagert sind:

$$x_i = \sum_{k=1}^p g_{ik} \beta_k + e_i \quad (7.48)$$

Wenn wir annehmen, dass die Fehler stochastisch unabhängig auftreten und einer Normalverteilung mit Mittelwert 0 und einer für alle Einzelexperimente gleichen Varianz σ^2 genügen, sind die Messergebnisse x_i Ergebnisse von stochastisch unabhängigen Zufallsexperimenten mit Mittelwerten $\mu_i = \sum_{k=1}^p g_{ik} \beta_k$ und Varianz σ^2 .

Der entsprechende statistische Raum besteht also aus dem Stichprobenraum $\mathcal{X} = \mathbb{R}^n$ mit Elementen $x = (x_1, x_2, \dots, x_n)^\top$ (wir fassen sie im folgenden als Spaltenvektoren auf), dem Parameterraum Θ mit Elementen $\theta = (\beta_1, \beta_2, \dots, \beta_p, \sigma^2)$ und einer Verteilungsannahme $\mathcal{P} = \{P_\theta ; \theta \in \Theta\}$ von absolutstetigen Verteilungen P_θ mit den Dichten

$$f(x; \theta) = f(x_1, \dots, x_n; \beta_1, \dots, \beta_p, \sigma^2) = f_1(x_1; \theta) \cdot f_2(x_2; \theta) \cdot \dots \cdot f_n(x_n; \theta)$$

wobei die Funktionen

$$f_i(t; \theta) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(t - \mu_i)^2}$$

die Dichten der $\mathcal{N}(\mu_i, \sigma^2)$ -Verteilungen mit den Mittelwerten $\mu_i = \sum_{k=1}^p g_{ik} \beta_k$ sind. Mit den Rechenregeln für die Exponentialfunktion ergibt sich

$$\begin{aligned} f(x; \theta) &= \left(\frac{1}{\sqrt{2\pi\sigma^2}} \right)^n e^{-\frac{1}{2\sigma^2} \sum_{i=1}^n (x_i - \mu_i)^2} \\ &= \left(\frac{1}{\sqrt{2\pi\sigma^2}} \right)^n e^{-\frac{1}{2\sigma^2} \|x - m\|^2} \end{aligned}$$

mit dem Vektor

$$m = \begin{pmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_n \end{pmatrix} = \begin{pmatrix} g_{11} & \cdots & g_{1p} \\ g_{21} & \cdots & g_{2p} \\ \vdots & & \vdots \\ g_{n1} & \cdots & g_{np} \end{pmatrix} \begin{pmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{pmatrix} = Gb$$

Wenn wir den Parametervektor jetzt noch in der Form $\theta = (b, \sigma^2)$ schreiben, erhalten wir

$$f(x; b, \sigma^2) = \left(\frac{1}{\sqrt{2\pi\sigma^2}} \right)^n e^{-\frac{1}{2\sigma^2} \|x - Gb\|^2} \quad (7.49)$$

Ein derartiger statistischer Raum wird als **lineares Modell für den Erwartungswert der Normalverteilung** bezeichnet. Die zusätzliche Voraussetzung, dass die Varianzen der Einzelexperimente sich nicht voneinander unterscheiden, heißt **Homoskedasizität**. Wir befassen uns also hier mit einem homoskedastischen linearen Modell. Dieses Modell enthält als Spezialfälle die beiden früher behandelten Fragestellungen bezüglich der Mittelwerte von Normalverteilungen. Der n -fachen Wiederholung eines Experiments mit $\mathcal{N}(\mu, \sigma^2)$ -Verteilung entspricht ein lineares Modell mit der $n \times 1$ -Matrix.

$$G = \begin{pmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{pmatrix}$$

und dem eindimensionalen Vektor $b = (\mu)$.

Der statistische Raum für den STUDENTschen t-Test zum Vergleich zweier Mittelwerte wird mit

$$G = \begin{pmatrix} 1 & 0 \\ \vdots & \vdots \\ 1 & 0 \\ 0 & 1 \\ \vdots & \vdots \\ 0 & 1 \end{pmatrix} \quad \text{und} \quad b = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}$$

erfasst.

7.1. Die Maximum-Likelihood-Schätzfunktionen

Analog zur Vorgehensweise im Abschnitt 3.1.1 bestimmen wir die Maximalstelle des Logarithmus der Likelihoodfunktion zur Dichte (7.49):

$$\begin{aligned} L^*(b, \sigma^2; x) &: = \ln f(x; b, \sigma^2) \\ &= -\frac{n}{2} \ln(2\pi) - \frac{n}{2} \ln(\sigma^2) - \frac{1}{2\sigma^2} \|x - Gb\|^2 \end{aligned}$$

Bei festgehaltener Varianz σ^2 wird diese Funktion in Abhängigkeit von b maximal, wenn $\|x - Gb\|^2$ minimal wird. Die Minimierung dieser Norm haben wir unter dem Stichwort *Methode der kleinsten Quadrate* gerade erst behandelt. Falls die $n \times p$ -Matrix G den Rang p besitzt, ist der Regressionsvektor

$$\hat{b}_x = B(x) = (G^\top G)^{-1} G^\top x \quad (7.50)$$

die eindeutig bestimmte Minimalstelle. Mit dem quadratischen Restfehler

$$\hat{q}_x = Q(x) = \|x - GB(x)\|^2 \quad (7.51)$$

wird - siehe wieder Abschnitt 3.1.1 — die Funktion

$$L^*(\hat{b}_x, \sigma^2; x) = -\frac{n}{2} \ln(2\pi) = \frac{n}{2} \ln(\sigma^2) - \frac{1}{2\sigma^2} \hat{q}_x$$

in Abhängigkeit von σ^2 bei

$$\hat{\sigma}_x^2 = \frac{1}{n} Q(x)$$

maximal. Der Maximum-Likelihood-Schätzwert ist somit $\hat{\theta}_x = (\hat{b}_x, \hat{\sigma}_x^2)$ und die Maximum-Likelihood-Schätzfunktion ist $T(x) = (B(x), \frac{1}{n} Q(x))$.

$(B(x), Q(x))$ ist eine minimal suffiziente Statistik für unsere Verteilungsannahme. Da $\hat{z} = GB(x)$ die Projektion von x auf den Untervektorraum \mathcal{L} der Spalten von G , der Vektor $z = Gb$ und damit auch $\hat{z} - z$ in \mathcal{L} liegt, ist $x - \hat{z}$ orthogonal zu $\hat{z} - z$ und nach dem Satz von Pythagoras gilt

$$\begin{aligned} \|x - Gb\|^2 &= \|x - z\|^2 = \|(x - \hat{z}) + (\hat{z} - z)\|^2 \\ &= \|x - \hat{z}\|^2 + \|\hat{z} - z\|^2 \\ &= \|x - GB(x)\|^2 + \|GB(x) - Gb\|^2 \\ &= Q(x) + \|G(B(x) - b)\|^2 \end{aligned}$$

Die Dichte (7.49) hängt also nur von den Größen $B(x)$, $Q(x)$, b und σ^2 ab.

7.2. Die Verteilung von B und Q

Zur Berechnung von Konfidenzintervallen und zur Durchführung von Tests benötigen wir Pivotvariable. Dazu müssen als Erstes die Verteilungen des Regressionsvektors und des quadratischen Restfehlers bestimmt werden.

Nach unserer Verteilungsannahme sind die Koordinationenvariablen $X_i(x) = x_i$ bezüglich der Wahrscheinlichkeit P_θ mit $\theta = (b, \sigma^2) = (\beta_1, \beta_2, \dots, \beta_p, \sigma^2)$ normalverteilt mit Mittelwerten $\mu_i = \sum_{k=1}^p g_{ik} \beta_k$ und Varianz σ^2 und außerdem stochastisch unabhängig.

Die Zufallsvariablen

$$E_i = \frac{1}{\sqrt{\sigma^2}}(X_i - \mu_i)$$

sind dann bezüglich P_θ stochastisch unabhängig und $\mathcal{N}(0, 1)$ -verteilt. Wir fassen sie zu dem n -dimensionalen Zufalls-Spaltenvektor E zusammen.

Aus $X_i = \mu_i + \sigma E_i$ mit $\sigma := \sqrt{\sigma^2}$ bzw. $x_i = \mu_i + \sigma E_i(x)$ ergibt sich in Vektordarstellung

$$\begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix} = \begin{pmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_n \end{pmatrix} + \sigma \begin{pmatrix} E_1(x) \\ E_2(x) \\ \vdots \\ E_n(x) \end{pmatrix}$$

bzw.

$$x = Gb + \sigma E(x) \tag{7.52}$$

Der quadratische Restfehler $Q(x)$ ist der minimale Wert des Abstandes $\|x - Gy\|^2$, also

$$Q(x) = \min_{y \in \mathbb{R}^p} \|x - Gy\|^2 = \min_{z \in \mathcal{L}} \|x - z\|^2$$

wobei \mathcal{L} der Unterraum des \mathbb{R}^n ist, der von den Spalten der Matrix G aufgespannt wird. Falls, wie wir angenommen haben, die Matrix G den Rang p besitzt (und $p < n$ ist), hat \mathcal{L} die Dimension p und es gibt eine Basis u_1, u_2, \dots, u_p dieses Unterraums, die aus p orthonormierten n -dimensionalen Spaltenvektoren besteht.

Diese Basis kann man durch $n - p$ weitere Spaltenvektoren $u_{p+1}, u_{p+2}, \dots, u_n$ zu einer Orthonormalbasis des \mathbb{R}^n ergänzen.

Die Matrix

$$U = \begin{pmatrix} u_1^\top \\ u_2^\top \\ \vdots \\ u_n^\top \end{pmatrix}$$

deren Zeilen aus diesen (transponierten) Vektoren gebildet werden, ist eine **Orthogonalmatrix**.

Ist z ein Vektor aus dem Unterraum \mathcal{L} , so ist z orthogonal zu $u_{p+1}, u_{p+2}, \dots, u_n$ und

$$Uz = \begin{pmatrix} u_1^\top z \\ \vdots \\ u_p^\top z \\ u_{p+1}^\top z \\ \vdots \\ u_n^\top z \end{pmatrix} = \begin{pmatrix} \tilde{z}_1 \\ \vdots \\ \tilde{z}_p \\ 0 \\ \vdots \\ 0 \end{pmatrix} =: \tilde{z}$$

Aus

$$z = U^\top U z = U^\top \tilde{z} = \tilde{z}_1 \cdot u_1 + \tilde{z}_2 \cdot u_2 + \dots + \tilde{z}_p u_p$$

ersieht man, dass die ersten p Komponenten des Vektors \tilde{z} die Entwicklungskoeffizienten des Vektors $z \in \mathcal{L}$ nach der Basis u_1, u_2, \dots, u_p sind.

Ferner gilt, da U eine Orthogonalmatrix ist:

$$\begin{aligned} \|x - z\|^2 &= \|U(x - z)\|^2 = \|Ux - Uz\|^2 \\ &= \|\tilde{x} - \tilde{z}\|^2 \\ &= \sum_{i=1}^p (\tilde{x}_i - \tilde{z}_i)^2 + \sum_{j=p+1}^n (\tilde{x}_j)^2 \end{aligned}$$

und daher

$$\begin{aligned} Q(x) &= \min_{z \in \mathcal{L}} \|x - z\|^2 \\ &= \min_{\tilde{z}_1, \dots, \tilde{z}_p} \left(\sum_{i=1}^p (\tilde{x}_i - \tilde{z}_i)^2 + \sum_{j=p+1}^n (\tilde{x}_j)^2 \right) \\ &= \sum_{j=p+1}^n (\tilde{x}_j)^2 \end{aligned}$$

wobei die Minimalstelle \hat{z} sich dadurch auszeichnet, dass ihre Entwicklungskoeffizienten $\tilde{z}_i = \tilde{x}_i$, sind. Andererseits wissen wir bereits, dass $\tilde{z} = GB(x)$ mit dem Regressionsvektor $B(x)$ aus (7.50), so dass

$$\begin{pmatrix} \tilde{x}_1 \\ \tilde{x}_2 \\ \vdots \\ \tilde{x}_p \\ 0 \\ \vdots \\ 0 \end{pmatrix} = U \hat{z} = UGB(x) \quad (7.53)$$

Mit dem Parametervektor $\theta = (b, \sigma^2)$ erhalten wir nach (7.52), dass

$$\tilde{x} = Ux = U(Gb + \sigma E(x)) = UGb + \sigma UE(x) =: \tilde{z} + \sigma \tilde{E}(x)$$

Da $Gb \in \mathcal{L}$, sind die Komponenten $\tilde{z}_{p+1} = \dots = \tilde{z}_n = 0$ und für $j = p+1, p+2, \dots, n$ gilt $\tilde{x}_j = \sigma \tilde{E}_j(x)$, wo $\tilde{E}_j(x)$ die j -te Komponente des Vektors $\tilde{E}(x) = UE(x)$ ist. Der quadratische Restfehler besitzt daher die Darstellung

$$Q(x) = \sum_{j=p+1}^n (\tilde{x}_j)^2 = \sigma^2 \sum_{j=p+1}^n \tilde{E}_j^2(x)$$

In Zufallsvariablen-Notation, also ohne das Argument x , erhalten wir somit

$$\frac{1}{\sigma^2}Q = \sum_{j=p+1}^n \tilde{E}_j^2 \quad (7.54)$$

Verteilung des quadratischen Restfehlers: Wie weiter oben angemerkt, sind die E_i bezüglich der Wahrscheinlichkeit P_θ stochastisch unabhängige $\mathcal{N}(0, 1)$ -verteilte Zufallsvariable. Da $\tilde{E} = UE$ mit einer Orthogonalmatrix U , gilt das Gleiche für die Komponenten \tilde{E}_j des Zufallsvektors \tilde{E} , so dass aus (7.54) folgt:

Satz 7.1 *Bezüglich der Wahrscheinlichkeit P_θ mit $\theta = (b, \sigma^2)$ ist die Zufallsvariable $\frac{1}{\sigma^2}Q$ Chi-Quadrat-verteilt mit $n - p$ Freiheitsgraden.*

Die χ^2 -Verteilung enthält nicht den Parameter θ , so dass daraus unmittelbar der

Satz 7.2 *Bei einem homoskedastischen linearen Modell ist*

$$Q_1(x, t) = \frac{1}{t}Q(x) = \frac{1}{t}\|x - GB(x)\|^2 \quad (7.55)$$

eine χ_{n-p}^2 -verteilte Pivotvariable für die Varianz σ^2

folgt.

Verteilung des Regressionsvektors: Nach (7.52) erhalten wir zum Parametervektor $\theta = (b, \sigma^2)$ für den Regressionsvektor die Darstellung

$$\begin{aligned} B(x) &= (G^\top G)^{-1}G^\top x \\ &= (G^\top G)^{-1}G^\top (Gb + \sigma E(x)) \\ &= (G^\top G)^{-1}G^\top Gb + \sigma(G^\top G)^{-1}G^\top E(x) \\ &= b + \sigma AE(x) \end{aligned}$$

oder

$$B = b + \sigma AE$$

mit der Matrix $A = (G^\top G)^{-1}G^\top$.

Da E bezüglich P_θ ein gaussischer Einheitsvektor und $AA^\top = (G^\top G)^{-1}G^\top G(G^\top G)^{-1} = (G^\top G)^{-1}$ ist, folgt aus dieser Darstellung

Satz 7.3 *Bezüglich der Wahrscheinlichkeit P_θ mit $\theta = (b, \sigma^2)$ ist der Regressionsvektor B ein normalverteilter Zufallsvektor mit dem Mittelwert b und der Kovarianzmatrix $\sigma^2 AA^\top = \sigma^2(G^\top G)^{-1}$.*

Multipliziert man die Gleichung (7.53) von links mit der Matrix $(G^\top G)^{-1}G^\top U^\top$, so erhält man

$$B(x) = (G^\top G)^{-1}G^\top U^\top \begin{pmatrix} \tilde{x}_1 \\ \vdots \\ \tilde{x}_p \\ 0 \\ \vdots \\ 0 \end{pmatrix} =: C \begin{pmatrix} \tilde{x}_1 \\ \vdots \\ \tilde{x}_p \\ 0 \\ \vdots \\ 0 \end{pmatrix}$$

Da $\tilde{x} = \tilde{z} + \sigma \tilde{E}(x)$, folgt weiter

$$B(x) = C\tilde{z} + \sigma C \begin{pmatrix} \tilde{E}_1 \\ \vdots \\ \tilde{E}_p \\ 0 \\ \vdots \\ 0 \end{pmatrix}$$

Der Zufallsvektor B hängt daher - was den Zufallsmechanismus anbelangt - nur von den Komponenten $\tilde{E}_1, \dots, \tilde{E}_p$ des gaussischen Einheitsvektors \tilde{E} ab, während sich der quadratische Restfehler Q aus den übrigen Komponenten $\tilde{E}_{p+1}, \dots, \tilde{E}_n$ zusammensetzt. Da die Komponenten eines gaussischen Einheitsvektors stochastisch unabhängig sind, erhalten wir als letzte Eigenschaft von B und Q noch den

Satz 7.4 *Der Regressionsvektor B und der quadratische Restfehler Q sind stochastisch unabhängige Zufallsgrößen.*

7.3. Lineare Parameterfunktionen

Im Rahmen eines linearen Modells ist man natürlich an statistischen Aussagen über die Parameter β_1, \dots, β_p interessiert.

Anstelle der einzelnen Parameter betrachten wir allgemeine Linearkombinationen dieser Parameter, d.h. Parameterfunktionen

$$\tau_a(\theta) = \tau_a(b, \sigma^2) = a^\top b = \sum_{j=1}^p a_j \beta_j \quad (7.56)$$

mit einem vorgegebenen Vektor $a^\top = (a_1, \dots, a_p)$ von Skalaren.

Eine aus dem Zusammenhang naheliegende Schätzfunktion für τ_a ist

$$T_a(x) = a^\top B(x) = \sum_{j=1}^p a_j B_j(x) \quad (7.57)$$

mit dem Regressionsvektor $B(x)$.

Als Linearkombination der Komponenten eines normalverteilten Zufallsvektors ist die Statistik T_a normalverteilt und besitzt bezüglich der Wahrscheinlichkeit P_θ den Erwartungswert

$$\mathcal{E}_\theta T_a = a^\top \mathcal{E}_\theta B = a^\top b = \tau_a(\theta)$$

T_a ist also eine für τ_a erwartungstreue Schätzfunktion.

Nach den Rechenregeln für Varianzen besitzt T_a die Varianz

$$\text{var}_\theta(T_a) = \text{var}_\theta(a^\top B) = a^\top \mathcal{C}_{B,\theta} a$$

mit der Kovarianzmatrix $\mathcal{C}_{B,\theta} = \sigma^2(G^\top G)^{-1}$ des Regressionsvektors B bezüglich P_θ . Also

$$\text{var}_\theta(T_a) = \sigma^2 \cdot a^\top (G^\top G)^{-1} a =: \sigma^2 v_a$$

Damit ist

$$H_0 = \frac{1}{\sqrt{\sigma^2 v_a}} (T_a - \tau_a(\theta)) = \frac{1}{\sqrt{\sigma^2 v_a}} (a^\top B - a^\top b)$$

eine $\mathcal{N}(0, 1)$ -verteilte Zufallsvariable und mit (7.54) erhält man

$$\begin{aligned} \sqrt{n-p} \frac{T_a - \tau_a(\theta)}{\sqrt{v_a Q}} &= \sqrt{n-p} \frac{\frac{1}{\sqrt{\sigma^2 v_a}} (T_a - \tau_a(\theta))}{\sqrt{\frac{1}{\sigma^2} Q}} \\ &= \frac{H_0}{\sqrt{\frac{1}{n-p} (\tilde{E}_{p+1}^2 + \dots + \tilde{E}_p^2)}} \end{aligned}$$

Da die Zufallsvariable H_0 als Funktion von B stochastisch unabhängig von Q ist, besagt die obige Gleichung, dass es sich bei

$$\sqrt{n-p} \frac{T_a - \tau_a(\theta)}{\sqrt{v_a \cdot Q}}$$

um eine t_{n-p} -verteilte Zufallsgröße handelt, die außer von x nur noch vom Wert $\tau_a(\theta)$ der Parameterfunktion abhängt, oder

Satz 7.5 *Die Funktion*

$$Q_a(x, t) = \sqrt{n-p} \frac{T_a(x) - t}{\sqrt{v_a Q(x)}} \tag{7.58}$$

ist eine t_{n-p} -verteilte Pivotvariable für die Parameterfunktion $\tau_a(\theta)$.

7.4. Konfidenzintervalle und Tests

Mit Hilfe der Pivotvariablen $Q_a(x, t)$ berechnet man Konfidenzintervalle und kritische Bereiche von Tests für die Parameterfunktion $\tau_a(\theta)$ gemäß den in den Abschnitten 4.5.2 bzw. 5.4 beschriebenen Vorgehensweisen, wobei die Verteilung F^Q der Pivotvariablen die t -Verteilung mit $n - p$ Freiheitsgraden und F^Q die zugehörige Verteilungsfunktion ist. Das Schema ist das Gleiche wie bei den Berechnungen für den Mittelwert der Normalverteilung.

Konfidenzbereiche

Für ein zweiseitiges Konfidenzintervall werden die Ungleichungen

$$-c \leq \sqrt{n-p} \frac{T_a(x) - t}{\sqrt{v_a Q(x)}} \leq c$$

in Ungleichungen für die Variable t umgeformt:

$$T_a(x) - c \sqrt{\frac{v_a Q(x)}{n-p}} \leq t \leq T_a(x) + c \sqrt{\frac{v_a Q(x)}{n-p}}$$

Ist γ die vorgegebene Konfidenzzahl, so ist c das $\frac{1+\gamma}{2}$ -Quantil der t_{n-p} -Verteilung, d.h. die Lösung der Gleichung

$$F^Q(c) = \frac{1+\gamma}{2}$$

Tests

Um eine einfache Nullhypothese $H_0 : \tau_a(\theta) = t_0$ gegen die Alternative $A : \tau_a(\theta) \neq t_0$ zu testen, setzt man einen Annahmebereich der Form

$$\bar{C} = \{x \in \mathbb{R}^n ; -c \leq Q_a(x, t_0) \leq c\}$$

an. Zu einer vorgegebenen Signifikanzzahl α ist c das $(1 - \frac{\alpha}{2})$ -Quantil der t_{n-p} -Verteilung, d.h. die Lösung von

$$F^Q(c) = 1 - \frac{\alpha}{2}$$

und der p -Wert ist durch

$$p(x) = 2 \left(1 - F^Q(|Q_a(x, t_0)|) \right)$$

gegeben.

7.5. Berechnung mit R

Die wichtigsten statistischen Kenngrößen für die lineare Regression werden in *R* mit dem Kommando `lm` berechnet. Konfidenzintervalle und *p*-Werte für die Parameterfunktion $\tau_a(\theta)$ erhält man aber nicht direkt. Dazu sind einige zusätzliche Schritte erforderlich. Wir erläutern die Vorgehensweise hier an dem Bremswegmodell $W \sim V + I(V^2)$. Der Datensatz ist im Anhang B.1 wiedergegeben. Zu berechnen sei ein Konfidenzintervall für den mittleren Bremsweg $w_0 = \beta_0 + \beta_1 v_0 + \beta_2 v_0^2$ bei Geschwindigkeit $v_0 = 35$ km/h, d.h. für die Parameterfunktion $\tau_a(\theta) = a^\top b$ mit $a = (1, v_0, v_0^2)^\top$.

Zur Berechnung der Größe $v_a = a^\top (G^\top G)^{-1} a$ benötigen wir die Matrix *G* für dieses Modell, die wir erhalten, wenn wir in das `lm`-Kommando zusätzlich noch das Argument `x=TRUE` einfügen.

Wir starten die *R*-Sitzung also mit

```
bw <- read.table("bw.txt", header = TRUE)
lmbw <- lm (W ~ V + I(V^2), x = TRUE)
v0 <- 35
a <- c(1, 35, 35^2)
```

Das Datenobjekt `lmbw` enthält insbesondere den Regressionsvektor $\hat{b} = B(w)$, den Residuenvektor $r = w - G\hat{b}$ und die Matrix *G*, die wir — um nachher nicht zuviel eintippen zu müssen — unter den Variablennamen `B`, `RV` und `G` speichern:

```
B <- lmbw$coefficients
RV <- lmbw$residuals
G <- lmbw$x
```

Der quadratische Restfehler wird aus `RV` berechnet:

```
Q <- sum(RV^2)
```

Falls man nur an Konfidenzintervallen für die einzelnen Parameterkomponenten β_k oder an Tests mit den einfachen Nullhypothesen $\beta_k = 0$ interessiert ist, braucht man nicht weiter zu rechnen. Man erhält sie mit den Befehlen

```
summary(lmbw)
confint(lmbw)
```

Im allgemeinen Fall müssen die Größen $T_a(w) = a^\top B(w)$ und $v_a = a^\top (G^\top G)^{-1} a$ berechnet werden. $T_a(w)$ ist das Skalarprodukt aus dem Vektor *a* und dem Regressionsvektor $B(x)$

```
Ta <- a %*% B
```

Für v_a berechnen wir den Vektor $y = (G^\top G)^{-1} a$ als Lösung des Gleichungssystems $(G^\top G)y = a$ und anschließend $v_a = a^\top y$:

```

y <- solve(t(G) %*% G, a)
va <- a %*% y

```

Schließlich benötigen wir noch das $\frac{1+\gamma}{2}$ -Quantil der t_{n-p} -Verteilung

```

n <- dim(G)[1]
p <- dim(G)[2]
gamma <- 0.95
c <- qt(0.5 * (1 + gamma), n-p)

```

Mit

```

d <- c * sqrt(va * Q / (n-p) )

```

sind die Randpunkte des Konfidenzintervalls dann

```

Ta - d ; Ta + d

```

Zur Berechnung des p -Werts $p(w)$ eines Tests mit der einfachen Nullhypothese $\tau_a(\theta) = t_0$, z.B. $t_0 = 7$, muss der Wert $Q_a(w, t_0)$ der Pivotvariablen berechnet werden:

```

t0 <- 7
qa <- sqrt(n-p) * (Ta - t0)/sqrt(va * Q)
p <- 2 * (1 - pt(abs(qa), n-p) )

```

8. Lineare Hypothesen

Im vorhergehenden Kapitel über lineare Regression gingen wir davon aus, dass die allgemeine Struktur des linearen Modells bekannt ist, d.h. die Messwerte sind von der Form $x_i = \mu_i + e_i$ mit Mittelwerten

$$\mu_i = \sum_{k=1}^p g_{ik} \beta_k \quad (8.59)$$

Dabei sind die Komponenten g_{ik} bekannt und die Gewichtungsfaktoren β_k sind die zu schätzenden Parameter.

Jetzt beschäftigen wir uns mit der Frage, wie man prüft, ob ein Ansatz für ein Modell zu den gemessenen Werten passt oder nicht. Aus statistischer Sicht bedeutet das, einen Signifikanztest mit der Nullhypothese (8.59) durchzuführen.

Beispiel 8.1 *Test eines Regressionsansatzes.* Angesichts der Messpunkte (t_i, x_i) in Abbildung 7 mit einer unabhängigen Variablen $t \in \mathbb{R}$ kann man sich zum Beispiel die Frage stellen, ob man sie durch einen Ansatz der Form $g(t) = \beta_0 + \beta_1 t + \beta_2 t^2$ ausreichend erklären kann.

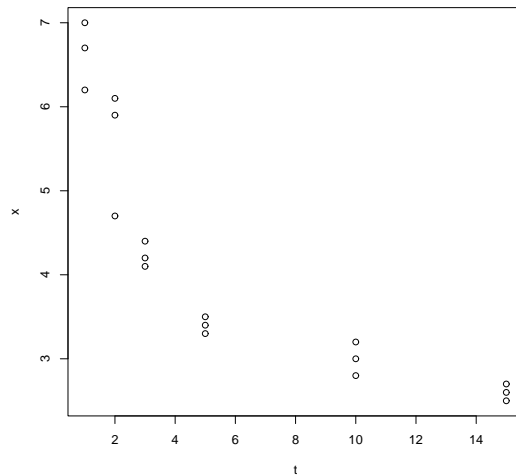


Abbildung 7: Test eines Regressionsansatzes

8.1. Die Datenstruktur

Wie man aus Abbildung 8 im Vergleich zu Abbildung 7 ersieht, reicht es für diese Fragestellung nicht aus, zu verschiedenen Werten t_i der unabhängigen Variablen jeweils nur

einen Wert x_i der abhängigen Variablen zu ermitteln, weil diese Messwerte keine Information darüber enthalten, wie groß der Abstand der zu bestimmenden Regressionskurve von den Messpunkten sein darf.

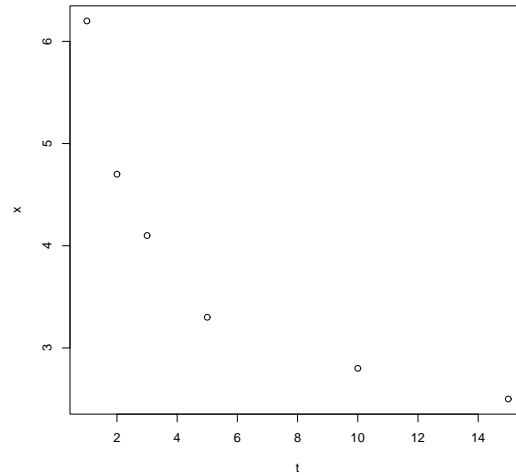


Abbildung 8: Daten ohne ausreichende Information

Führt man mit dem Wert t_i mehrere Experimente mit Ergebnissen $x_{i1}, x_{i2}, \dots, x_{in_i}$ durch, so liefert die Streuung der Werte x_{ij} einen Anhaltspunkt für diesen Abstand. Allgemein nehmen wir für dieses Kapitel an, dass r Gruppen von Messwerten,

$$\begin{array}{ccccccc}
 x_{11} & x_{12} & x_{13} & \dots & x_{1n_1} & & \\
 x_{21} & x_{22} & \dots & x_{2n_2} & & & \\
 \vdots & & & & & & \\
 x_{r1} & x_{r2} & x_{r3} & \dots & x_{rn_r} & &
 \end{array} \tag{8.60}$$

die sogenannten **Klassen**, gegeben sind, wobei die einzelnen Messwerte von der Form

$$x_{ik} = \mu_i + e_{ik} \quad k = 1, 2, \dots, n_i$$

mit einem zufällig zustande gekommenen Fehler e_{ik} und einem für alle Werte der Klasse i gleichen (unbekannten) Wert μ_i sind.

Wir fassen diese Größen zu Vektoren zusammen:

$$\begin{pmatrix} x_{11} \\ \vdots \\ x_{1n_1} \\ \text{---} \\ x_{21} \\ \vdots \\ x_{2n_2} \\ \text{---} \\ \vdots \\ \text{---} \\ x_{r1} \\ \vdots \\ x_{rn_r} \end{pmatrix} = \begin{pmatrix} \mu_1 \\ \vdots \\ \mu_1 \\ \text{---} \\ \mu_2 \\ \vdots \\ \mu_2 \\ \text{---} \\ \vdots \\ \text{---} \\ \mu_r \\ \vdots \\ \mu_r \end{pmatrix} + \begin{pmatrix} e_{11} \\ \vdots \\ e_{1n_1} \\ \text{---} \\ e_{21} \\ \vdots \\ e_{2n_2} \\ \text{---} \\ \vdots \\ \text{---} \\ e_{r1} \\ \vdots \\ e_{rn_r} \end{pmatrix} \quad (8.61)$$

oder kurz

$$x = m + e \quad (8.62)$$

mit Vektoren x, m, e aus dem \mathbb{R}^n mit $n = n_1 + n_2 + \dots + n_r$. Der Vektor m der unbekanntenen Mittelwerte ist aufgrund der Datenerhebung nicht vollkommen beliebig, sondern ein Element eines linearen Teilraums \mathcal{L} des \mathbb{R}^n mit der Dimension $r < n$.

8.2. Die Nullhypothese

Bevor wir die allgemeine Form der Nullhypothese beschreiben, leiten wir sie anhand von zwei typischen Problemstellungen her.

8.2.1. Regressionsansatz als Nullhypothese

Unter der Annahme, dass der Regressionsansatz (8.59) zutrifft, besitzt der Vektor m die spezielle Gestalt

$$\begin{pmatrix} \mu_1 \\ \vdots \\ \mu_1 \\ \hline \mu_2 \\ \vdots \\ \mu_2 \\ \hline \vdots \\ \hline \mu_r \\ \vdots \\ \mu_r \end{pmatrix} = \begin{pmatrix} g_{11} & \cdots & g_{1p} \\ \vdots & & \vdots \\ g_{11} & \cdots & g_{1p} \\ \hline g_{21} & \cdots & g_{2p} \\ \vdots & & \vdots \\ g_{21} & \cdots & g_{2p} \\ \hline \vdots & & \vdots \\ \hline g_{r1} & \cdots & g_{rp} \\ \vdots & & \vdots \\ g_{r1} & \cdots & g_{rp} \end{pmatrix} \begin{pmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{pmatrix}$$

oder kurz

$$m = Gb$$

mit einer fest vorgegebenen $n \times p$ -Matrix G und einem beliebigen p -dimensionalen Vektor b , wobei gemäß der Philosophie der linearen Modelle die Anzahl p der freien Parameter kleiner ist als die Anzahl r der Messwertklassen, die der Anzahl der unterschiedlichen Zeilen der Matrix G entspricht.

Bezeichnet $\mathcal{L}_0 = \{m = Gb ; b \in \mathbb{R}^p\}$ den von den Spalten der Matrix G aufgespannten Untervektorraum, so ist $\mathcal{L}_0 \subset \mathcal{L}$ und wir können die Nullhypothese in der Form

$$H_0 : m \in \mathcal{L}_0$$

mit der Alternative $A : m \in \mathcal{L} \setminus \mathcal{L}_0$ schreiben. Falls die Matrix G den Rang p besitzt, ist die Dimension des Raums \mathcal{L}_0 gleich p und es gilt $\mathcal{L}_0 \subset \mathcal{L} \subset \mathbb{R}^n$ mit $p < r < n$.

8.2.2. Die einfache Varianzanalyse

Zur Veranschaulichung der Problemstellung betrachten wir das folgende

Beispiel 8.2 In einem Artikel von O.H. Latter in der Zeitschrift *Biometrika* aus dem Jahr 1901 findet sich die folgende Tabelle der Längen [mm] von Kuckuckseiern, die in den Nestern dreier Vogelarten gefunden wurden:

<i>Braune Grasmücke</i>	22.0	23.9	20.9	23.8	25.0	24.0	21.7	23.8
	22.8	23.1	23.1	23.5	23.0	23.0		
<i>Rotkehlchen</i>	21.8	23.0	23.3	22.4	22.4	23.0	23.0	23.0
	23.9	22.3	22.0	22.6	22.0	22.1	21.1	23.0
<i>Zaunkönig</i>	19.8	22.1	21.5	20.9	22.0	21.0	22.3	21.0
	20.3	20.9	22.0	20.0	20.8	21.2	21.0	

Gibt es einen signifikanten Zusammenhang zwischen der Größe der Kuckuckseier und der Vogelart, in deren Nestern sie ausgebrütet werden?

Dies ist ein Beispiel für die sogenannte **einfache Varianzanalyse**: Man untersucht den Einfluss eines **Faktors** (hier die Vogelart) auf bestimmte Objekte (Kuckuckseier), wobei der Faktor r verschiedene **Stufen** oder **Levels** besitzt. In unserem Beispiel ist $r = 3$ mit den Levels 1 = Braune Grasmücke, 2 = Rotkehlchen und 3 = Zaunkönig. Für jeden Level i ermittelt man eine Klasse $x_{i1}, x_{i2}, \dots, x_{in_i}$ von Messwerten, wobei $x_{ij} = \mu_i + e_{ij}$ mit einem level- oder klassenspezifischen Mittelwert μ_i angenommen wird. Der Aussage, dass der Faktor keinen Einfluss besitzt, entspricht der Hypothese

$$H_0 : \mu_1 = \mu_2 = \dots = \mu_r = \mu,$$

die man mit den oben eingeführten Bezeichnungen auch in der Form

$$m \in \mathcal{L}_0 = \left\{ \begin{pmatrix} \mu \\ \mu \\ \vdots \\ \mu \end{pmatrix} = \begin{pmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{pmatrix} \mu ; \mu \in \mathbb{R} \right\}$$

mit $\dim(\mathcal{L}_0) = 1$ schreiben kann.

8.2.3. Lineare Hypothesen

Ganz allgemein können wir unser Testproblem wie folgt formulieren:

Gegeben sind ein Ergebnisvektor $x \in \mathbb{R}^n$ und zwei Untervektorräume $\mathcal{L}_0 \subset \mathcal{L} \subset \mathbb{R}^n$ mit den Dimensionen $p = \dim(\mathcal{L}_0)$ und $r = \dim(\mathcal{L})$, wobei $p < r < n$.

Es ist

$$x = m + e$$

mit einem Mittelwertvektor m und einem „Fehler“-Vektor e , wobei lediglich bekannt ist, dass $m \in \mathcal{L}$.

Es ist zu prüfen, ob die Nullhypothese

$$H_0 : m \in \mathcal{L}_0$$

angenommen oder widerlegt werden kann.

Dazu benötigt man einen Schätzwert $T(x)$ für den Abstand

$$d(m, \mathcal{L}_0) = \min_{z \in \mathcal{L}_0} \|m - z\|$$

des (unbekannten) Vektors m vom Raum \mathcal{L}_0 auf der Basis des Ergebnisses x , anhand dessen entschieden wird, ob $d(m, \mathcal{L}_0) = 0$ oder $d(m, \mathcal{L}_0) > 0$.

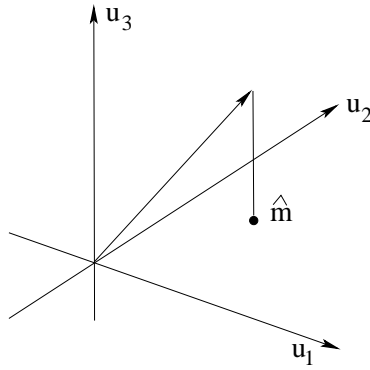


Abbildung 9: Schätzwert für m

8.3. Schätzung des Abstands

In diesem Abschnitt wird ein statistisch „vernünftiges“ Maß $T(x)$ für den Abstand des Vektors $m \in \mathcal{L}$ vom Teilraum $\mathcal{L}_0 \subset \mathcal{L}$ hergeleitet, das auf der alleinigen Kenntnis des Ergebnisvektors x beruht.

Zur Veranschaulichung betrachten wir den Fall $n = 3$, $r = 2$ und $p = 1$. In dem in Abbildung 9 dargestellten kartesischen Koordinatensystem mit den Basisvektoren u_1, u_2 und u_3 sei \mathcal{L} die Ebene, die von u_1 und u_2 aufgespannt wird und \mathcal{L}_0 die Gerade, in der der Vektor u_1 liegt.

Da m unbekannt ist, benötigt man dafür einen Schätzwert. Wie bei der Methode der kleinsten Quadrate wählen wir dafür wieder den Vektor \hat{m} aus dem Raum \mathcal{L} , der von x den geringsten Abstand besitzt, d.h. die Projektion des Vektors x auf den Raum \mathcal{L} , der durch $\|x - \hat{m}\|^2 = \min_{z \in \mathcal{L}} \|x - z\|^2$ bzw. die Eigenschaft charakterisiert ist, dass der Differenzvektor $x - \hat{m}$ orthogonal zum Raum \mathcal{L} ist.

Den Abstand des Vektors \hat{m} vom Teilraum \mathcal{L}_0 wiederum erhält man dadurch, dass man die Projektion $\hat{\hat{m}}$ von \hat{m} auf \mathcal{L}_0 berechnet:

$$\|\hat{m} - \hat{\hat{m}}\|^2 = \min_{z \in \mathcal{L}_0} \|\hat{m} - z\|^2$$

Dieser Abstand liefert aber keine zuverlässige Aussage darüber, ob der Mittelwertvektor m einen positiven Abstand von \mathcal{L}_0 besitzt, denn in der Differenz $\hat{m} - \hat{\hat{m}}$ sind Anteile enthalten, die von den Zufallsfehlern e_{ik} herrühren.

Um die Größenordnung dieser Anteile abzuschätzen, benötigt man Informationen über die Größenordnung der Zufallsfehler. Die erhält man, wie man sich an der Abbildung 9 veranschaulicht, aus dem Differenzvektor $x - \hat{m}$, der orthogonal zu \mathcal{L} ist und somit allein durch die Zufallsfehler bestimmt ist.

Geht man davon aus, dass die Größenordnung der Zufallsfehler in jeder Koordinatenrichtung in etwa gleich ist, so dürfte der unbekannte Vektor m innerhalb eines Kreises um \hat{m} in der u_1 - u_2 -Ebene mit dem Radius $\|x - \hat{m}\|$ liegen.

Die Abbildungen 10 und 11 zeigen auf, wie man unter dieser Annahme argumentieren

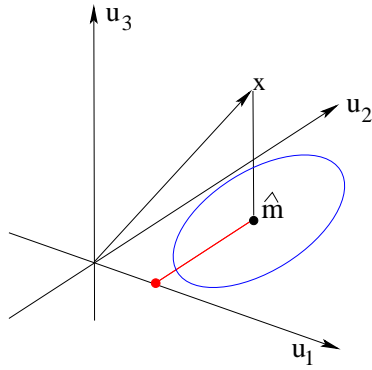


Abbildung 10: m nicht in \mathcal{L}_0

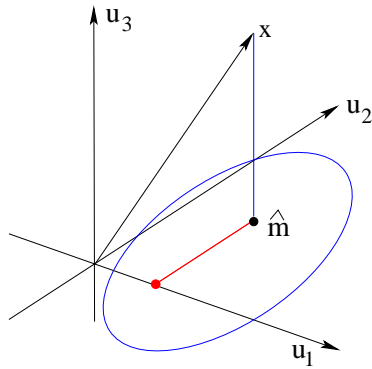


Abbildung 11: m aus \mathcal{L}_0 ist nicht auszuschließen

kann: Liegt die Gerade \mathcal{L}_0 ganz ausserhalb dieses Kreises, so kann m nicht in \mathcal{L}_0 liegen. Durchschneidet diese Gerade den Kreis, dann ist nicht auszuschließen, dass m in \mathcal{L}_0 liegt. Für ein Abstandsmaß $T(x)$ ergibt sich daraus, dass man den geschätzten Abstand $\|\hat{m} - \hat{m}\|$ relativ zur Größenordnung $\|x - \hat{m}\|$ der Zufallsfehler sehen bzw. einen Quotienten

$$\frac{\|\hat{m} - \hat{m}\|^2}{\|x - \hat{m}\|^2}$$

verwenden sollte.

Dies ist in dieser Form allerdings nur für den Fall $p = 1, r = 2, n = 3$ korrekt.

Im allgemeinen Fall muss man berücksichtigen, dass $x - \hat{m}$ in einem Teilraum der Dimension $n - r$ und $\hat{m} - \hat{m}$ in einem der Dimension $r - p$ liegt, so dass der Zähler des obigen Quotienten das Normquadrat eines $(r - p)$ -dimensionalen Vektors, d.h. eine Summe aus $r - p$ Summanden darstellt und der Nenner $n - r$ Summanden besitzt.

Um diese Abstände vergleichbar zu machen, muss man von den Summen zu Mittelwerten übergehen, d.h. die Summen durch die Anzahl der jeweiligen Summanden dividieren. Als

Ansatz für ein Abstandsmaß erhalten wir somit

$$T(x) = \frac{\frac{1}{r-p} \|\hat{m}_l - \hat{m}\|^2}{\frac{1}{n-r} \|x - \hat{m}\|^2} \quad (8.63)$$

8.4. Berechnung des Abstandsmaßes

Zur Berechnung des Nenners des Abstandsmaßes $T(x)$ muss ein Minimierungsproblem gelöst werden.

$$\|x - \hat{m}\|^2 = \min_{z \in \mathcal{L}} \|x - z\|^2$$

Haben die Vektoren $z \in \mathcal{L}$ die in (8.61) dargestellte Struktur, so ist

$$\|x - z\|^2 = \sum_{i=1}^r \sum_{k=1}^{n_i} (x_{ik} - \mu_i)^2$$

und es müssen diejenigen $\mu_1, \mu_2, \dots, \mu_r$ bestimmt werden, die diese Summe minimieren.

$$\begin{aligned} & \min_{z \in \mathcal{L}} \|x - z\|^2 \\ &= \min_{\mu_1, \mu_2, \dots, \mu_r} \sum_{i=1}^r \sum_{k=1}^{n_i} (x_{ik} - \mu_i)^2 \\ &= \sum_{i=1}^r \min_{\mu_i} \sum_{k=1}^{n_i} (x_{ik} - \mu_i)^2 \end{aligned}$$

Wie von der Berechnung des Maximum-Likelihood-Schätzwerts für den Mittelwert der Normalverteilung her bekannt, wird das Minimum einer Quadratsumme

$$Q(t) = \sum_{j=1}^m (y_j - t)^2$$

beim Mittelwert

$$\bar{y} = \frac{1}{m} (y_1 + \dots + y_m)$$

angenommen. Im vorliegenden Fall ist also

$$\mu_i = \bar{x}_{i\cdot} := \frac{1}{n_i} (x_{i1} + x_{i2} + \dots + x_{in_i}) \quad (8.64)$$

zu wählen und es ist

$$\|x - \hat{m}\|^2 = \sum_{i=1}^r \sum_{k=1}^{n_i} (x_{ik} - \bar{x}_{i\cdot})^2 \quad (8.65)$$

Zur Berechnung des Zählers von $T(x)$ verwendet man die Tatsache, dass $x - \hat{m}$ orthogonal zu allen Vektoren aus \mathcal{L} ist und daher der Satz von Pythagoras benutzt werden kann. Für $z \in \mathcal{L}_0 \subset \mathcal{L}$ ist $\hat{m} - z \in \mathcal{L}$ und damit

$$\begin{aligned}\|x - z\|^2 &= \|(x - \hat{m}) + (\hat{m} - z)\|^2 \\ &= \|x - \hat{m}\|^2 + \|\hat{m} - z\|^2\end{aligned}$$

Daraus folgt

$$\begin{aligned}\min_{z \in \mathcal{L}_0} \|x - z\|^2 &= \min_{z \in \mathcal{L}_0} (\|x - \hat{m}\|^2 + \|\hat{m} - z\|^2) \\ &= \|x - \hat{m}\|^2 + \min_{z \in \mathcal{L}_0} \|\hat{m} - z\|^2 \\ &= \|x - \hat{m}\|^2 + \|\hat{m} - \hat{m}\|^2\end{aligned}$$

Dies bedeutet einmal, dass der Vektor \hat{m} die Projektion von x auf den Teilraum \mathcal{L}_0 ist

$$\min_{z \in \mathcal{L}_0} \|x - z\|^2 = \|x - \hat{m}\|^2$$

und zum anderen, dass

$$\|\hat{m} - \hat{m}\|^2 = \|x - \hat{m}\|^2 - \|x - \hat{m}\|^2$$

Mit den in der Literatur üblichen Bezeichnungen

$$\gamma^2(x) = \|x - \hat{m}\|^2 = \sum_{i=1}^r \sum_{k=1}^{n_i} (x_{ik} - \bar{x}_{i\cdot})^2 \quad (8.66)$$

und

$$\gamma_0^2(x) = \|x - \hat{m}\|^2 = \min_{z \in \mathcal{L}_0} \|x - z\|^2 \quad (8.67)$$

erhält man für das Abstandsmaß die Rechenformel

$$T(x) = \frac{\frac{1}{r-p} (\gamma_0^2(x) - \gamma^2(x))}{\frac{1}{n-r} \gamma^2(x)} \quad (8.68)$$

Die Größe $\gamma^2(x)$ heißt die **Intraklassenvarianz**, in der englischsprachigen Literatur „**Sum of squares within subjects (SSW)**“. Sie ist bis auf die Vorfaktoren die Summe der Stichprobenvarianzen der Messwertklassen und stellt somit ein Maß für die Streuung der Messwerte *innerhalb* der einzelnen Klassen um den Klassenmittelwert dar.

Explizite Formeln für die Größe $\gamma_0^2(x)$ hängen natürlich von der Gestalt des Teilraums \mathcal{L}_0 ab. Wir werden sich in einem späteren Abschnitt für einige typische Fragestellungen herleiten.

8.5. Statistische Eigenschaften

Zur Herleitung der Verteilung der Statistik T gehen wir von der allgemeinen Situation aus, dass ein Ergebnisvektor

$$x = (x_1, x_2, \dots, x_n)^\top \in \mathbb{R}^n$$

vorliegt, der von der Form

$$x = m + e$$

mit einem Mittelwertvektor m und einem Fehlervektor e ist.

Über m ist bekannt, dass es in einem Teilraum $\mathcal{L} \subset \mathbb{R}^n$ der Dimension $r < n$ liegt.

Zu prüfen ist die Hypothese, dass m in einem kleineren Teilraum $\mathcal{L}_0 \subset \mathcal{L}$ der Dimension $p < r$ liegt.

Über den Fehlervektor

$$e = (e_1, e_2, \dots, e_n)^\top$$

wird angenommen, dass seine Komponenten Werte von stochastisch unabhängigen $\mathcal{N}(0, \sigma^2)$ -verteilten Fehlervariablen sind, die alle den Mittelwert 0 und die — unbekannte, aber für alle Komponenten gleiche — Varianz σ^2 besitzen.

Der entsprechende statistische Raum setzt sich daher zusammen aus dem \mathbb{R}^n als Stichprobenraum, der Menge Θ der Parameter $\theta = (m, \sigma^2)$ mit $m \in \mathcal{L}$ und $\sigma^2 > 0$ sowie der Verteilungsannahme \mathcal{P} von Wahrscheinlichkeiten P_θ , die dadurch charakterisiert ist, dass Koordinatenvariablen X_i bezüglich der Verteilung $P_{(m, \sigma^2)}$ eine $\mathcal{N}(\mu_i, \sigma^2)$ -Verteilung besitzen, wobei μ_i die i -te Komponente des Vektors m ist, und außerdem stochastisch unabhängig sind.

Im folgenden seien $\theta = (m, \sigma^2)$ mit $m = (\mu_1, \mu_2, \dots, \mu_n)^\top$ fest und die Statistiken als Zufallsgrößen auf dem Wahrscheinlichkeitsraum $(\mathbb{R}^n, \mathcal{B}_n, P_{(m, \sigma^2)})$ aufzufassen.

Die Koordinatenvariablen $X_i(x) = x_i$ sind $\mathcal{N}(\mu_i, \sigma^2)$ -verteilt, so dass die

$$E_i(x) = \frac{1}{\sigma} (X_i(x) - \mu_i) = \frac{1}{\sigma} (x_i - \mu_i)$$

mit $\sigma := \sqrt{\sigma^2}$ eine $\mathcal{N}(0, 1)$ -Verteilung besitzen. Der Ergebnisvektor x hat somit die Form

$$\begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix} = \begin{pmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_n \end{pmatrix} + \sigma \begin{pmatrix} E_1(x) \\ E_2(x) \\ \vdots \\ E_n(x) \end{pmatrix}$$

oder kurz

$$x = m + \sigma E(x),$$

wobei $E(x)$ auf $(\mathbb{R}^n, \mathcal{B}_n, P_{(m, \sigma^2)})$ ein gaussischer Einheitsvektor ist.

Zur Herleitung der Verteilung der Statistik $T(x)$ nehmen wir eine Basis u_1, u_2, \dots, u_n von orthonormierten Vektoren des \mathbb{R}^n als gegeben an, die die folgenden Eigenschaften besitzen:

- Die ersten p dieser Vektoren — u_1, u_2, \dots, u_p — bilden eine Basis des Teilraums \mathcal{L}_0
- Die ersten r dieser Vektoren — u_1, u_2, \dots, u_r — bilden eine Basis des Teilraums \mathcal{L} .

Dann sind die Vektoren $u_{r+1}, u_{r+2}, \dots, u_n$ orthogonal zu allen Vektoren $z \in \mathcal{L}$ und $u_{p+1}, u_{p+2}, \dots, u_n$ orthogonal zu den $z \in \mathcal{L}_0$.

Für die Orthogonalmatrix

$$U = \begin{pmatrix} u_1^\top \\ u_2^\top \\ \vdots \\ u_n^\top \end{pmatrix}$$

mit den u_j als Zeilen gilt dann

1. Ist $z \in \mathcal{L}$, so ist

$$\tilde{z} = Uz = \begin{pmatrix} u_1^\top z \\ \vdots \\ u_r^\top z \\ \hline u_{r+1}^\top z \\ \vdots \\ u_n^\top z \end{pmatrix} = \begin{pmatrix} \tilde{z}_1 \\ \vdots \\ \tilde{z}_r \\ \hline 0 \\ \vdots \\ 0 \end{pmatrix}$$

Die \tilde{z}_j sind dabei die Entwicklungskoeffizienten des Vektors z nach der Basis u_1, \dots, u_r .

2. Ist $z \in \mathcal{L}_0$, so ist

$$\tilde{z} = Uz = \begin{pmatrix} u_1^\top z \\ \vdots \\ u_p^\top z \\ \hline u_{p+1}^\top z \\ \vdots \\ u_r^\top z \\ \hline u_{r+1}^\top z \\ \vdots \\ u_n^\top z \end{pmatrix} = \begin{pmatrix} \tilde{z}_1 \\ \vdots \\ \tilde{z}_p \\ \hline 0 \\ \vdots \\ 0 \\ \hline 0 \\ \vdots \\ 0 \end{pmatrix}$$

wobei die \tilde{z}_j die Entwicklungskoeffizienten des Vektors z nach der Basis u_1, \dots, u_p sind. Da U eine Orthogonalmatrix ist, erhält sie die euklidische Norm. Es ist

$$\|x - z\|^2 = \|U(x - z)\|^2 = \|Ux - Uz\|^2 =: \|\tilde{x} - \tilde{z}\|^2$$

Ist $z \in \mathcal{L}$, so ist

$$\|x - z\|^2 = \|\tilde{x} - \tilde{z}\|^2 = \sum_{i=1}^r (\tilde{x}_i - \tilde{z}_i)^2 + \sum_{k=r+1}^n (\tilde{x}_k)^2$$

und

$$\begin{aligned} \gamma^2(x) &= \min_{z \in \mathcal{L}} \|x - z\|^2 \\ &= \min_{\tilde{z}_1, \dots, \tilde{z}_r} \left(\sum_{i=1}^r (\tilde{x}_i - \tilde{z}_i)^2 + \sum_{k=r+1}^n (\tilde{x}_k)^2 \right) \\ &= \sum_{k=r+1}^n (\tilde{x}_k)^2 \end{aligned}$$

Denn die Quadratsumme $\sum_{i=1}^r (\tilde{x}_i - \tilde{z}_i)^2$ wird genau dann minimal, wenn man die \tilde{z}_i gleich den \tilde{x}_i setzt.

Für $z \in \mathcal{L}_0$ ist

$$\|x - z\|^2 = \sum_{i=1}^p (\tilde{x}_i - \tilde{z}_i)^2 + \sum_{j=p+1}^r (\tilde{x}_j)^2 + \sum_{k=r+1}^n (\tilde{x}_k)^2$$

Daraus ergibt sich auf die entsprechende Weise

$$\begin{aligned} \gamma_0^2(x) &= \min_{z \in \mathcal{L}_0} \|x - z\|^2 \\ &= \sum_{j=p+1}^r (\tilde{x}_j)^2 + \sum_{k=r+1}^n (\tilde{x}_k)^2 \end{aligned}$$

Für unser Abstandsmaß $T(x)$ folgt daraus

$$\begin{aligned} T(x) &= \frac{\frac{1}{r-p} \left(\gamma_0^2(x) - \gamma^2(x) \right)}{\frac{1}{n-r} \gamma^2(x)} = \\ &= \frac{\frac{1}{r-p} \sum_{j=p+1}^r (\tilde{x}_j)^2}{\frac{1}{n-r} \sum_{k=r+1}^n (\tilde{x}_k)^2} \end{aligned}$$

Gemäß der Verteilungsannahme besitzt der Ergebnisvektor die Darstellung $x = m + \sigma E(x)$ mit $m \in \mathcal{L}$, $\sigma = \sqrt{\sigma^2}$ und einem gaussischen Einheitsvektor $E(x)$. Dann ist

$$\tilde{x} = Ux = Um + \sigma UE(x) = \tilde{m} + \sigma \tilde{E}(x)$$

wobei die Größen \tilde{m} und $\tilde{E}(x)$ die folgenden Eigenschaften besitzen.

1. Da U eine Orthogonalmatrix, ist auch $\tilde{E}(x)$ ein gaussischer Einheitsvektor, d.h. seine Komponenten $\tilde{E}_i(x)$ sind stochastisch unabhängige $\mathcal{N}(0, 1)$ -verteilte Zufallsvariable.

2. m liegt im Teilraum \mathcal{L} , so dass für die Komponenten des Vektors $\tilde{m} = (\tilde{\mu}_1, \tilde{\mu}_2, \dots, \tilde{\mu}_n)^\top$ gilt.

$$\tilde{\mu}_k = 0 \text{ für } k > r$$

Die Komponenten des Vektors \tilde{x} besitzen daher die Gestalt

$$\tilde{x}_i = \begin{cases} \tilde{\mu}_i + \sigma \tilde{E}_i(x) & \text{für } i = 1, 2, \dots, r \\ \sigma \tilde{E}_i(x) & \text{für } i = r + 1, \dots, n \end{cases}$$

und für das Abstandsmaß $T(x)$ gilt

$$T(x) = \frac{\frac{1}{r-p} \sum_{j=p+1}^r \left(\tilde{\mu}_j + \sigma \tilde{E}_j(x) \right)^2}{\frac{1}{n-r} \sigma^2 \sum_{k=r+1}^n \left(\tilde{E}_k(x) \right)^2}$$

Liegt der Vektor m insbesondere im Teilraum \mathcal{L}_0 , so ist $\tilde{\mu}_j = 0$ auch für $j = p + 1, \dots, r$ und es ist

$$T(x) = \frac{\frac{1}{r-p} \sum_{j=p+1}^r \left(\tilde{E}_j(x) \right)^2}{\frac{1}{n-r} \sum_{k=r+1}^n \left(\tilde{E}_k(x) \right)^2}$$

Die rechte Seite enthält keine Parameteranteile, sondern ist lediglich eine Funktion der stochastisch unabhängigen $\mathcal{N}(0, 1)$ -verteilten Zufallsvariablen \tilde{E}_j . Aus der Struktur der Formel ersieht man, dass die Verteilung dieser Zufallsvariable die F-Verteilung mit $(r - p, n - r)$ Freiheitsgraden ist.

8.6. Kritischer Bereich

Entsprechend der Interpretation der Statistik (8.68) als Abstandsmaß wird der kritische Bereich als Menge der Form

$$\mathcal{C} = \{x ; T(x) > c\}$$

angesetzt. Für die Parameter $\theta = (m, \sigma^2)$ aus der **Nullhypothese** ist die Verteilung der Statistik T bekannt. Für diese θ gilt

$$P_\theta(\mathcal{C}) = P_\theta(T > c) = P^T(c, \infty) = 1 - F^T(c)$$

mit der $f_{(r-p, n-r)}$ -Verteilung P^T bzw. der zugehörigen Verteilungsfunktion F^T .

Zu einer vorgegebenen Signifikanzzahl α ist daher c so zu wählen, dass $F^T(c) = 1 - \alpha$ und der p -Wert berechnet sich zu

$$p(x) = 1 - F^T(T(x))$$

8.7. Die einfache Varianzanalyse

Wie in Abschnitt 8.2.2 dargestellt, ist die Nullhypothese bei der einfachen Varianzanalyse durch den Untervektorraum

$$\mathcal{L}_0 = \left\{ z = \begin{pmatrix} \mu \\ \mu \\ \vdots \\ \mu \end{pmatrix} ; \mu \in \mathbb{R} \right\}$$

mit der Dimension $p = 1$ gegeben. Zur Berechnung des Zählerterms $\|\hat{m} - \widehat{\hat{m}}\|^2$ der Statistik $T(x)$ kennt man bereits aus Abschnitt 8.4 den Vektor \hat{m} :

$$\hat{m} = \underbrace{(\bar{x}_{1\cdot}, \dots, \bar{x}_{1\cdot})}_{n_1\text{-mal}} \underbrace{(\bar{x}_{2\cdot}, \dots, \bar{x}_{2\cdot})}_{n_2\text{-mal}} \underbrace{(\bar{x}_{r\cdot}, \dots, \bar{x}_{r\cdot})}_{n_r\text{-mal}}^\top$$

Für $z \in \mathcal{L}_0$ gilt somit

$$\|\hat{m} - z\|^2 = \sum_{i=1}^r \sum_{k=1}^{n_i} (\bar{x}_{i\cdot} - \mu)^2 = \sum_{i=1}^r n_i (\bar{x}_{i\cdot} - \mu)^2 =: g(\mu)$$

Die Minimalstelle der Funktion $g(\mu)$ erhält man aus $g'(\mu) = -2 \sum_{i=1}^r n_i (\bar{x}_{i\cdot} - \mu) = 0$ bzw.

$$\sum_{i=1}^r n_i \bar{x}_{i\cdot} = \left(\sum_{i=1}^r n_i \right) \mu$$

Da $\sum_{i=1}^r n_i = n$ und $\sum_{i=1}^r n_i \bar{x}_{i\cdot} = \sum_{i=1}^r \sum_{k=1}^{n_i} x_{ik}$, ist die Minimalstelle gleich

$$\mu = \frac{1}{n} \sum_{i=1}^r \sum_{k=1}^{n_i} x_{ik} =: \bar{x}_{\cdot\cdot},$$

d.h. gleich dem arithmetischen Mittel aller Messwerte. Es ist daher

$$\begin{aligned}
\|\widehat{m} - \widehat{m}\|^2 &= \min_{z \in \mathcal{L}_0} \|\widehat{m} - z\|^2 \\
&= \min_{\mu \in \mathbb{R}} \sum_{i=1}^r n_i (\bar{x}_{i.} - \mu)^2 \\
&= \sum_{i=1}^r n_i (\bar{x}_{i.} - \bar{x}_{..})^2
\end{aligned}$$

Diese Größe heißt die **Interklassenvarianz**. Sie ist ein Maß dafür, wie stark die Klassenmittelwerte $\bar{x}_{i.}$ um den gemeinsamen Mittelwert streuen, d.h. wie unterschiedlich sie sind.

Die Statistik $T(x)$ hat damit die Gestalt

$$T(x) = \frac{\frac{1}{r-1} \sum_{i=1}^r n_i (\bar{x}_{i.} - \bar{x}_{..})^2}{\frac{1}{n-r} \sum_{i=1}^r \sum_{k=1}^{n_i} (x_{ik} - \bar{x}_{i.})^2} = \frac{\frac{1}{r-1} \text{Interklassenvarianz}}{\frac{1}{n-r} \text{Intraklassenvarianz}}$$

Diese Formel erklärt auch, weshalb die Methode die einfache *Varianz*-Analyse heißt, obwohl Aussagen über Mittelwerte gemacht werden. Nicht die theoretischen Varianzen sind gemeint, sondern Stichprobenvarianzen als Maß für die Unterschiedlichkeit der Mittelwerte.

Berechnung mit R

Zur Darstellung der Vorgehensweise verwenden wir wieder das Beispiel 8.2. Der Datensatz für R ist im Anhang B.2 wiedergegeben. Stehen in einer Spalte eines „data frame“, wie hier in der Spalte `Art`, keine Zahlen sondern Namen, die mit einem Buchstaben beginnen, so fasst R diese als die Levels eines Faktors auf. In unserem Beispiel haben wir also einen Faktor `Art` mit den Levels `B`, `R` und `Z`. Hat man die Levels im Datensatz mit Ziffern wie 1, 2, ... gekennzeichnet, so muss diese Spalte beim Einlesen mit `read.table` als Faktor deklariert werden. Wie das geht, entnimmt man der Online-Hilfe von R .

Nach dem Einlesen des Datensatzes mit

```
kk <- read.table("kk.txt", header = TRUE)
```

enthält die Spalte `Lge` die gemessenen Kuckuckseierlängen und die Faktor-Spalte `Art` die zugehörigen Vogelarten. Die Abhängigkeit der Länge von der Vogelart wird durch das Modell $Lge \sim Art$ beschrieben und die einfache Varianzanalyse wird mit dem Befehl

```
kkav <- aov(Lge ~ Art, data = kk)
```

durchgeführt. Das Ergebnis erhält man mit

```
summary(kkav)
```

In unserem Beispiel erhält man

```
          Df  Sum Sq Mean Sq F value    Pr(>F)
Art          2  31.1119  15.5560  22.329 2.479e-07 ***
Residuals   42  29.2605   0.6967
---
```

```
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Die erste Zeile enthält die Werte des Zählers und die zweite die des Nenners der Statistik $T(x)$. Die Spalte **Df** gibt die Anzahl der Freiheitsgrade ($r-1$ bzw. $n-r$) an, die Spalte **Sum Sq** die Werte der Inter- bzw. Intraklassenvarianz und die Spalte **Mean Sq** die durch die jeweiligen Freiheitsgrade dividierten Varianzen. **F value** ist der Wert $T(x)$ und **Pr(>F)** der p -Wert des Tests.

Da in unserem Beispiel sogar noch zu einem Signifikanzniveau von $\alpha = 0,001$ signifikante Unterschiede festgestellt wurden, möchte man natürlich durch einen paarweisen Vergleich untersuchen, wo die Unterschiede liegen.

Eine Möglichkeit dazu ist die Methode der „Honest Significant Differences“ von Tukey:

```
TukeyHSD(kkav)
```

mit dem Ergebnis

```
Tukey multiple comparisons of means
 95% family-wise confidence level
```

```
Fit: aov(formula = Lge ~ Art, data = kk)
```

```
$Art
      diff      lwr      upr    p adj
R-B -0.5580357 -1.300147  0.1840751 0.1732815
Z-B -1.9942857 -2.747852 -1.2407194 0.0000003
Z-R -1.4362500 -2.165048 -0.7074517 0.0000619
```

Die Spalte **diff** enthält die paarweisen Differenzen der Klassenmittelwerte, **lwr** und **upr** die unteren und oberen Grenzen der Konfidenzintervalle für diese Differenzen und **p adj** die *korrigierten* p -Werte der Tests auf die Nullhypothese, dass die Differenz gleich Null ist. Eine Korrektur ist deshalb erforderlich, weil mehrere Konfidenzintervalle bzw. Tests gleichzeitig an einem Datensatz durchgeführt werden. Auf Einzelheiten zu diesem Problem wollen wir hier nicht eingehen und verweisen stattdessen auf [10].

8.8. Test eines Regressionsansatzes

Um zu testen, ob die Nullhypothese

$$H_0 : \mu_i = \sum_{k=1}^p g_{ik} \beta_k \quad \text{für } i = 1, 2, \dots, r$$

bzw.

$$H_0 : m \in \mathcal{L}_0 = \{z = Gb ; b \in \mathbb{R}^p\}$$

für die Ergebnisse

$$x_{ij} = \mu_i + e_{ij} \quad \text{für } i = 1, 2, \dots \quad \text{und } j = 1, 2, \dots, n_i$$

angenommen werden kann, berechnet man $T(x)$ gemäß der Formel (8.68):

$$T(x) = \frac{\frac{1}{r-p}(\gamma_0^2(x) - \gamma^2(x))}{\frac{1}{n-r}\gamma^2(x)}$$

mit der Intraklassenvarianz

$$\gamma^2(x) = \|x - \hat{m}\|^2 = \sum_{i=1}^r \sum_{k=1}^{n_i} (x_{ik} - \bar{x}_{i.})^2$$

und

$$\gamma_0^2(x) = \min_{z \in \mathcal{L}_0} \|x - z\|^2 = \min_{b \in \mathbb{R}^p} \|x - Gb\|^2 = \hat{q}$$

d.h dem quadratischen Restfehler \hat{q} des Regressionsansatzes.

Berechnung mit R

In R gibt es (meines Wissens) keine vorgefertigte Prozedur für diese Problemstellung. Man muss also $\gamma^2(x)$, $\gamma_0^2(x)$ und daraus $T(x)$ explizit berechnen und anschließend den p-Wert ermitteln. Zur Demonstration der Vorgehensweise verwenden wir Beispiel 8.1 mit dem im Anhang B.3 abgedruckten Datensatz. Die Nullhypothese lautet, dass sich die Punkte durch ein Polynom zweiten Grades approximieren lassen.

```
a10 <- read.table("a10.txt",header=TRUE)
attach(a10)
n <- length(x)
```

Zur Berechnung der Intraklassenvarianz kann man den Befehl für die einfache Varianzanalyse verwenden, wenn man einen Faktor zur Verfügung hat, der die Messwertklassen unterscheidet, die durch gleichen Wert der unabhängigen Variablen t charakterisiert sind. Dazu wandeln wir einfach die Spalte t zu einem Faktor mit ganzzahligen Levels um:

```
tfac <- factor(t)
r <- length(levels(tfac))
```

Mit

```
a10.aov <- aov(x ~ tfac)
gamma <- sum(residuals(a10.aov)^2)
```

berechnet man die Intraklassenvarianz.

Zur Berechnung von $\gamma_0^2(x)$ wenden wir den Regressionsansatz der Nullhypothese auf \mathbf{x} an und berechnen den quadratischen Restfehler:

```
a10.lm <- lm(x ~ t + I(t^2))
p <- 3
gamma0 <- sum(residuals(a10.lm)^2)
```

Dann setzen wir $T(x)$ aus den berechneten Größen zusammen:

```
zaehler <- (gamma0 - gamma)/(r - p)
nenner <- gamma/(n - r)
Tx <- zaehler/nenner
```

Unter der Nullhypothese ist $T(x)$ F-verteilt mit $(r-p, n-r)$ Freiheitsgraden. Den p-Wert des Tests erhält man daher mit der Anweisung

```
px <- 1.0 - pf(Tx, r-p, n-r)
```

In unserem Beispiel erhält man $p(x) = 0.002990162$. Bezüglich aller gängigen Signifikanzzahlen ist die Nullhypothese also widerlegt. Ein Regressionsansatz der Form $g(t) = \beta_0 + \beta_1 t + \beta_2 t^2$ passt also nicht zu den Messergebnissen, was auch optisch offensichtlich wird, wenn man die Regressionskurve gegen die Messwerte plottet (Abbildung 12, linkes Bild).

Wendet man einen Regressionsansatz $g(t) = \beta_0 + \beta_1 t + \beta_2 \frac{1}{\sqrt{t}}$ an, d.h.

```
a10.lm <- lm(x ~ t + I(1/sqrt(t)))
```

so erhält man einen p-Wert $p(x) = 0.1298062$ mit der in Abbildung 12 rechts dargestellten Kurve. Dieser Ansatz wird also durch den Test nicht widerlegt, die Nullhypothese wird angenommen.

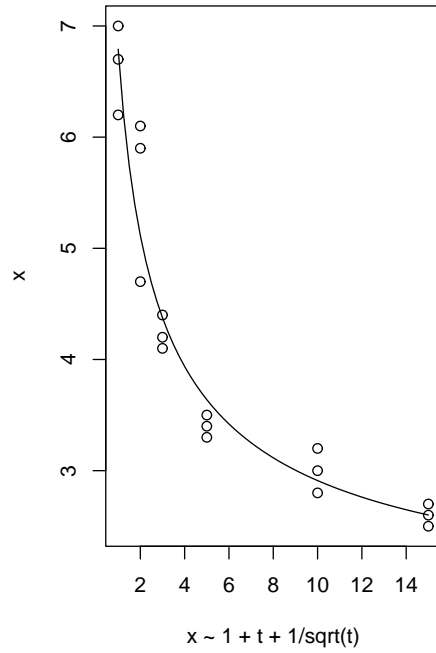
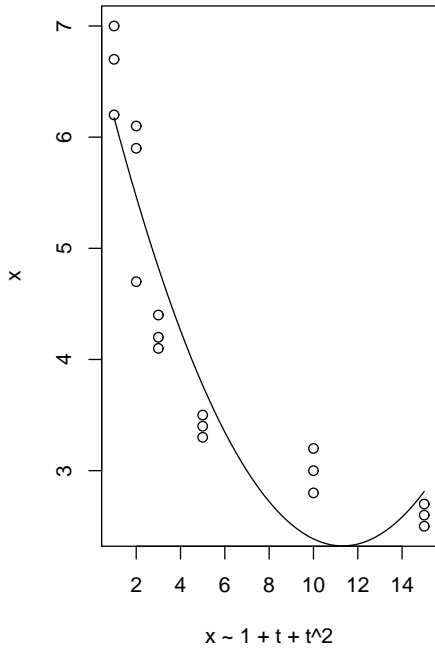


Abbildung 12: Test eines Regressionsansatzes

A. Statistische Berechnungen mit R

R ist eine objektorientierte und interpretierte Sprache und Programmierumgebung für Datenanalyse und Graphik, die unter der *GNU General Public License* frei verfügbar ist. Sie kann für alle gängigen Betriebssysteme und im Quellcode von <http://www.r-project.org> heruntergeladen werden.

Dort findet man auch die Einführung [17] in diese Programmiersprache und Handbücher im PDF-Format. Die Software enthält außerdem ein umfangreiches Hilfesystem. Eine ausführliche Darstellung der Programmiermöglichkeiten findet man in dem Buch [8] von Uwe Ligges, die Graphikprogrammierung beschreibt Paul Murrell [12] und speziell die Anwendung in der Statistik Brian S. Everitt und Torsten Hothorn [5].

In diesem Abschnitt werden zur Veranschaulichung der Arbeitsweise nur einige wenige Konzepte und Befehle vorgestellt, die in den Beispielen dieses Skripts Anwendung finden.

A.1. Vektoren

Das Basisobjekt in R ist der Vektor. Man erzeugt einen Vektor zum Beispiel durch die Zusammenfassung seiner Komponenten durch den Verkettungsoperator `c` wie bei

```
x <- c(10.4,5.6,3.1,6.4,21.7)
```

oder liest die Komponenten aus einer Datei ein

```
x <- scan("beispiel.dat")
```

Der Operator `<-` ist der Zuweisungsoperator, der den Operatoren `=` oder `:=` bei anderen Programmiersprachen entspricht. Man kann statt dessen meist, aber nicht immer das Symbol `=` verwenden. Um Probleme zu vermeiden, sollte man aber konsequent bei `<-` bleiben.

Arithmetische Operatoren

Die arithmetischen Operatoren `+`, `-`, `*`, `/` und `^` werden immer komponentenweise auf Vektoren angewandt (für Matrix-Vektor-Berechnungen gibt es eigene Operatoren). Haben die beteiligten Vektoren unterschiedliche Länge, so werden die kürzeren periodisch auf die größte Länge aufgefüllt.

Die Anweisung `1/x` liefert somit den Vektor

```
0.09615385 0.17857143 0.32258065 0.15625000 0.04608295
```

Mit dem Verkettungsoperator kann man auch Vektoren zusammenhängen.

```
> y <- c(x,0,x)
```

```
> y
```

```
[1] 10.4 5.6 3.1 6.4 21.7 0.0 10.4 5.6 3.1 6.4 21.7
```

und `v <- 2*x + y + 1` ergibt

```
[1] 32.2 17.8 10.3 20.2 66.1 21.8 22.6 12.8 16.9 50.8 43.5
```

Funktionen

`length(x)` liefert die Anzahl der Komponenten des Vektors `x`; `min(x)`, `max(x)` die minimale bzw. maximale Komponente, `mean(x)` das Stichprobenmittel und `var(x)` die Stichprobenvarianz, wie man aus dem Vergleich mit der Formel

```
sum( (x - mean(x))^2 ) / (length(x) - 1)
```

ersieht. `sort(x)` liefert einen Vektor, der die Komponenten von `x` in aufsteigender Reihenfolge enthält.

Reguläre Folgen

Anstelle von ausführlichen Erläuterungen zum Generieren von regulären Zahlenfolgen, die sich in [17] finden, geben wir hier einen kurzen selbsterklärenden Ausschnitt aus einer R-Sitzung wieder:

```
> 1:30
[1] 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25
[26] 26 27 28 29 30
> n <- 10
> # Achtung! : hat hoehere Prioritaet als +,-,*
> 1:n-1
[1] 0 1 2 3 4 5 6 7 8 9
> 1:(n-1)
[1] 1 2 3 4 5 6 7 8 9
> seq(2,10)
[1] 2 3 4 5 6 7 8 9 10
> seq(10,2)
[1] 10 9 8 7 6 5 4 3 2
> 10:2
[1] 10 9 8 7 6 5 4 3 2
> seq(-5,5,by=0.2)
[1] -5.0 -4.8 -4.6 -4.4 -4.2 -4.0 -3.8 -3.6 -3.4 -3.2 -3.0 -2.8 -2.6 -2.4 -2.2
[16] -2.0 -1.8 -1.6 -1.4 -1.2 -1.0 -0.8 -0.6 -0.4 -0.2 0.0 0.2 0.4 0.6 0.8
[31] 1.0 1.2 1.4 1.6 1.8 2.0 2.2 2.4 2.6 2.8 3.0 3.2 3.4 3.6 3.8
[46] 4.0 4.2 4.4 4.6 4.8 5.0
> rep(c(1,2),times=3)
[1] 1 2 1 2 1 2
```

Boolesche Vektoren

Boolesche Vektoren haben die Komponenten `TRUE` und `FALSE`. Sie ergeben sich aus dem komponentenweisen Vergleich mit den logischen Operatoren `<`, `<=`, `>`, `>=`, `==`, `!=`, `&`, `|`:


```

> x > 13
[1] FALSE FALSE FALSE FALSE TRUE
> !(x <= 13)
[1] FALSE FALSE FALSE FALSE TRUE
> x & y
[1] TRUE TRUE TRUE TRUE TRUE FALSE TRUE TRUE TRUE TRUE TRUE

```

Indexvektoren und Teilmengen

Teilmengen der Komponenten eines Vektors erhält man durch Angabe der gewünschten Indices in eckigen Klammern

```

> x[2]
[1] 5.6
> x[c(1,3,5)]
[1] 10.4 3.1 21.7
> x[1:2]
[1] 10.4 5.6
> x[-(1:2)]
[1] 3.1 6.4 21.7
> c("x","y")[rep(c(1,2,2,1),times=2)]
[1] "x" "y" "y" "x" "x" "y" "y" "x"
> paste(c("X","Y"),1:10,sep="")
[1] "X1" "Y2" "X3" "Y4" "X5" "Y6" "X7" "Y8" "X9" "Y10"

```

oder durch einen Booleschen Vektor der gleichen Länge:

```

> y <- rep(c(1,-1),times=4)
> y
[1] 1 -1 1 -1 1 -1 1 -1
> y[y<0]
[1] -1 -1 -1 -1
> # Auch eine Zuweisung ist moeglich:
> y[y<0] <- -y[y<0]
> # liefert das Gleiche wie die Funktion
> abs(y)
[1] 1 1 1 1 1 1 1 1

```

A.2. Wahrscheinlichkeitsrechnung

R kann Dichten, Verteilungsfunktionen und Quantile der für die Statistik wichtigsten Verteilungen berechnen. Ist `xyz` der Name der Verteilung, so ist `dxyz(t,...)` der Wert der Dichte, `pxyz(t,...)` der Wert der Verteilungsfunktion an der Stelle `x` bzw. an den Stellen der Komponenten des Vektors `x` und `qxyz(alpha,...)` das `alpha`-Quantil. ... steht dabei für die Parameter der Verteilung, wie Mittelwert oder Freiheitsgrade.

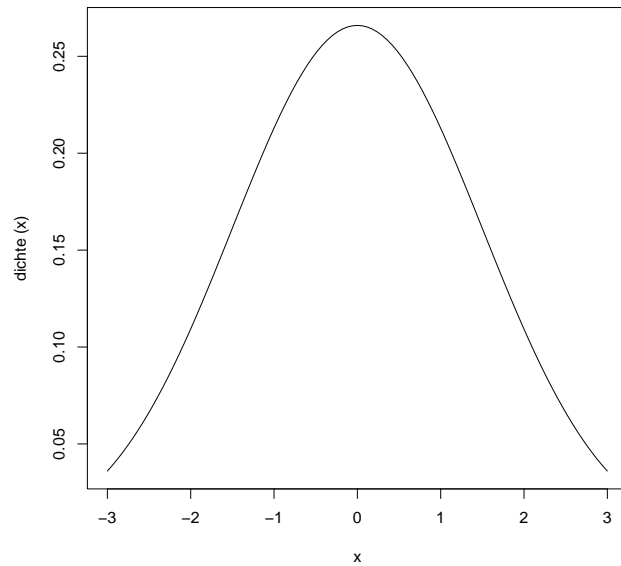


Abbildung 13: Dichte der Normalverteilung

Den Wert der Verteilungsfunktion der Normalverteilung mit Mittelwert μ und Varianz σ^2 an der Stelle t erhält man z.B. durch `pnorm(t,mu,sigma)`.

Funktionsplots

Um den Graphen einer Funktion graphisch auszugeben, darf die Funktion nur von einer Variablen abhängen. Zum Plot der Dichte der Normalverteilung mit Mittelwert 0 und Varianz 1.5^2 definieren wir zuerst die Funktion `dichte` und plotten sie anschließend aus:

```
> dichte <- function(t) { dnorm(t,0,1.5) }
> curve(dichte,from=-3,to=3)
```

Das von R gelieferte Resultat ist in Abbildung 13 dargestellt.

A.3. Statistische Kenngrößen von Stichproben

Der nachstehend verwendete Vektor `x` enthält die 59 Messwerte aus der Tabelle 1 auf Seite 5 (Bestimmung der Schmelztemperaturen von Wachsproben).

Einen ersten Überblick verschafft das Kommando `summary`:

```
> summary(x)
  Min. 1st Qu.  Median    Mean 3rd Qu.   Max.
62.85  63.36   63.53   63.59  63.84   64.42
```

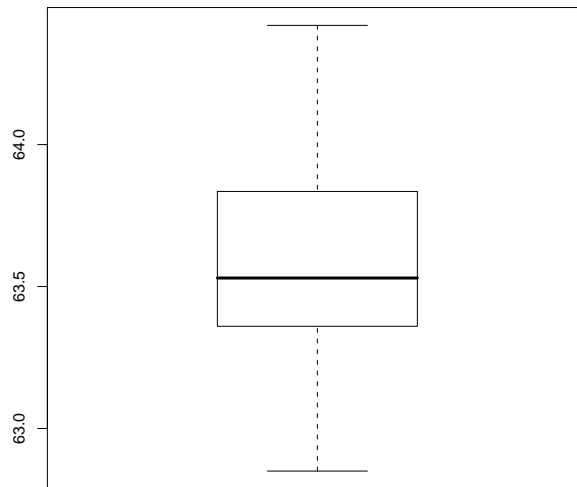


Abbildung 14: Der Boxplot der Stichprobe

das den kleinsten und größten Wert sowie die 25%-, 50%- und 75%-Quantile der Stichprobe ausgibt. (Das $x\%$ -Quantil ist die Zahl, unterhalb der $x\%$ der Stichprobenwerte liegen.

Das graphische Gegenstück dazu ist der Boxplot (Abb. 14)

```
> boxplot(x, range=0)
```

Mit `mean(x)` und `sd(x)` erhält man Mittelwert und Standardabweichung.

```
> mean(x)
[1] 63.58881
> sd(x)
[1] 0.3472209
```

Ein Histogramm erhält man mit `hist(x)` (Abb. 15). Dabei werden die absoluten Häufigkeiten nach oben abgetragen. Will man das Histogramm z.B. mit der Dichte der Normalverteilung vergleichen, empfiehlt sich ein zusätzliches Argument, durch das das Histogramm auf Gesamtfläche 1 skaliert wird (Abb. 16): `hist(x, freq=FALSE)`.

Möchte man den Graphen der Dichte über das Histogramm legen, so geht das nicht mit dem High-Level-Graphikbefehl `curve`, der das Histogramm löscht und den Graphen darstellt. Man muss eine Folge von Punkten auf dem Graphen erzeugen und mit dem Low-Level-Befehl `lines` einen Polygonzug konstruieren, der diese Punkte miteinander verbindet.

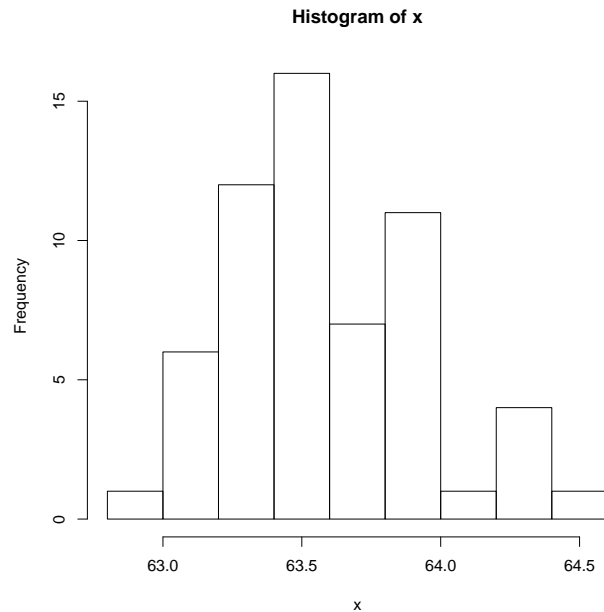


Abbildung 15: Histogramm mit absoluten Häufigkeiten

```

> hist(x,freq=FALSE)
> dichte <- function(t) { dnorm(t,mean(x),sd(x)) }
> curve(dichte,from=min(x),to=max(x))
> hist(x,freq=FALSE)
> dichte <- function(t) { dnorm(t,mean(x),sd(x)) }
> y <- seq(from=min(x),to=max(x),by=0.1)
> z <- dichte(y)
> lines(y,z,type="l",col="red")

```

Die empirische Verteilungsfunktion

Entsprechend der Definition $F(t) = P(X \leq t)$ der Verteilungsfunktion der Verteilung einer Zufallsvariablen X ist die empirische Verteilungsfunktion $F_x(t)$ einer Stichprobe an der Stelle t gleich der relativen Häufigkeit der Komponenten x_i von x mit $x_i \leq t$. In R heißt sie `ecdf(t)`. Einen Plot erhält man durch

```

> plot(ecdf(x),do.points=FALSE,verticals=TRUE)

```

Die Argumente werden in der Online-Hilfe zu `plot` erläutert. Zum Vergleich legen wir die Verteilungsfunktion der Normalverteilung darüber, wobei wir den Abszissenvektor `y` von oben benutzen (Abb. 17)

```

> z <- pnorm(y,mean(x),sd(x))
> lines(y,z,type="l",col="blue")

```

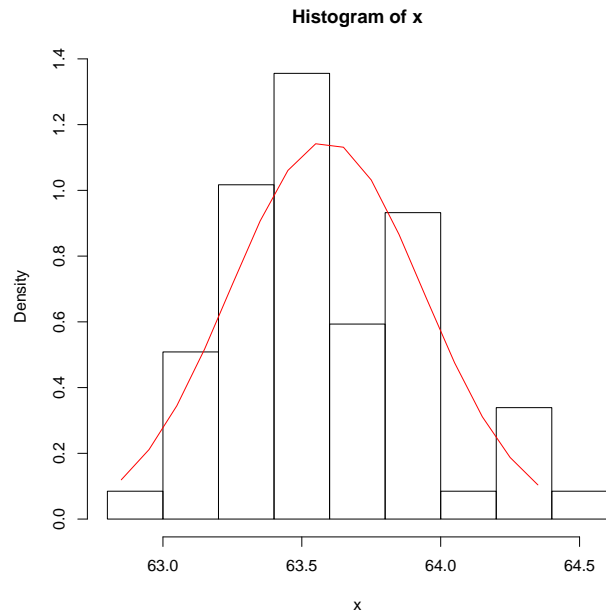


Abbildung 16: Vergleich des skalierten Histogramms mit einer Dichte

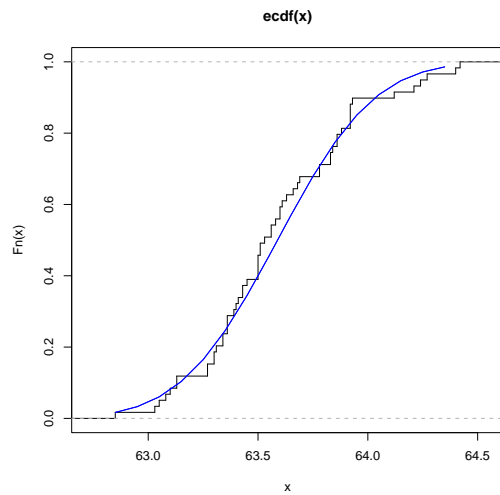


Abbildung 17: Die empirische Verteilungsfunktion

Die Stichprobenquantile

Die Stichprobenquantile sind die Lösungen der Gleichung $F_x(t) = q$. Nachdem die empirische Verteilungsfunktion eine Sprungfunktion ist, gibt es keine eindeutige Lösung,

wobei es verschiedene Konventionen für die Auswahl *des* q -Quantils gibt, die man in R mit dem Argument `type=1,2,...` wählen kann (s. Online-Hilfe). voreingestellt ist `type=7`.

```
> quantile(x,prob=c(0.2,0.4,0.6,0.8))
      20%   40%   60%   80%
63.328 63.500 63.608 63.868
```

Der Q-Q-Plot

Der Q-Q-Plot (Quantil-Quantil-Plot) dient der visuellen Überprüfung, ob eine bestimmte Verteilungsannahme zu einer Stichprobe passt. Beim Vergleich mit der Normalverteilung werden die Quantile der empirischen Verteilungsfunktion an den Stellen x_i gegen die entsprechenden Quantile der $\mathcal{N}(0,1)$ -Verteilung abgetragen. Je näher die Punkte an einer Geraden liegen, desto besser ist die Normalverteilungsannahme erfüllt. Q-Q-Plot mit Vergleichsgerade erhält man mit

```
> qqnorm(x); qqline(x)
```

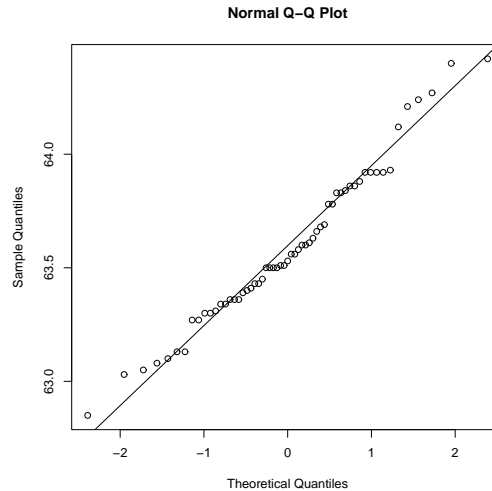


Abbildung 18: Q-Q-Plot mit Normalverteilung

(s. Abb. 18). Man kann auch mit anderen Verteilungstypen vergleichen, z.B. mit der t -Verteilung mit 5 Freiheitsgraden (Abb. 19):

```
> qqplot(qt(ppoints(250),df=5),x); qqline(x)
```

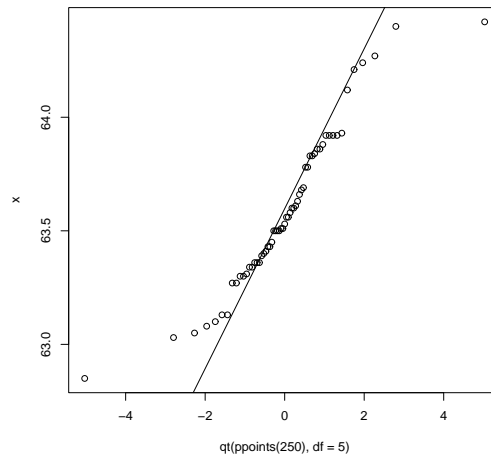


Abbildung 19: Q-Q-Plot mit t-Verteilung

A.4. Tabellen

Das übliche Speicherformat für statistische Daten sind Tabellen (tables, data frames). Bei einer Studie zur Wirkung von „Saugrüsseln“ an Benzinzapfsäulen wurden zum Beispiel bei einer Anzahl von Tankvorgängen jeweils die Größen

X1 : Temperatur im Fahrzeugtank

X2 : Temperatur des eingefüllten Benzins

X3 : Dampfdruck im Fahrzeugtank

X4 : Dampfdruck in der Zuleitung

Y : Menge der entwichenen Kohlenwasserstoffe

gemessen und in der folgenden Art in eine Datei (`mreg.dat`) geschrieben:

X1	X2	X3	X4	Y
28.00	33.00	3.00	3.49	22.00
24.00	48.00	2.78	3.22	27.00
33.00	53.00	3.32	3.42	29.00
29.00	52.00	3.00	3.28	29.00
33.00	44.00	3.28	3.58	27.00
31.00	36.00	3.10	3.26	24.00
32.00	34.00	3.16	3.16	23.00
34.00	35.00	3.22	3.22	23.00
33.00	51.00	3.18	3.18	26.00
		.		
		.		
		.		
35.00	35.00	3.23	3.12	22.00
36.00	35.00	3.27	3.27	23.00
37.00	35.00	3.26	3.12	22.00
37.00	36.00	3.26	3.12	23.00

Zum Einlesen einer Tabelle in R dient das Kommando `read.table`. Dabei muss man angeben, ob die erste Zeile Spaltennamen enthält oder bereits zu den Messdaten zählt:

```
> mr <- read.table("mreg.txt",header=TRUE)
```

Das Kommando `str` gibt die Struktur der Tabelle aus und `summary` die statistischen Kennzahlen der Spalten:

```
> str(mr)
'data.frame': 125 obs. of 5 variables:
 $ X1: num 28 24 33 29 33 31 32 34 33 33 ...
 $ X2: num 33 48 53 52 44 36 34 35 51 50 ...
 $ X3: num 3 2.78 3.32 3 3.28 3.1 3.16 3.22 3.18 3.12 ...
 $ X4: num 3.49 3.22 3.42 3.28 3.58 3.26 3.16 3.22 3.18 3.02 ...
 $ Y : num 22 27 29 29 27 24 23 23 26 19 ...
> summary(mr)
      X1          X2          X3          X4          Y
Min.   :24.00   Min.   :33.0   Min.   :2.590   Min.   :2.520   Min.   :16.00
1st Qu.:37.00   1st Qu.:50.0   1st Qu.:3.370   1st Qu.:3.430   1st Qu.:25.00
Median :60.00   Median :60.0   Median :4.110   Median :4.110   Median :30.00
Mean   :57.32   Mean   :56.5   Mean   :4.307   Mean   :4.253   Mean   :30.58
3rd Qu.:61.00   3rd Qu.:62.0   3rd Qu.:4.480   3rd Qu.:4.450   3rd Qu.:34.00
Max.   :93.00   Max.   :92.0   Max.   :7.600   Max.   :7.450   Max.   :58.00
```

Die einzelnen Spalten kann man mit `mr$X1`, `mr$X2`, ... , `mr$Y` ansprechen. Außerdem ist `mr` eine Matrix. Die Zeilen- und Spaltenzahl kann man mit `dim` erfahren:

```
> dim(mr)
[1] 125 5
```

Einzelne Komponenten oder Teilmatrizen erhält man wie bei Vektoren, wobei man bei den Spalten wahlweise den Index oder den Namen benutzen kann:

```
> mr[3,2]
[1] 53
> mr[3,"X2"]
[1] 53
> mr[, "X1"]
 [1] 28 24 33 29 33 31 32 34 33 33 34 37 36 34 36 39 34 35 36 34 59 59 60 60 60
 [26] 60 59 57 57 60 61 62 60 60 59 60 61 59 60 60 60 60 59 60 59 61 62 58 61
 [51] 59 60 61 62 60 60 90 92 90 90 92 92 92 92 91 62 62 61 61 59 59 93 92 88
 [76] 91 92 91 90 61 63 60 62 60 59 60 60 59 60 60 61 59 59 61 61 59 59 62 61 60
 [101] 59 60 60 60 60 61 60 60 62 60 59 40 37 37 34 35 35 37 35 34 35 35 36 37 37
> mr[4:8,]
      X1 X2  X3  X4  Y
```



```

4 29 52 3.00 3.28 29
5 33 44 3.28 3.58 27
6 31 36 3.10 3.26 24
7 32 34 3.16 3.16 23
8 34 35 3.22 3.22 23

```

A.5. Faktoren

Im der Tabelle B.2, die zum Beispiel 8.2 der Vorlesung gehört, bei dem der Einfluss der Vogelart auf die Größe von Kuckuckseiern untersucht wird, taucht eine neue Größe auf:

```

> kk <- read.table("kk.txt",header=TRUE)
> str(kk)
'data.frame': 45 obs. of 2 variables:
 $ Art: Factor w/ 3 levels "B","R","Z": 1 1 1 1 1 1 1 1 1 1 ...
 $ Lge: num 22 23.9 20.9 23.8 25 24 21.7 23.8 22.8 23.1 ...

```

Die Variable `Art`, die die Vogelart beschreibt, ist ein *Faktor* (*factor*) mit den *Ausprägungen* (*Stufen*, *levels*). Die Werte sind keine Zahlen, sondern Namen ohne Ordnungsstruktur.

Die statistischen Kenngrößen der gemessenen Längen möchte man natürlich getrennt nach Vogelarten haben. Dazu dient die Anweisung `tapply`, die einen Befehl auf die nach Faktorstufen getrennten Teilvektoren ausführt:

```

> tapply(kk$Lge, kk$Art, summary)
$B
  Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
20.90  22.85   23.10   23.11  23.80   25.00

$R
  Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
21.10  22.08   22.50   22.56  23.00   23.90

$Z
  Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
19.80  20.85   21.00   21.12  21.75   22.30

> tapply(kk$Lge, kk$Art, mean)
  B      R      Z
23.11429 22.55625 21.12000

```

Getrennte Boxplots erhält man z.B. mit

```

> boxplot(Lge ~ Art, data=kk, varwidth=TRUE)

```

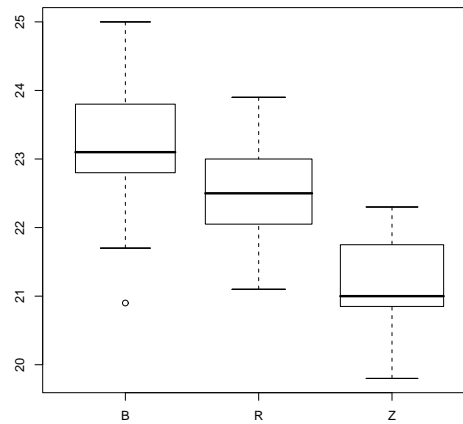


Abbildung 20: Boxplots pro Faktorstufe

mit dem in Abb. 20 dargestellten Ergebnis.

Komplexe Graphiken

Zum Abschluss noch ein Beispiel, wie man mehrere Bilder auf eine Graphikseite bringt: Den obigen Boxplot und die Q-Q-Plots der drei Messwertgruppen. Dazu wird eine 2×3 -Matrix konstruiert, die angibt, wieviele Bilder es geben soll und wie sie verteilt sind.

```
matrix(c(1,1,1,2,3,4),nrow=2,byrow=TRUE)
```

erzeugt aus dem Vektor (1,1,1,2,3,4), in der die Komponenten zeilenweise zugeordnet werden. Die drei Einsen in der ersten Zeile bedeuten, dass das erste Bild die erste Zeile ausfüllt und die Komponenten 2, 3 und 4, dass die folgenden drei Bilder von links nach rechts in die zweite Zeile positioniert werden. Der Befehl `layout` weist R an, die Grafikausgabe der vier aufeinanderfolgenden Plotbefehle entsprechend dieser Positionsangaben auszuführen.

```
> layout(matrix(c(1,1,1,2,3,4),nrow=2,byrow=TRUE))
> boxplot(Lge ~ Art, data=kk,varwidth=TRUE)
> qqnorm(kk$Lge[kk$Art=="B"]); qqline(kk$Lge[kk$Art=="B"])
> qqnorm(kk$Lge[kk$Art=="R"]); qqline(kk$Lge[kk$Art=="R"])
> qqnorm(kk$Lge[kk$Art=="Z"]); qqline(kk$Lge[kk$Art=="Z"])
```

Für die passende Beschriftung der einzelnen Bilder verweisen wir auf die Online-Hilfe und die angegebene Literatur.

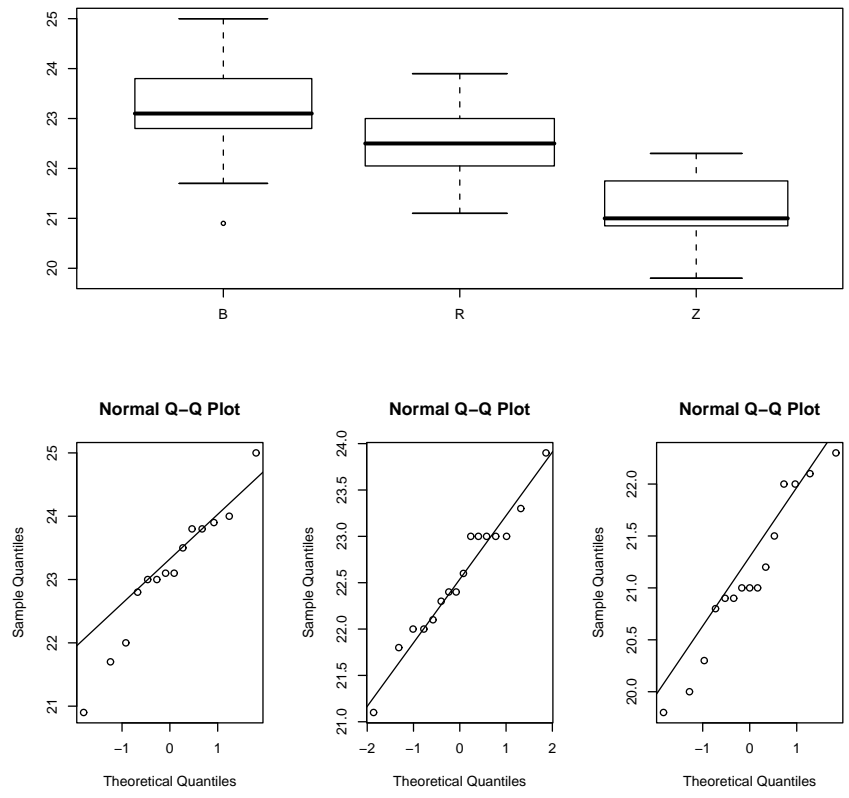


Abbildung 21: Mehrere Plots auf einer Graphik

B. Beispiel-Datensätze

B.1. bw.txt: Bremsweg in Abhängigkeit von der Geschwindigkeit

V	W	V	W
3.9	0.6	30.9	4.4
9.9	1.1	30.9	5.2
9.9	2.0	30.9	6.3
14.3	1.1	38.3	3.5
14.3	2.0	38.3	5.1
19.6	1.1	38.3	5.6
19.6	2.5	38.3	6.0
19.6	3.5	47.6	5.5
24.3	1.2	47.6	6.5
24.3	2.9	47.6	10.7
24.3	4.5	54.9	8.5
26.7	3.6	54.9	11.4
26.7	4.5	60.3	11.4
26.7	5.2	60.3	14.4

B.2. kk.txt: Länge von Kuckuckseiern bei verschiedenen Vogelarten

Art	Lge	Art	Lge	Art	Lge
B	22.0	R	21.8	Z	19.8
B	23.9	R	23.0	Z	22.1
B	20.9	R	23.3	Z	21.5
B	23.8	R	22.4	Z	20.9
B	25.0	R	22.4	Z	22.0
B	24.0	R	23.0	Z	21.0
B	21.7	R	23.0	Z	22.3
B	23.8	R	23.0	Z	21.0
B	22.8	R	23.9	Z	20.3
B	23.1	R	22.3	Z	20.9
B	23.1	R	22.0	Z	22.0
B	23.5	R	22.6	Z	20.0
B	23.0	R	22.0	Z	20.8
B	23.0	R	22.1	Z	21.2
		R	21.1	Z	21.0
		R	23.0		

B.3. a10.txt: Test eines Regressionsansatzes

t	x
1	6.2
1	6.7
1	7.0
2	4.7
2	5.9
2	6.1
3	4.1
3	4.2
3	4.4
5	3.3
5	3.4
5	3.5
10	2.8
10	3.0
10	3.2
15	2.5
15	2.6
15	2.7

Literatur

- [1] G. P. Beaumont. *Intermediate Mathematical Statistics*. Chapman and Hall, London, 1980.
- [2] Karl Bosch. *Angewandte Mathematische Statistik*. Vieweg, Reinbek bei Hamburg, 1976.
- [3] Peter Deuffhard und Andreas Hohmann. *Numerische Mathematik I*. Berlin, New York, 2002.
- [4] Annette J. Dobson. *An Introduction to Generalized Linear Models*. Chapman & Hall/CRC, Boca Raton, second edition, 2002.
- [5] Brian S. Everitt and Thorsten Hothorn. *A Handbook of Statistical Analyses Using R*. Chapman & Hall/CRC, Boca Raton, 2006.
- [6] Friedrich Graef. *Wahrscheinlichkeitsrechnung I für Informatiker und Ingenieure*. Vorlesungsskript, <http://www.am.uni-erlangen.de/home/graef/wr1>, 2003.
- [7] Erwin Kreyszig. *Statistische Methoden und ihre Anwendungen*. Vandenhoeck & Ruprecht, Göttingen, 1975.
- [8] Uwe Ligges. *Programmieren mit R*. Springer, Berlin, 2007.
- [9] Günter Menges. *Grundriss der Statistik, Teil 1: Theorie*. Westdeutscher Verlag, Opladen, 1972.
- [10] R. G. Miller. *Simultaneous Statistical Inference*. Springer, New York, 1981.
- [11] Alexander M. Mood, Franklin A. Graybill, and Duane C. Boes. *Introduction to the Theory of Statistics*. McGraw-Hill, Tokyo, 1974.
- [12] Paul Murrell. *R Graphics*. Chapman & Hall/CRC, Boca Raton, 2006.
- [13] John A. Rice. *Mathematical Statistics and Data Analysis*. Duxbury Press, Belmont, 1995.
- [14] Siegfried Schach und Thomas Schäfer. *Regressions- und Varianzanalyse*. Springer, Berlin, 1978.
- [15] Henry Scheffé. *The Analysis of Variance*. Wiley, New York, 1959.
- [16] Leopold Schmetterer. *Introduction to Mathematical Statistics*. Springer, Berlin, second edition, 1974.
- [17] W. N. Venables, D. M. Smith, and the R Development Core Team. *An Introduction to R*. <http://www.r-project.org>, 2006.

- [18] H. Witting und G. Nölle. *Angewandte Mathematische Statistik*. Teubner, Stuttgart, 1970.
- [19] Herrmann Witting. *Mathematische Statistik*. Teubner, Stuttgart, 1966.