STATISTIK

G.Dikta

20.08.2003

Vorwort: Ich möchte mich an dieser Stelle ganz besonders bei Frau Anke Graf bedanken, die den größten Teil des Textes TEXte und die Grafiken erstellt hat.

Inhaltsverzeichnis

Kapitel 1

Einleitung

Im folgenden Abschnitt werden wir anhand einiger Beispiele versuchen, den Begriff Statistik näher zu beschreiben.

1.1 Einführende Beispiele

Beispiel 1.1.1. Firma hat 100.000 Glühlampen hergestellt. Es soll vor Verkauf die Qualität festgestellt werden:

a) Wie kann man die Qualität einer Glühlampe messen?
z.B. funktionsfähig, Lebensdauer, ...
b) Wieviel Glühlampen sollen untersucht werden?
c) Wie sollen die Glühlampen zur Untersuchung ausgewählt werden?
d) Wie soll die Untersuchung durchgeführt werden?
(gleiche Bedingungen)
e) Wie sollen die Daten dargestellt werden?
f) Welche Schlüsse lassen ich ziehen?
g) Wie zuverlässig sind die Schlüsse?

Bemerkung 1.1.2. Statistik:

- Deskriptive Statistik: Darstellung der gewonnenen Daten hinsichtlich der Fragestellung. Methoden: Tabellen, Grafiken, Maßzahlen, ... (vgl. 1.1.1 e).
- Analytische Statistik: Schluss auf Gesetzmäßigkeiten. Methode: Wahrscheinlichkeitstheorie (vgl. 1.1.1 a,b,c,d,j,g).

Beispiel 1.1.3. a) Schätzen des Fischbestands eines Sees zur Bestimmung des ökologischen Zustands.
b) Zusammenhang zwischen Legeleistung und Gewicht von Hennen.
c) Zusammenhang zwischen Ausdehnung eines Metallstabs und der Temperatur.
d) Unterschiede bzgl. der Reißfestigkeit zweier Drahtsorten.
e) Zusammenhang zwischen Anzahl der Störche und der Geburten in einem Dorf.

Kapitel 2

Deskriptive Statistik

Die Aufgabe der deskriptiven Statistik ist die aussagekräftige Darstellung der erhobenen Daten.

2.1 Untersuchungseinheiten, Merkmale und Ausprägungen

Beispiel 2.1.1. Anhand des Beispiels "Zusammenhang zwischen Legeleistung und Gewicht von Hennen" erläutern wir einige allgemeine Begriffe

Hennen	-	Untersuchungseinheiten
Gewicht und Eier pro Monat (reelle Zahl)	-	Merkmale
tatsächlich ermitteltes Gewicht	-	Ausprägung des Merkmals Gewicht
tatsächliche Anzahl gelegter Eier pro Monat (rationale Zahl)	-	Ausprägung des Merkmals "Eier/Monat"

Die Ausprägungen der Merkmale der einzelnen Untersuchungseinheiten sind die Daten.

Man unterschiedet je nach Art des Merkmals unterschiedliche Merkmalstypen, deren Ausprägungen unterschiedliche Skalen besitzen.

- quantitativ: Z.B. Alter, Gewicht, Temperatur, Lebensdauer, ...
- qualitativ: Z.B. Haarfarbe, Geschlecht, Fabrikat, ...

Quantitative und qualitative Merkmale unterscheiden sich bzgl. der Skala (Maßeinstellung).

Nominalskala:

- Skala ist diskret (endlich viele Einteilungen).
- Daten sind nicht vergleichbar, es gibt keine Rangfolge.
- z.B.: Geschlecht: w/m
Farbe: blau/weiß/rot

Ordinalskala:

- Skala ist diskret (oder abzählbar unendlich).
- Daten lassen sich ordnen und unterliegen einer Rangfolge.
- Abstände lassen sich nicht interpretieren.
- z.B.: EG-Qualitätsnorm für Äpfel: Extra, I, II, III
  Noten: 1-6

Metrische Skala:

- Daten lassen sich ordnen und Abstände sind interpretierbar (gleiche Differenzen auf der Skala entsprechen gleichen Differenzen beim Merkmal).
- z.B.: Gewicht, Größe, Lebensdauer, Temperaturen

2.2 Darstellung von eindimensionalen Daten

Beispiel 2.2.1. Zur Analyse der innerartlichen Variabilität wurden die Flügellängen von n = 25 Insekten [mm] gemessen. Daten:$x_1, x_2, ..., x_{25}$

Urliste:

3.8	3.6	4.3	3.5	4.1	4.4	4.5	3.6	3.8
3.3	4.3	3.9	4.3	4.4	4.1	3.6	4.2	3.9
3.8	4.4	3.8	4.7	3.8	3.6	4.3		

Kleinstes Datum: $x_{10} = 3.3$; Größte Daten: $x_{22} = 4.7$. Skala: metrisch; 3.3 - 4.7; Unterteilung 0.1.

Bemerkung 2.2.2. Urliste ist unübersichtlich. Daher ist es nützlich, mehrere Daten zu einer neuen Kenngröße zusammenzufassen. Allgemein bezeichnet man eine Funktion, die auf Daten anwendbar ist, als Statistik.

Beispiel 2.2.3. Im Folgenden geben wir einige Beispiele bekannter Statistiken an:

a) $\frac{1}{25}\sum\limits_{i=1}^{25} x_i \to$ mittlere Flügellänge ist eine Statistik.

b) Bezeichne A die Menge der möglichen Ausprägungen eines Merkmals X und sei $B \subset A$. Dann wird durch

$$1_B(x) = \begin{cases} 1 & \text{falls x} \in B \\ 0 & \text{sonst} \end{cases}$$

die <u>Indikatorvariable</u> zu B definiert.
Sind dann $x_1, ..., x_n$ n gemessene Daten, dann bezeichnet

$$H_n(B) := \sum_{i=1}^{n} 1_B(x_i)$$

die <u>absolute Häufigkeit</u> des Auftretens in B (Anzahl der Daten in B).

$$h_n(B) := n^{-1} H_n(B)$$

bezeichnet die <u>relative Häufigkeit</u> des Auftretens in B

c) Sind die Daten der Größe nach geordnet, kann man sogenannte Summenhäufigkeiten betrachten. Zu jedem $x \in A(= \mathbb{R})$, A - Menge der möglichen Ausprägungen, wird die relative Anzahl der Daten gemessen, die $\leq x$ sind. Die sich ergebende Funktion nennt man empirische Verteilungsfunktion (evf).
Seien $x_1, ..., x_n$ Daten, dann bezeichnen wir die evf. hier mit F_n, wobei:

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

Beispiel 2.2.4. Fortsetzung von Beispiel 2.2.1

Tabelle der Häufigkeiten

Skalenwert	Häufigkeit		Summenhäufigkeit	emp. VF
	abs	rel.		
a_j	$H_n(a_j)$	$h_n(a_j)$	$\sum_{k=1}^{j} H_n(a_k)$	$F_n(a_j)$
$a_1 = 3.3$	1	0.04	1	$\frac{1}{25} = 0.04$
$a_2 = 3.4$	0	0	1	$\frac{1}{25} = 0.04$
$a_3 = 3.5$	1	0.04	2	$\frac{2}{25} = 0.08$
$a_4 = 3.6$	4	0.16	6	$\frac{6}{25} = 0.24$
$a_5 = 3.7$	0	0	0	$\frac{6}{25} = 0.24$
$a_6 = 3.8$	5	0.2	11	$\frac{11}{25} = 0.44$
$a_7 = 3.9$	2	0.08	13	$\frac{13}{25} = 0.52$
$a_8 = 4.0$	0	0	13	$\frac{13}{25} = 0.52$
$a_9 = 4.1$	2	0.08	15	$\frac{15}{25} = 0.6$
$a_{10} = 4.2$	1	0.04	16	$\frac{16}{25} = 0.64$
$a_{11} = 4.3$	4	0.16	20	$\frac{20}{25} = 0.8$
$a_{12} = 4.4$	3	0.12	23	$\frac{23}{25} = 0.92$
$a_{13} = 4.5$	1	0	24	$\frac{24}{25} = 0.96$
$a_{14} = 4.6$	0	0.04	24	$\frac{24}{25} = 0.96$
$a_{15} = 4.7$	1	0.04	25	$\frac{25}{25} = 1$

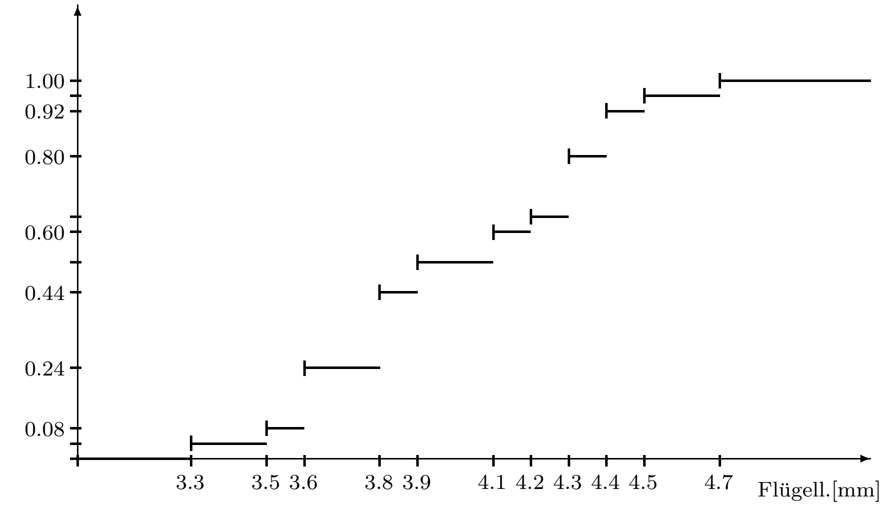


Abbildung 2.1: Empirische Verteilungsfunktion

Aus der Darstellung lassen sich sofort zwei typische Fragen beantworten:

a) Wieviel % der Insekten haben eine Flügellänge, die $\leq$ 3.9mm ist?
($F_n(3.9) = 0.52$) $\rightarrow$ 52 %

b) Welches ist die kleinste Flügellänge, die $\geq$ 40% der gemessenen Flügellänge ist?
(3.8mm)

Bemerkung 2.2.5. Besitzen die Daten viele verschiedene Werte, dann ist es übersichtlicher, benachbarte Werte in Klassen (oder zu Gruppen) zusammenzufassen.

Klassifizierte Häufigkeitstabelle:

Klasse		Klassenmitte	Häufigkeiten	
Nr.	Beschr.		abs.	rel.
1	$3.3 \leq x \leq 3.6$	3.45	2	0.08
2	$3.6 \leq x \leq 3.9$	3.75	9	0.36
3	$3.9 \leq x \leq 4.2$	4.05	4	0.16
4	$4.2 \leq x \leq 4.5$	4.35	8	0.32
5	$4.5 \leq x \leq 4.8$	4.65	2	0.08

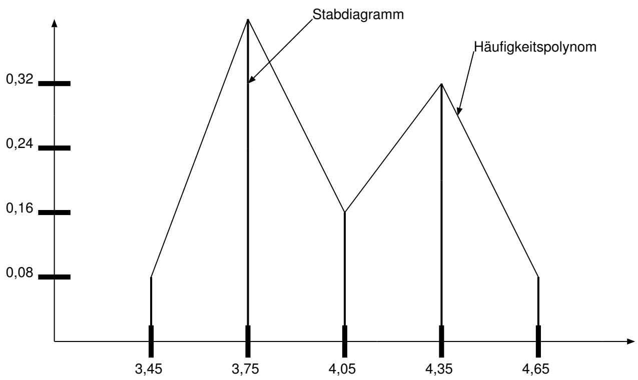
Die Klassen sollten gleiche Breite besitzen, falls keine sachlichen Gründe dagegen sind. (z.B. Klausurpunkt $\rightarrow$ Note)
Anzahl der Klassen sollte möglichst klein sein, wobei aber die spezifische Information der Daten noch sichtbar sein muss (Manipulation!).
Ohne Vorinformation wählt man häufig als Klassenbreite:

$$\text{b} = \frac{x_{\text{größtes}} - x_{\text{kleinstes}}}{1 + 3.32 \underbrace{\log(n)}_{\ln_{10}(n)}}$$

$$\text{Im Beispiel also: b} = \frac{4.7 - 3.3}{1 + 3.32 \log(25)} \approx 0.25$$

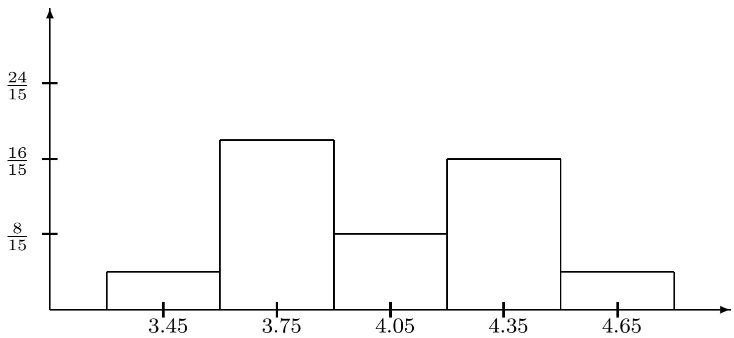
Bemerkung 2.2.6. Graphische Darstellung der relativen Häufigkeiten

a) Stabdiagramm und Häufigkeitspolygon:



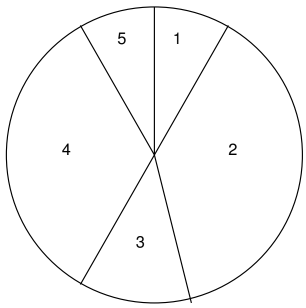
Aus dem Stabdiagramm lässt sich eine spezifische Eigenschaft, nämlich eine zweigipflige Verteilung, erkennen. (Bei den Insekten wurde nicht zwischen m/w unterschieden).

b) Histogramm:



Die Fläche zu jeder Klasse entspricht jeweils der relativen Häufigkeit der Klasse, d.h. die gesamte Fläche des Histogramms entspricht 1.

c) Kreisdiagramm:



Kreis wird in Flächen zerlegt, die proportional zu den rel. Häufigkeiten sind.

2.3 Maßzahlen für eindimensionale Daten

Man unterscheidet verschiedene Typen von Statistiken (Funktion der Daten). Ein Typ ist das Lokationsmaß bzw. Lageparameter.

Definition 2.3.1. Ein Lokationsmaß ist eine Maßzahl (Statistik) für einen zentralen Wert, um den sich die Daten gruppieren.

Beispiel 2.3.2. Gegeben seien die Daten $x_1, x_2, ..., x_n$.

a) Bei metrischer Skala:

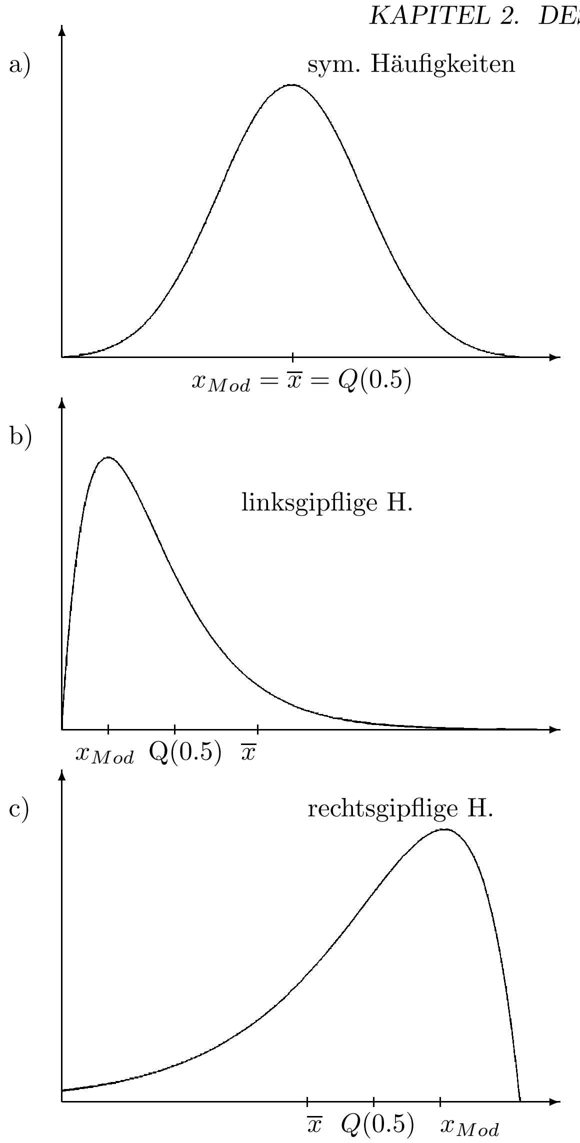
$$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \quad \text{arithmetisches Mittel}$$

b) Bezeichnen $x_{(1)} \leq x_{(2)} \leq ... \leq x_{(n)}$ die geordneten Daten (Ordnungsstatistik).

$$\text{Median:} \quad \mathrm{Q}(0.5) := \begin{cases} x_{\left(\frac{n+1}{2}\right)}, & \text{n ungerade} \\ \frac{1}{2}\left(x_{\left(\frac{n}{2}\right)} + x_{\left(\frac{n+2}{2}\right)}\right), & \text{n gerade} \end{cases}$$

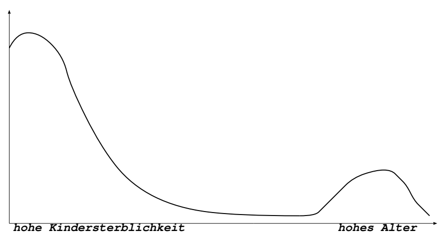
c) Falls nominalskaliertes Merkmal, dann scheiden Median und arithmetisches Mittel aus. Es bleibt der Modalwert (Modus) x_{mod}. Dies ist der Wert mit der größten Häufigkeit.

Bemerkung 2.3.3. Was messen $\overline{x}$, Q(0.5), x_{mod}?



Bemerkung 2.3.4. Lokationsmaße sind nur bei eingipfliger Verteilung der Häufigkeiten sinnvoll! Betrachte hierzu als Beispiel

"Lebensdauer der Menschen im Mittelalter" (vgl.Hartung S.37 ff)



Lagemaße führen zu Fehlinformation! Um mehrgipflige Verteilungen zu untersuchen, muss der Median zur Quantilfunktion verallgemeinert werden.

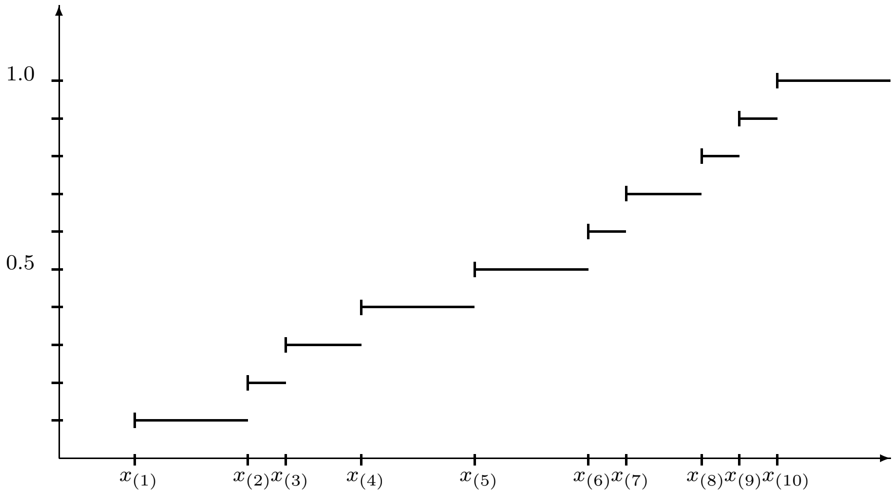
Definition 2.3.5. Gegeben seien die Daten $x_1, ..., x_n$ und $x_{(1)}, ..., x_{(n)}$ die zugehörige Ordnungsstatistik. Für $0 \leq \alpha \leq 1$ heißt

$$Q(\alpha) := \Big\{ x_{(k)} \; : \; \text{k ist kleinste natürliche Zahl mit } k \geq n \cdot \alpha$$

das zugehörige α-Quantil.

Beispiel 2.3.6.

a) Ist n ungerade, dann ist der Median das 0.5 - Quantil.
b) Seine $x_1, ..., x_{10}$: 10 gemessene Daten.



Q(0.5) $= x_{(5)}$
Q(0.85) $= \Big\{ x_{(k)}$: k kleinste natürliche Zahl mit k $\geq 10 \cdot 0.85 = 8.5 = x_{(9)}$

c) Für die Lebensdauer im Mittelalter sind $Q(0.1)$, $Q(0.5)$, $Q(0.9)$ zusammengenommen viel aussagekräftiger als $\overline{x}, x_{mod}$ oder Q(0.5) allein.

Neben dem Lagemaß ist man natürlich auch daran interessiert, die Streuung der Daten um einen entsprechenden Ort als Maßzahl zu erfassen.

Definition 2.3.7. Sei $x_{(1)} \leq x_{(2)} \leq ... \leq x_{(n)}$ die Ordnungsstatistik zu den Daten $x_1, x_2, ..., x_n$. Dann bezeichnet

$$R := x_{(n)} - x_{(1)}$$

die Spannweite (Range) der Messreihe.

Bemerkung 2.3.8. R ist von den extremen Ordnungsgrößen $x_{(n)}, x_{(1)}$ abhängig, daher anfällig bzgl. Ausreißern.

Definition 2.3.9. Gegeben sei die Ordnungsstatistik $x_{(1)} \leq x_{(2)} \leq ... \leq x_{(n)}$. Dann bezeichnet

$$V := Q(\frac{3}{4}) - Q(\frac{1}{4})$$

den Quartialabstand (d.h. Spanne der mittleren 50% der Daten).

Weitere Möglichkeiten: Berechne den mittleren "Abstand" der Daten zu einem festen Wert.

$$\text{z.B.} \quad \frac{1}{n}\sum_{i=1}^{n}|x_i - c| \quad \text{bzw.} \quad \frac{1}{n}\sum_{i=1}^{n}(x_i - c)^2$$

c muss dabei so gewählt werden, dass die mittleren Abstände minimal sind!

Beim ersten Abstand ergibt sich für c der Median und beim zweiten das arithmetische Mittel! (vgl. Mathe).

Definition 2.3.10. Die Größe

$$d_n := \frac{1}{n}\sum_{i=1}^{n}|x_i - Q(0.5)|$$

heißt mittlere absolute Abweichung vom Median.

Die Größe

$$s_n^2 := \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2, \quad \text{mit} \quad \overline{x} := \frac{1}{n}\sum_{i=1}^{n}x_i$$

wird die Stichprobenvarianz bzw.

$$s_k = \sqrt{s_n^2}$$

die Standardabweichung der Stichprobe genannt.

Bemerkung 2.3.11. Für die Stichprobenvarianz gilt folgender Zusammenhang:

$$s_n^2 = \frac{1}{n}\sum_{i=1}^{n}x_i^2 - \overline{x}^2$$

Beweis. Der Beweis ergibt sich aus

$$\begin{aligned} s_n^2 &= \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i^2 - 2x_i\overline{x} + \overline{x}^2) \\ &= \frac{1}{n}\sum_{i=1}^{n}x_i^2 - 2\overline{x}\underbrace{\frac{1}{n}\sum_{i=1}^{n}x_i}_{\overline{x}} + \underbrace{\frac{1}{n}\sum_{i=1}^{n}\overline{x}^2}_{\overline{x}^2} \end{aligned}$$

□

Beispiel 2.3.12. Intuitiv würde man erwarten, dass die Körpergrößen 6-jähriger Mädchen stärker streuen als die Körpergrößen erwachsener Frauen, da die 6-jährigen Mädchen sich sicherlich hinsichtlich der Wachstumsentwicklung voneinander unterscheiden. Der Vergleich der Standardabweichungen zweier Messreihen zeigt aber das Gegenteil:

Größe	n	$\overline{x}$	s_n
6-jährige Mädchen	77	112,6	4,64
20-jährige Frauen	51	162,6	5,12

Betrachtet man dagegen $\frac{s_n}{\overline{x}}$, so ergibt sich:

$$\begin{array}{rrcrcl} \text{6-jähr.} & s_n/\overline{x} & = & 4,64/112,6 & \approx & 4,12 \cdot 10^{-2} \\ \text{20-jähr.} & s_n/\overline{x} & = & 5,12/162,6 & \approx & 3,15 \cdot 10^{-2} \end{array}$$

Dieser Vergleich entspricht jetzt unseren Erwartungen.

Definition 2.3.13. Seien $\overline{x}$ und s_n arithmetisches Mittel und Standardabweichung einer Messreihe. Dann bezeichnet

$$cv := \frac{s_n}{|\overline{x}|}$$

den Variationskoeffizienten.

2.4 Darstellung von zweidimensionalen Daten

Betrachte zunächst

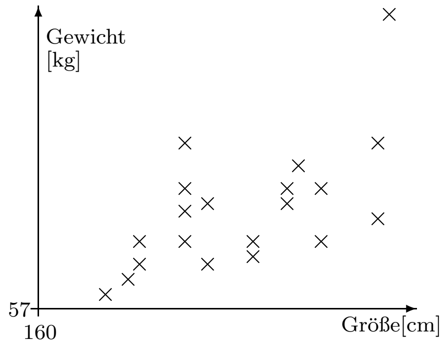
Beispiel 2.4.1. Gewicht und Größe von 20 Schülern einer Klasse:

Schüler(i)	Größe	Kateg.	Rang(R_i)	Gewicht	Kateg.	Rang(Q_i)
1	174	m	9,5	62	l	3,5
2	182	g	15	75	m	16
3	178	m	11,5	63	l	5
4	190	g	20	95	s	20
5	172	k	6,5	69	l	10
6	165	k	1	58	l	1
7	172	k	6,5	78	m	17,5
8	189	g	18,5	84	s	19
9	168	k	3,5	62	l	3,5
10	181	g	13,5	70	m	11,5
11	172	k	6,5	72	m	14
12	178	m	11,5	65	l	7,5
13	174	m	9,5	70	m	11,5
14	184	g	16,5	65	l	7,5
15	189	g	18,5	78	m	17,5
16	167	k	2	60	l	2
17	172	k	6,5	65	l	7,5
18	184	g	16,5	72	m	14
19	168	k	3,5	65	l	7,5
20	181	g	13,5	72	m	14

Es stellt sich hier sofort die Frage nach einem möglichen Zusammenhang zwischen Größe und Gewicht.

Um eine Idee bzgl. eines Zusammenhangs zu erhalten, stellen wir die Daten grafisch dar.

Definition 2.4.2. Fassen wir das Gewicht als Funktion (Zusammenhang) der Größe auf, dann ergibt sich das folgende Streudiagramm: Aus diesem Streudiagramm lässt sich schon erkennen, dass es eine wachsende Tendenz gibt, d.h. je größer desto schwerer.



metrische Daten!

Natürlich kann man ein Streudiagramm nur dann benutzen, wenn die Daten metrisch sind. Handelt es sich um gruppierte Daten, dann benutzt man zur Darstellung eine Kontingenztafel.

Definition 2.4.3. Da unsere Daten metrisch sind, werden sie zunächst in Klassen eingeteilt.

Klassifizieren der Größe:

$$\text{Kl. Breite} \quad b_{Gr.} = \frac{190-165}{1+3.32\ln_{10}(20)} \approx 4.5 \approx 5$$

oder etwas willkürlich

$$\text{klein} := [165, 173[\quad \text{mittel} := [173, 181[\quad \text{groß} := [181, 191[$$

Klassifizieren des Gewichts:

$$\text{leicht} := [58, 70[\quad \text{mittel} := [70, 82[\quad \text{schwer} := [82, 95[$$

Nachdem die Daten so gruppiert wurden, können wir sie in eine Kontingenztafel eintragen. Es ergibt sich jetzt für die Daten:

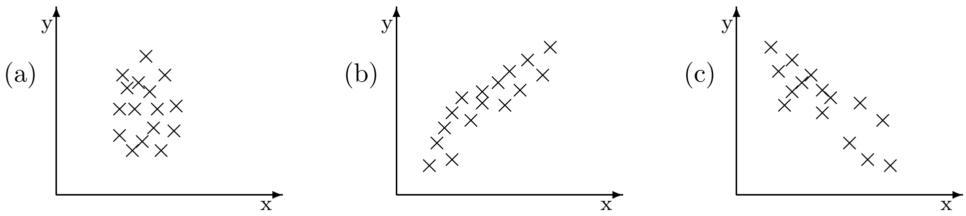
Größe / Gewicht	klein	mittel	groß	Zeilensumme	Zeilenhäufigkeit
leicht	6	3	1	10	$\frac{10}{20} = 0.5$
mittel	2	1	5	8	$\frac{8}{20} = 0.4$
schwer	-	-	2	2	$\frac{2}{20} = 0.1$
Spaltensumme	8	4	8	20	
Spaltenhäufigkeit	$\frac{8}{20} = 0.4$	$\frac{4}{20} = 0.2$	$\frac{8}{20} = 0.4$		

D.h. 6 Daten gehören sowohl in die Gewichtsklasse "leicht" wie zugleich auch in die Klasse "klein" hinsichtlich der Körpergröße, usw.

2.5 Maßzahlen für zweidimensionale Daten

Im Folgenden seien (x_1, y_1), ..., (x_n, y_n) zweidimensionale metrische Daten.

Betrachte die folgenden Streudiagramme:

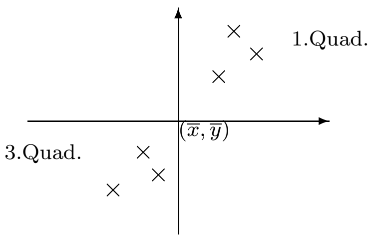


Offensichtlich erhalten wir

(a) richtungsloser Punkthaufen
(b) $x \uparrow \quad \Rightarrow \quad$ y $\uparrow$ wachsende Tendenz + Störung
(c) $x \uparrow \quad \Rightarrow \quad$ y $\downarrow$ fallende Tendenz + Störung

Wie lässt sich diese Tendenz messen? Als Motivation zur Gewinnung einer geeigneten Maßzahl betrachte:

Lege Koordinatenursprung in $(\overline{x}, \overline{y})$
$\Rightarrow$ (x_i, y_i) im 1. oder 3. Quadranten, falls $(x_i - \overline{x})(y_i - \overline{y}) > 0$
$\Rightarrow$ wachsende Tendenz, wenn viele dieser Produkte > 0



Setze daher:

$$\begin{aligned} r_{x,y} : &= \frac{\sum_{i=1}^{n} (x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_{i=1}^{n} (x_i - \overline{x})^2 \sum_{j=1}^{n} (y_i - \overline{y})^2}} \\ &= \frac{\frac{1}{n} \sum_{i=1}^{n} (x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\frac{1}{n} \sum_{i=1}^{n} (x_i - \overline{x})^2} \sqrt{\frac{1}{n} \sum_{j=1}^{n} (y_i - \overline{y})^2}} \\ &=: \frac{s_{xy}}{s_x s_y} \end{aligned}$$

Definition 2.5.1. Falls zwischen x und y ein linearer Zusammenhang besteht und die Daten metrisch sind, dann wird durch

$$r_{x,y} := \frac{s_{x,y}}{s_x s_y} \quad \text{, mit} \quad s_{x,y} := \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y}),$$

der Korrelationskoeffizient von Bravais-Pearson definiert. s_x bzw. s_y bezeichnen in der Formel die Standardabweichung der $x-$ bzw. $y-$Daten.

Bemerkung 2.5.2. Zur Berechnung von $r_{x,y}$ benutzt man

$$r_{x,y} = \frac{n\sum\limits_i x_i y_i - \left(\sum\limits_i x_i\right)\left(\sum\limits_j y_j\right)}{\left[\left(n\sum\limits_i x_i^2 - \left(\sum\limits_i x_i\right)^2\right)\left(n\sum\limits_j y_j^2 - \left(\sum\limits_j y_j\right)^2\right)\right]^{\frac{1}{2}}}$$

Beispiel 2.5.3. Für die Daten aus Beispiel 2.4.1 ergibt sich:

$$\sum_{i=1}^{20} x_i y_i = 248754, \quad \sum_{i=1}^{n} x_i^2 = 627718, \quad \sum_{i=1}^{n} y_i^2 = 99508, \quad \left(\sum_{i=1}^{n} x_i\right)^2 = 3540^2, \quad \left(\sum_{i=1}^{n} y_i\right)^2 = 1400^2$$

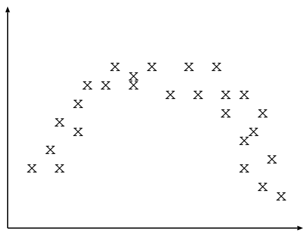
$$r_{x,y} = \frac{20 \cdot 248754 \cdot 3540 \cdot 1400}{\left[\left(20 \cdot 627718 - 3540^2\right)\left(20 \cdot 99508 - 1400^2\right)\right]^{\frac{1}{2}}} = ... \approx 0.73$$

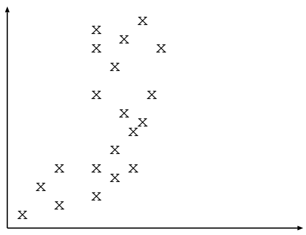
Also ein positiver Zusammenhang, d.h. "je größer desto schwerer"

Bemerkung 2.5.4. (i) $r_{x,y} \in$ [-1,1]

$$r_{x,y} \approx \begin{cases} -1 & \text{starker fallender Zusammenhang} \\ 0 & \text{kein Zusammenhang} \\ 1 & \text{starker positiver Zusammenhang} \end{cases}$$

(ii) Falls Streudiagramme die folgende Form haben, ist $r_{x,y}$ sinnlos, da x-y nicht linear! In diesem





Fall nimmt man Rangkorrelation.

(iii) Vorsicht bei Scheinkorrelation! z-Dorfgröße, x-Störche, y-Geburten

$$\begin{tikzcd} & x \\ z \arrow[ur] \arrow[dr] & \\ & y \end{tikzcd}$$

z↑ ⇒ x↑ und y↑

⇒ pos. Zusammenhang zwischen x und y!

(iv) Falls y = a + bx (d.h. nicht gestört)

$$\begin{aligned}
r_{x,y} &= \frac{\left(\frac{1}{n}\sum_i^n (x_i-\overline{x})(y_i-\overline{y})\right)}{\sqrt{\frac{1}{n}\sum_i^n (x_i-\overline{x})^2}\sqrt{\frac{1}{n}\sum_j^n (y_j-\overline{y})^2}} \\
&= \frac{\frac{1}{n}\sum_i^n (x_i-\overline{x})\left[(a+bx_i)-(a+b\overline{x})\right]}{\sqrt{\frac{1}{n}\sum_i^n (x_i-\overline{x})^2}\sqrt{\frac{1}{n}\sum_j^n \left[(a+bx_i)-(a+b\overline{x})\right]^2}} \\
&= \frac{b\frac{1}{n}\sum_i^n (x_i-\overline{x})^2}{\sqrt{\frac{1}{n}\sum_i^n (x_i-\overline{x})^2}\,|b|\sqrt{\frac{1}{n}\sum_i^n (x_i-\overline{x})^2}} \\
&= \frac{b}{|b|} = \begin{cases} 1 & \text{falls b} > 0 \\ -1 & \text{falls b} < 0 \end{cases}
\end{aligned}$$

Definition 2.5.5. Seien $x_1, ..., x_n$ die erhobene Daten und $x_{(1)} \leq ... \leq x_{(n)}$ zugehörige Ordnungsstatistik. Dann bezeichnet

$$R_i := \{\text{k:} \quad \text{falls } x_i = x_{(k)}$$

den Rang von x_i in der Stichprobe $x_1, ..., x_n$.

Mit Hilfe der Ränge lässt sich nun ein weiterer Korrelationskoeffizient definieren, der nicht mehr die stark einschränkende Linearität zwischen den $x-$ und $y-$Daten voraussetzt.

Definition 2.5.6. Seien (x_1, y_1), ..., (x_n, y_n) erhobene Daten (metrisch oder ordinal) und bezeichne R_i den Rang von x_i in der $x-$Stichprobe bzw. Q_i den Rang von y_i in der $y-$Stichprobe. Betrachte nun (R_1, Q_1), ..., (R_n, Q_n). Besteht zwischen x und y ein monotoner Zusammenhang, so lässt sich dieser mit

$$r_{sp} = \frac{\frac{1}{n}\sum_i^n (R_i-\overline{R})(Q_i-\overline{Q})}{\sqrt{\frac{1}{n}\sum_i^n (R_i-\overline{R})^2}\sqrt{\frac{1}{n}\sum_i^n (Q_j-\overline{Q})^2}}$$

dem Rangkorrelationskoeffizient von Spearman untersuchen. Dabei ist

$$\overline{R} := \frac{1}{n}\sum_i^n R_i \quad \text{bzw.} \quad \overline{Q} := \frac{1}{n}\sum_i^n Q_i$$

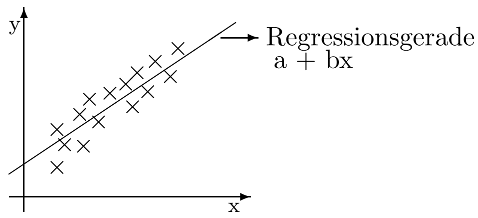
Bemerkung 2.5.7. Falls Daten in der Stichprobe identisch sind, so werden die zugehörigen Ränge durch Mittelbildung bestimmt. Z.B.:

$$\begin{aligned}
& x_1 = 3,\ x_2 = 7,\ x_3 = 9,\ x_4 = 7,\ x_5 = 7 \\
\Rightarrow \quad & R_1 = 1,\ R_2 = \frac{2+3+4}{3} = 3,\ R_3 = 5,\ R_4 = R_5 = 3
\end{aligned}$$

Beispiel 2.5.8. Als Beispiel zur Bestimmung der Ränge vgl. die Tabelle aus Bsp. 2.4.1! Für den Rangkorrelationskoeffizienten ergibt sich in diesem Beispiel: $r_{sp} \approx 0.697$

Bemerkung 2.5.9. Maßzahlen für den "Zusammenhang von nominalen Merkmalen" werden später, d.h. beim χ^2-Test, behandelt.

Falls das Streudiagramm einen linearen Zusammenhang y = a + bx rechtfertigt, kann man durch den Punkthaufen eine Gerade legen, die "möglichst gut" zu den Daten passt.



Satz 2.5.10. Die Regressionsgerade wird nach der Methode der kleinsten Quadrate bestimmt, d.h. für a, b muss gelten

$$\sum_{i=1}^{n} (y_i - (a + bx_i))^2 = \min_{\alpha,\beta} \sum_{i=1}^{n} (y_i - (\alpha + \beta x_i))^2$$

Die Lösung dieser Minimierung führt zu (vgl. Mathe)

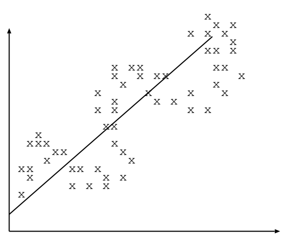
$$\begin{aligned} b &= \frac{\sum_i (x_i - \overline{x})(y_i - \overline{y})}{\sum_i (x_i - \overline{x})^2} = \frac{n \sum_i x_i y_i - (\sum_i x_i)(\sum_i y_i)}{n \sum_i x_i^2 - (\sum_i x_i)^2} \\ a &= \overline{y} - b\overline{x} \\ \overline{x} &= \frac{1}{n} \sum_i x_i \\ \overline{y} &= \frac{1}{n} \sum_i y_i \end{aligned}$$

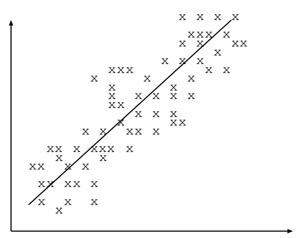
Die Werte

$$y_i - (a + bx_i), \quad i = 1, ..., n$$

werden Residuen (Fehler!) genannt.

Bemerkung 2.5.11. Ob ein linearer Zusammenhang überhaupt vorliegt, lässt sich durch Betrachtung der Residuen folgern. Falls kein linearer Zusammenhang vorliegt, versuche Transformation, z.B. $\ln(y)$!





Bemerkung 2.5.12. Sei $g(x) := a + bx$ die ermittelte Regressionsgerade. Dann gilt:

$$g(\overline{x}) = a + b\overline{x} = (\overline{y} - b\overline{x}) + b\overline{x} = \overline{y}$$

Die Stichprobenvarianz der $y-$Daten, $s_y^2 = 1/n \sum_{i=1}^{n}(y_i - \overline{y})$, misst die durchschnittliche Veränderung der $y-$Daten. Betrachte jetzt

$$\frac{1}{n}\sum_{i=1}^{n}\left(g(x_i) - g(\overline{x})\right)^2 = \frac{1}{n}\sum_{i=1}^{n}\left(g(x_i) - \overline{y}\right)^2$$

Diese Größe ist ein Maß für die durchschnittlich erklärbare Veränderung in den $y-$Daten. Eine erklärbare Veränderungen in einem $y-$Datum wird dabei ausschließlich durch die entsprechende Veränderung in dem zugehörigen $x-$Datum hervorgerufen, d.h.

$$\overline{x} \to x_i \Rightarrow g(\overline{x}) \to g(x_i)$$

Definition 2.5.13. Ein Maß für die Qualität der Anpassung ist das Bestimmtheitsmaß, definiert durch

$$B := \frac{\text{erklärbare Änderung}}{\text{vollständige Änderung}} = \frac{\sum_i (a + bx_i - \overline{y})^2}{\sum_j (y_j - \overline{y})^2}$$

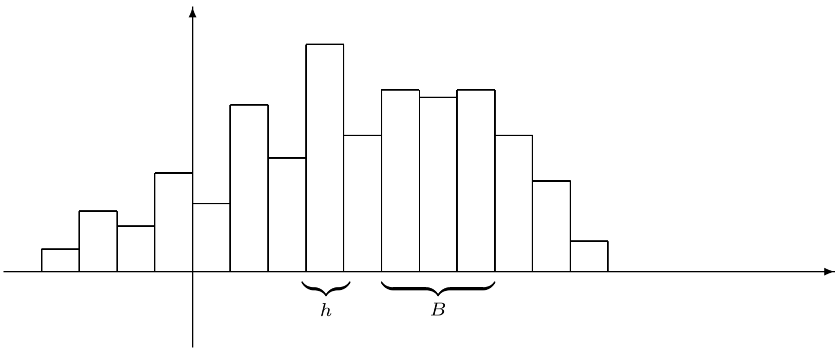
Für B gilt: $B \in [0,1]$ und B $\approx 1 \Rightarrow$ Anpassung hat hohe Qualität.

Kapitel 3

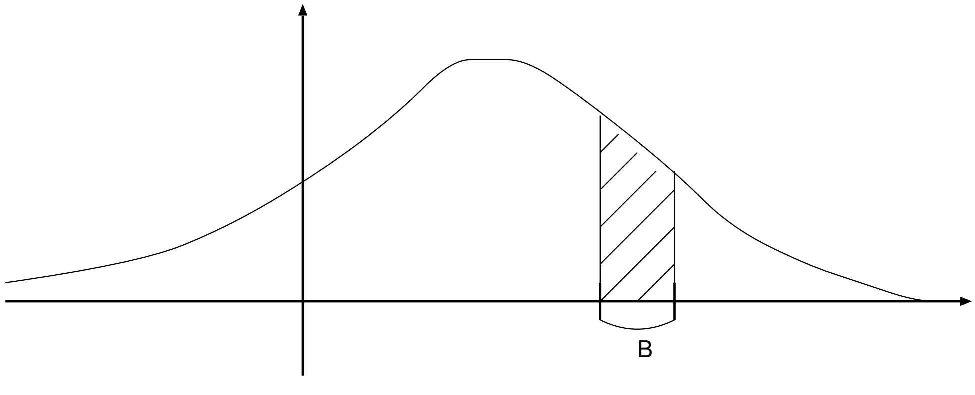
Grundbegriffe der Wahrscheinlichkeitsrechnung

3.1 Einführung

Betrachte zunächst ein Histogramm. Wählt man h hinreichend klein und erhöht den Stichprobenum-



fang, dann kommt häufig eine glatte Funktion f heraus (Dichte).



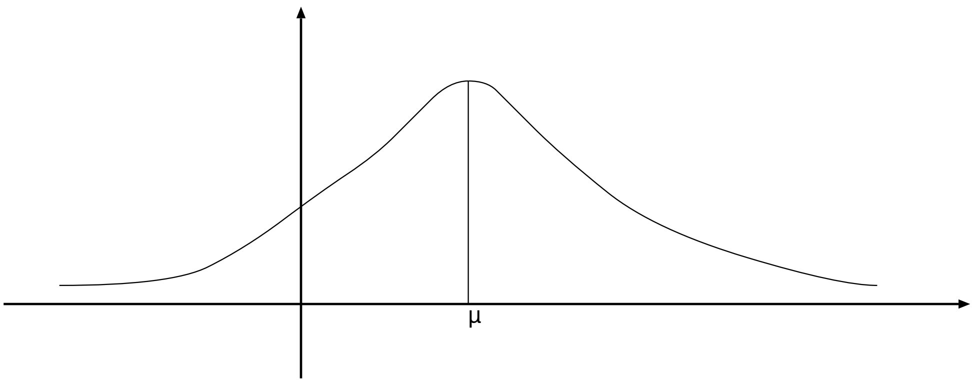
Über einer Menge B ist die Fläche unter dem Histogramm die relative Anzahl der Daten, die nach B fallen. Diese lässt sich nun durch ein kleines Gedankenexperiment zu einem intuitiven Begriff der W. formulieren. Nehmen wir an, dass die erhobenen Daten $x_1, \ldots, x_n$, die zu obigem Histogramm geführt haben, die *gemessenen Körpergrößen von n zufällig ausgewählten Personen aus einer Grundgesamtheit von N Personen* seien. Stellen wir uns weiter vor, dass wir jede einzelne Messung auf einem Zettel notiert haben und ihn anschließend in einen "Topf" gelegt haben. B entspricht einem Intervall, sagen wir $B = [a, b]$. Ziehen wir nun aus diesem Topf einen Zettel zufällig, dann entspricht die Fläche unter dem Histogramm über dem Intervall B der W., dass auf dem Zettel eine Zahl zwischen a und b steht.

Die Funktion f hatten wir erhalten, indem wir den Stichprobenumfang n vergrößert und gleichzeitig h verkleinert haben. Im Rahmen des Gedankenexperimentes interpretieren wir das Vergrößern des Stichprobenumfangs damit, dass wir alle möglichen N Körpergrößen auf Zettel geschrieben haben, die anschließend in den Topf gewandert sind. Ziehen wir jetzt einen Zettel aus den Topf zufällig heraus, dann entspricht jetzt die Fläche unter f über dem Intervall B der W., dass auf dem Zettel eine Zahl zwischen a und b steht. Da sich aber in dem Topf alle Körpergrößen der Grundgesamtheit auf Zetteln notiert befinden, können wir das "zufällige Ziehen des Zettels aus dem Topf" auch interpretieren als *Messung der Körpergröße einer zufällig ausgewählten Person der Grundgesamtheit.* D.h. die Fläche unter f über B entspricht der theoretischen Wahrscheinlichkeit (W), dass eine Messung (oder ein Experiment) einen Wert ergibt, der in B liegt.

In der Statistik nimmt man häufig an, dass ein solches f existiert, d.h. es gibt eine Funktion f, so dass die Fläche über B unter f die W. angibt, dass ein Experiment ein Ergebnis (Datum) liefert, das nach B fällt.

Beachte: $f \geq 0$ und $\int\limits_{-\infty}^{\infty} f(x)dx = 1$

Häufig ist dieses f symmetrisch: Das Symmetriezentrum μ entspricht dann dem theoretischen Lokati-



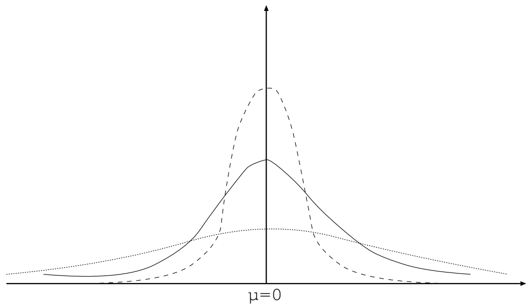
onsmaß, während $\overline{x}$ dem empirischen entspricht. Es ist zu erwarten, dass $\overline{x} = \overline{x}_n \longrightarrow \mu \quad$ für $\quad n \to \infty$. In der klassischen Statistik nimmt man an, dass f eine ganz bestimmte Gestalt (parametrische Form) hat.

Wichtige Beispiele:

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

Wir sehen: Bei kleinem σ ist die Hauptmasse mehr konzentriert als bei großem σ. Den gleichen Effekt haben wir bei der Standardabweichung

$$s_n := \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2}$$



$\cdots\cdots$: kleines σ.

In der Tat: $s_n \longrightarrow \sigma (n \to \infty)$.

Die schließende Statistik hat die Aufgabe, eine “Verbindung” zwischen Stichproben und Grundgesamtheit zu erstellen. Im Wesentlichen werden, basierend auf der Stichprobe, Hypothesen bzgl. der Grundgesamtheit überprüft (Testtheorie) oder Eigenschaften der Grundgesamtheit geschätzt (Schätztheorie).

Bemerkung 3.1.1. Vergleich zwischen dem Konzept der klassischen Analysis aus der Grundvorlesung und der Stochastik.

Mathe	Stochastik
Funktion f:	Zufallsvariable (Experiment) X:
$\underbrace{D_f}_{\text{wichtig u. bekannt}} \ni x \mapsto y := \underbrace{f(x)}_{\text{berechnet}}$	$\underbrace{D_x}_{\text{unwichtig u. häufig unbekannt}} =: \Omega \ni \omega \mapsto \underbrace{X(\omega)}_{\text{liefert das Experiment}} = x \in M$
Analyse von f: - Wertetabelle - Graph - Kurvendisk.	Analyse von X: - grafische Darstellung der Daten x - versucht zu ergründen, wie “wahrscheinlich” es ist, dass X in eine Menge B eines Systems $\mathcal{B}$ fällt.

In folgenden Beispiel werden wir anhand eines Würfelexperiments einige Formeln zum Rechnen mit Wahrscheinlichkeiten motivieren.

Beispiel 3.1.2. “Würfeln”

a) Das Experiment

$$X : \Omega \to M := \{1, 2, 3, 4, 5, 6\}$$

beschreibe das Würfeln mit einem “fairen” Würfel. M ist der <u>Ereignisraum</u>. $\mathcal{B} = \mathcal{P}(M) :=$ <u>Potenzmenge</u> von M, d.h. die Menge aller Teilmengen von M.

$$\mathbb{P}(B) := \mathbb{P}(X \in B) = W(X \in B), \quad B \in \mathcal{B}$$

die Wahrscheinlichkeit, dass X in B liegt, ist gesucht. Dies lässt sich auch durch

$$\begin{aligned} \mu : \mathcal{B} &\to [0, 1] \\ B &\mapsto \mu(B) := \mathbb{P}(X \in B) \end{aligned}$$

ausdrücken.

Da der Würfel fair ist, gilt:

$$\begin{aligned}
\mu(\{1\}) &= \mu(\{2\}) = \ldots = \mu(\{6\}) = \frac{1}{6} \\
\mu(\{1,2\}) &= \text{W. eine"1" oder "2" zu werfen} \\
&= \mu(\{1\}) + \mu(\{2\}) \\
\mu(\emptyset) &= \text{W. keine Zahl zwischen "1" und "6"} \\
&= 0 \\
\mu(M) &= \text{W. eine Zahl zwischen "1" und "6"} \\
&= 1
\end{aligned}$$

Bezeichnung: $\complement B :=$ { x ∈ M : x ∉ B }

$$\begin{aligned}
\mu(\complement\{1,2\}) &= \mu(M \setminus \{1,2\}) \\
&= \text{W. weder "1" noch "2"} \\
&= \mu(\{3,4,5,6\}) \\
&= \frac{4}{6} \\
&= 1 - \frac{2}{6} \\
&= \mu(M) - \mu(\{1,2\})
\end{aligned}$$

b) Das Experiment

$$X : \Omega \longrightarrow M^2 := \{(x,y) : x \in M, y \in M\}$$

beschreibt "zweimal nacheinander würfeln". Die relevanten Mengen sind wieder aus $\mathcal{B} = \mathcal{P}(M^2)$, der Potenzmenge von M^2. Die Mengenfunktion μ hat die Form:

$$\begin{aligned}
\mu : \quad \mathcal{B} &\rightarrow [0,1] \\
B &\mapsto \mu(B) := \frac{|B|}{|M^2|} = \frac{|B|}{6^2}
\end{aligned}$$

Seien jetzt

$$\begin{aligned}
B &:= \{(1,x) \in M^2 : x \neq 2\} && \text{- 1-ter Wurf eine "1" und 2-ter Wurf keine "2"} \\
A &:= \{(1,x) : x \in M\} \\
C &:= \{(1,2)\}.
\end{aligned}$$

Offensichtlich gilt jetzt $C \subset A$ und $B = A \setminus C$. Ferner ergibt sich aus der Definition von μ:

$$\mu(A \setminus C) = \frac{|A \setminus C|}{6^2} = \frac{5}{6^2}$$

Andererseits erhalten wir aus der Definition von μ auch:

$$\begin{aligned}
\mu(A) &= \frac{|A|}{6^2} = \frac{6}{6^2} \\
\mu(C) &= \frac{|C|}{6^2} = \frac{1}{6^2}
\end{aligned}$$

Zusammengefasst ergibt sich also aus $C \subset A$ dann

$$\mu(A \backslash C) = \frac{5}{6^2} = \frac{6}{6^2} - \frac{1}{6^2} = \mu(A) - \mu(C)$$

c) Fortsetzung zu (b)

$$\begin{array}{rcll} B & := & \{(x,6) : x \in \{1,2\}\} & \text{- erst “1” oder “2”, dann “6”} \\ A & := & \{(x,y) : x \in 1,2 \,\text{und}\, y \in M\} & \text{- erst “1” oder “2”, dann egal} \\ C & := & \{(x,6) : x \in M\} & \text{- erst egal, dann “6”} \end{array}$$

Offensichtlich gilt: $B = A \cap C$. Wenden wir wieder die Definition von μ an, so erhalten wir:

$$\mu(B) = \frac{2}{6^2}\ , \quad \mu(A) = \frac{12}{6^2}\ , \quad \mu(C) = \frac{6}{6^2}$$

Die Ereignisse A und C sind voneinander unabhängig. Dieses spiegelt sich im folgenden Sachverhalt wieder.

$$\mu(A \cap C) = \frac{1}{18} = \frac{1}{3}\frac{1}{6} = \mu(A)\mu(C)$$

Definition 3.1.3. Seien $\Omega \neq \emptyset$ und $\mathcal{A} \subset \mathcal{P}(\Omega)$ ein Mengensystem mit den Eigenschaften

a) $\Omega \in \mathcal{A}$
b) $A \in \mathcal{A} \Rightarrow \complement A \in \mathcal{A}$
c) $A_n \in \mathcal{A}$ für alle $n \in \mathbb{N} \Rightarrow \bigcup_{n=1}^{\infty} A_n \in \mathcal{A}$.

Ferner sei $\mathbb{P} : \mathcal{A} \to [0,1]$ eine Abbildung mit den Eigenschaften:

$$\begin{array}{ll} \text{(i)} & \mathbb{P}(\Omega) = 1 \\ \text{(ii)} & \mathbb{P}(A) \geq 0 \quad \text{für alle}\, A \in \mathcal{A} \\ \text{(iii)} & \mathbb{P}\left(\bigcup_{n=1}^{\infty} A_n\right) = \sum_{n=1}^{\infty} \mathbb{P}(A_n) \quad \text{für alle Folgen}\, (A_n)_{n\in\mathbb{N}} \subset \mathcal{A} \quad \text{für die} \\ & A_n \cap A_m = \emptyset\,, \quad \text{falls}\, n \neq m. \end{array}$$

Man nennt dann $(\Omega, \mathcal{A}, \mathbb{P})$ einen Wahrscheinlichkeitsraum, $\mathbb{P}$ ein Wahrscheinlichkeitsmaß und das Mengensystem $\mathcal{A}$ eine σ-Algebra.

Beispiel 3.1.4. $(M, \mathcal{P}(M), \mu)$ aus Beispiel 3.1.2 sind Wahrscheinlichkeitsräume.

Definition 3.1.5. Unter einer Zufallsvariable oder auch Experiment X versteht man eine Abbildung

$$\begin{array}{rrcl} X: & (\Omega, \mathcal{A}, \mathbb{P}) & \longrightarrow & (M, \mathcal{B}) \\ & \omega & \mapsto & X(\omega) \end{array}$$

von dem Wahrscheinlichkeitsraum $(\Omega, \mathcal{A}, \mathbb{P})$ in den Ereignisraum $(M, \mathcal{B})$ mit der Eigenschaft, dass

$$\{\omega \,:\, X(\omega) \in B\} \in \mathcal{A}$$

für jede Menge B aus (der σ-Algebra) dem Mengensystem $\mathcal{B}$. Durch

$$\begin{array}{rrcl} \mu: & \mathcal{B} & \to & [0,1] \\ & B & \mapsto & \mu(B) := \mathbb{P}(X \in B) \end{array}$$

wird durch die Zufallsvariable X ein neuer W.Raum $(M, \mathcal{B}, \mu)$ definiert. μ heißt die Verteilung der Zufallsvariable X.

3.2 Kombinatorik

Beispiel 3.2.1. Setze $\Omega := \{\omega_1, \omega_2, ..., \omega_n\}$ und $\mathcal{A} := \mathcal{P}(\Omega)$. Dann ist $\mathcal{A}$ eine σ-Algebra. Einen W. Raum erhält man dann durch Festlegen von

$$\mathbb{P}\left(\{\omega_i\}\right) := p_i \qquad \text{, i=1,..,n}$$

wobei $p_i \geq 0$ vorgegebene Werte mit

$$\sum_{i=1}^{n} p_i = 1, \quad \mathbb{P}(A) = \sum_{w_i \in A} \mathbb{P}\left(\{w_i\}\right) = \sum_{w_i \in A} p_i$$

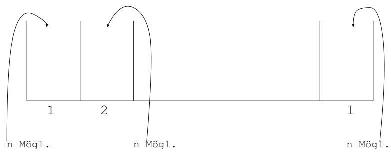
Spezialfall: $p_1 = p_2 = ... = p_n \equiv p$, d.h. alle Elementarwahrscheinlichkeiten sind gleich. Dann gilt:

$$\begin{aligned} 1 &= \sum_{i=1}^{n} p_i = \sum_{i=1}^{n} p = np \\ p &= \frac{1}{n} = \frac{1}{|\Omega|} \quad \text{d.h. } \underline{\text{Laplace-Modell}} \\ \mathbb{P}(A) &= \sum_{w_i \in A} p_i = \sum_{w_i \in A} \frac{1}{|\Omega|} = \frac{|A|}{|\Omega|} \\ &= \frac{|\text{für A günstige Fälle}|}{|\text{mögliche Fälle}|} \end{aligned}$$

Um solche W. zu berechnen, muss man die Anzahl der Elemente von Mengen A, d.h. $|A|$ bestimmen. Dabei sind kombinatorische Formeln hilfreich.

Im Folgenden nehmen wir an, dass $Z \neq \emptyset$ mit $|Z| = n$ und $n \in \mathbb{N}$.

a) $\overline{V}_{n,l}$ - Anzahl geordneter Stichproben aus Z vom Umfang l mit Wiederholungen. Offensichtlich

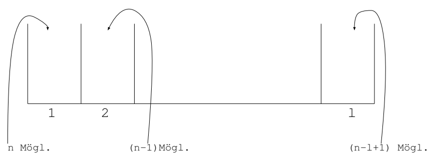


erhalten wir:

$$\overline{V}_{n,l} = n^l$$

Beispiel: Totto

b) $V_{n,l}$ - Anzahl geordneter Stichproben aus Z vom Umfang l ohne Wiederholungen. In diesem Fall



ergibt sich:

$$\begin{aligned} V_{n,l} &= n(n-1)...(n-l+1) \\ &= \frac{n(n-1)...(n-l+1)(n-l)...1}{(n-l)(n-l-1)...1} \\ &= \frac{n!}{(n-l)!} \end{aligned}$$

Falls $n = l$ erhalten wir als Spezialfall

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

die Anzahl von möglichen Anordnungen, <u>Permutation</u> genannt, von n Elementen.

c) $C_{n,l}$ - Anzahl ungeordneter Stichproben aus Z vom Umfang l ohne Wiederholungen. Diese Anzahl entspricht der Anzahl der l-elementigen Teilmengen einer n-elementigen Menge.

$$\begin{aligned} l! \cdot C_{n,l} &= V_{n,l} \\ \Rightarrow C_{n,l} &= \frac{V_{n,l}}{l!} = \frac{n!}{(n-l)!l!} \\ &= \binom{n}{l} \quad := \underline{\text{Binomialkoeffizienten}} \end{aligned}$$

Beispiel: Lotto $\binom{49}{6}$

d) $\overline{C}_{n,l}$ - Anzahl ungeordneter Stichproben aus Z vom Umfang l mit Wiederholungen. Diese Anzahl ist gleichbedeutend mit der Anzahl der Möglichkeiten, l Chips auf n Fächer zu verteilen.

$$\overline{C}_{n,l} = \binom{n+l-1}{l}$$

z.B. $n = 4$, $l = 3$.
Wieviel Möglichkeiten gibt es nun? Betrachte dazu die Anordnung: 3 $*$ und 3 $|$, i.F. als Trenn-

	$**$		$*$		$\Leftrightarrow$ 2K in F1 und 1K in F3.
Fach	1	2	3	4	

wand bezeichnet. Dann erhalten wir:

$$* * | \quad | * | \Leftrightarrow \text{2K in F1 und 1K in F3}$$

Insgesamt gibt es $4 - 1 + 3 = (n - 1 + l)!$ Anordnungen. Jede Vertauschung der Trennwände, 3! bzw. $(n-1)!$, und jede Vertauschung der $*$, 3! bzw. $l!$, führt zur gleichen Belegung. Also:

$$\Rightarrow \overline{C}_{n,l} = \frac{(n-1+l)!}{(n-1)!\,l!} = \binom{n-1+l}{l}$$

e) $P_{n,n_1,n_2,...,n_k}$ - Anzahl unterschiedlicher Anordnungen der Länge n, wenn von den n Elementen n_1 identisch, n_2 identisch, ..., n_k identisch sind.

$$P_n, n_1, ..., n_k = \frac{n!}{n_1!...n_k!}$$

Beispiel: $n = 10$ Kugeln, $n_1 = 3$ weiße, $n_2 = 2$ schwarze, $n_3 = 4$ blaue und eine rote Kugel:

$$P_{10,3,2,4} = \frac{10!}{3! * 2! * 4!} = \frac{5 * 6 * 7 * 8 * 9 * 10}{3! * 2!} = 5 * 7 * 4 * 9 * 10$$

Beispiel 3.2.2. Berechnungen von W. im Laplace-Modell:

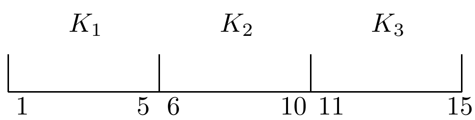
a) W. genau 3 Richtige im Lotto (6 aus 49) (Zusatzzahl nicht berücksichtigt)

$$\Omega := \text{Menge der 6-elementigen Teilmengen von } \{1, ..., 49\}$$
$$\Rightarrow \quad p = \frac{1}{|\Omega|} = \frac{1}{\binom{49}{6}}$$
$$A = \{\omega \in \Omega : \omega \text{ enthält genau 3 Richtige}\}$$
$$|A| = \binom{6}{3}\binom{43}{3}$$
$$\Rightarrow \mathbb{P}(A) = \frac{|A|}{|\Omega|} = \frac{\binom{6}{3}\binom{43}{3}}{\binom{49}{6}} = \frac{\frac{6!}{3!3!}\frac{43!}{3!40!}}{\frac{49!}{6!43!}} = ... \approx 0.018$$

b) "Qualitätskontrolle" Warenlieferung von $m-$Stück enthalte k schlechte Stücke. Mit welcher W. hat eine Stichproben vom Umfang n genau j schlechte Stücke?

$$|\Omega| = \binom{m}{n} \quad \text{ungeordnete Stichprobe vom Umfang } n \text{ ohne Wiederholungen bei } m-\text{elementiger Menge.}$$
$$A := \{\omega \in \Omega : \omega \text{ enthält genau } j \text{ schlechte Stücke}\}$$
$$\mathbb{P}(A) = \frac{|A|}{|\Omega|} = \frac{\binom{k}{j}\binom{m-k}{n-j}}{\binom{m}{n}}$$

c) 15 neue Schüler sollen gleichmäßig auf drei Klassen verteilt werden. Unter den 15 sind drei "Schlauköpfe". Mit welcher W. bekommt jede Klasse einen?



Ω - unterschiedliche Klassenbelegungen

$$\Rightarrow |\Omega| = \frac{15!}{5!5!5!} \quad \begin{array}{l} \rightarrow \text{ Permutationen der 15 Schüler} \\ \rightarrow \text{ Permutationen innerhalb der Fächer } K_1, K_2, K_3 \end{array}$$

$A = \{\omega \in \Omega : \text{jeweils 1 "Schlaukopf" pro Klasse } K_i\}$

3! Möglichkeiten, die 3 "Schlauköpfe auf die Klassen zu verteilen

$$\Rightarrow |A| = \frac{3!12!}{4!4!4!}$$
$$\Rightarrow \mathbb{P}(A) = \frac{\frac{3!12!}{4!4!4!}}{\frac{15!}{5!5!5!}}$$

3.3 Bedingte Wahrscheinlichkeit, Unabhängigkeit

Beispiel 3.3.1. Raucher in der Oberstufe eines Gymnasiums verteilt nach Geschlecht

	weibl.	männl.	$\sum$
Raucher	32	30	62
Nichtraucher	23	15	38
$\sum$	55	45	100

S - zufällig ausgewählte Schüler.

$$\Rightarrow \mathbb{P}(S = R) = \frac{62}{100}$$

Falls wir wissen, dass der Schüler weiblich ist, ergibt sich:

$$\Rightarrow W(S = R) = \frac{32}{55}$$

Wie hängt diese W. mit $(\Omega, \mathcal{A}, \mathbb{P})$ zusammen?

$$A := \{S = R\} \quad , \ B := \{S = W\}$$

Wenn die Bedingung B erfüllt ist, d.h. "ausgewählter Schüler ist weiblich", d.h. S $\in$ B, kann das Ereignis

"ausgewählter Schüler ist Raucher", $S \in A$

nur eintreten, wenn

$$\begin{aligned} & S \in A \cap B \\ \underset{imLapl.M}{\Rightarrow} \quad & \mathbb{P}(A|B) = \text{"W. von A unter der Bedingung B"} \\ & = \frac{|A \cap B|}{|B|} = \frac{|A \cap B|}{|\Omega|} \frac{|\Omega|}{|B|} \\ & = \mathbb{P}(A \cap B) \frac{1}{\mathbb{P}(B)} \\ & = \frac{32}{100} \frac{100}{55} = \frac{32}{55} \end{aligned}$$

Definition 3.3.2. Sei $(\Omega, \mathcal{A}, \mathbb{P})$ ein Wahrscheinlichkeitsraum, $B \in \mathcal{A}$ mit $\mathbb{P}(B) > 0$. Dann wird durch

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

die bedingte Wahrscheinlichkeit von A unter B definiert.

Bemerkung 3.3.3. $(\Omega,\ \mathcal{A},\ \mathbb{P}(\cdot \,|\, B))$ ist selbst wieder ein W.-Raum.

Beispiel 3.3.4. Ein Produkt wird auf zwei unterschiedlichen Maschinen gefertigt:

$$\begin{aligned} M_1 \quad & - \quad \text{fertigt 5000/Tag mit 10\% Ausschuss} \\ M_2 \quad & - \quad \text{fertigt 6000/Tag mit 8\% Ausschuss} \end{aligned}$$

Wie groß ist die W., dass ein zufällig ausgewähltes Produkt beim Warenausgang defekt ist?

$$\begin{aligned} A &:= \{\omega : \omega \text{ auf } M_1 \text{ produziert}\}, \\ B &:= \{\omega : \omega \text{ auf } M_2 \text{ produziert}\} \\ \Omega &:= A \cup B \quad \text{(Tagesproduktion)} \\ D &:= \{\omega : \text{Ware ist defekt}\} \\ (\Omega, \mathcal{A}, \mathbb{P}) &\ - \ \text{Laplace-Modell} \end{aligned}$$

Dann erhalten wir:

$$\mathbb{P}(A) = \frac{5000}{5000 + 6000} = \frac{5}{11}$$

und damit dann weiter:

$$\mathbb{P}(B) = 1 - \mathbb{P}(A) = \frac{6}{11}$$

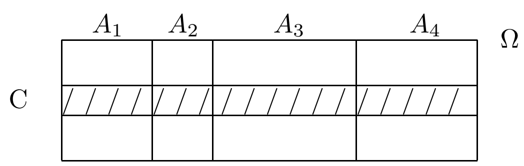
Es ist ferner bekannt, dass $\mathbb{P}(D|A) = 0.1$ und $\mathbb{P}(D|B) = 0.08$. Daraus folgt abschließend:

$$\begin{aligned}
\mathbb{P}(D) &= \mathbb{P}(D \cap \Omega) = \mathbb{P}(D \cap (A \cup B)) \\
&\underset{deMorgan}{=} \mathbb{P}((D \cap A) \cup (D \cap B)) = \mathbb{P}(D \cap A) + \mathbb{P}(D \cap B) \\
&= \frac{\mathbb{P}(D \cap A)}{\mathbb{P}(A)}\mathbb{P}(A) + \frac{\mathbb{P}(D \cap B)}{\mathbb{P}(B)}\mathbb{P}(B) \\
&= \mathbb{P}(D|A)\mathbb{P}(A) + \mathbb{P}(D|B)\mathbb{P}(B) \\
&= 0.1\frac{5}{11} + 0.08\frac{6}{11} \\
&\approx 0.089
\end{aligned}$$

Satz 3.3.5. Formel von der totalen Wahrscheinlichkeit Sei $(\Omega, \mathcal{A}, \mathbb{P})$ ein Wahrscheinlichkeitsraum und $A_1, \ldots, A_n$, jeweils $\in \mathcal{A}$, paarweise disjunkte Ereignisse (d.h. $i \neq j \Rightarrow A_i \cap A_j = \emptyset$) mit $\bigcup\limits_{i=1}^{n} A_i = \Omega$ und $\mathbb{P}(A_i) > 0$ für alle $i = 1, ..., n$. Für alle $C \in \mathcal{A}$ gilt dann:

$$\mathbb{P}(C) = \sum_{i=1}^{n} \mathbb{P}(C|A_i)\mathbb{P}(A_i)$$

Beweis. "Zur Vorstellung" Zunächst folgt aus den Voraussetzungen, dass



$$C = \bigcup_{i=1}^{n} (A_i \cap C).$$

Mit dieser Zerlegung erhalten wir weiter:

$$\begin{aligned}
\mathbb{P}(C) &= \mathbb{P}\left(\bigcup_{i=1}^{n} (C \cap A_i)\right) \\
&\underset{C \cap A_i p.d.}{=} \sum_{i=1}^{n} \mathbb{P}(C \cap A_i) \\
&= \sum_{i=1}^{n} \mathbb{P}(C|A_i)\mathbb{P}(A_i)
\end{aligned}$$

□

Satz 3.3.6. Bayessche Formel Unter den Voraussetzungen von Satz 3.3.5 gilt für $C \in \mathcal{A}$ mit $\mathbb{P}(C) > 0$:

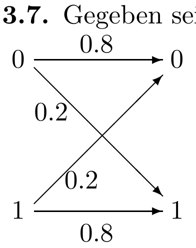
$$\mathbb{P}(A_k|C) = \frac{\mathbb{P}(C|A_k)\mathbb{P}(A_k)}{\sum\limits_{i=1}^{n} \mathbb{P}(C|A_i)\mathbb{P}(A_i)} \quad , \text{ k = 1, ..., n}$$

Beweis.

$$\mathbb{P}(A_k|C) \quad = \quad \frac{\mathbb{P}(A_k \cap C)}{\mathbb{P}(C)} = \frac{\mathbb{P}(C|A_k)\mathbb{P}(A_k)}{\sum\limits_{i=1}^{n} \mathbb{P}(C|A_i)\mathbb{P}(A_i)}$$

Interpretation: Arzt beobachtet Symptom C. C kann durch Krankheit A_1 oder A_2.... oder A_n verursacht sein. $\mathbb{P}(A_k|C)$ - W. dass für Symptom C die Krankheit A_k verantwortlich ist. □

Beispiel 3.3.7. Gegeben sei ein binärer Nachrichtenkanal mit Übertragungsschema



Sender sendet nur 0000000 oder 1111111 und zwar mit W. 0.1 bzw. 0.9

Es wird 0100110 empfangen. Wie groß ist die W., dass $0\dots0$ gesendet wurde?

$$\begin{aligned}
\mathbb{P}(0\dots0|0100110) &= \quad ? \\
&\underset{(3.3.6)}{=} \frac{\mathbb{P}(0100110|0\dots0)\mathbb{P}(0\dots0)}{\mathbb{P}(0\dots0)\mathbb{P}(0100110|0\dots0) + \mathbb{P}(1\dots1)\mathbb{P}(0100110|1\dots1)} \\
&= \frac{(0.8^4 0.2^3)0.1}{0.1(0.8^4 0.2^3) + 0.9(0.2^4 0.8^3)} \\
&= \frac{4}{13}
\end{aligned}$$

d.h. trotz “0” Mehrheit wird man 0100110 als $1\dots1$ zu dekodieren haben.

Beispiel 3.3.8. Gegeben sei eine Urne mit 2 weißen und 3 schwarzen Kugeln. Definiere die Ereignisse:

$$\begin{aligned}
A: &\quad - \quad \text{“Beim ersten Ziehen eine weiße Kugel”} \\
B: &\quad - \quad \text{“Beim zweiten Ziehen eine schwarze Kugel”}
\end{aligned}$$

(i) Ziehen ohne Zurücklegen

$$\Rightarrow \mathbb{P}(A \cap B) \quad = \quad \mathbb{P}(A)\mathbb{P}(B|A) = \frac{2}{5}\frac{3}{4} = \frac{3}{10}$$

(ii) Ziehen mit Zurücklegen

$$\Rightarrow \mathbb{P}(A) = \frac{2}{5}, \mathbb{P}(B) = \frac{3}{5}, \mathbb{P}(A \cap B) = \frac{6}{25}$$

$$\text{also} \quad \mathbb{P}(A \cap B) = \frac{6}{25} = \frac{2}{5}\frac{3}{5} = \mathbb{P}(A)\mathbb{P}(B)$$

Beim Ziehen mit Zurücklegen hat die erste Ziehung "keinerlei Einfluss" auf die zweite Ziehung, d.h. beide Ziehungen sind unabhängig voneinander.

Definition 3.3.9. Sei $(\Omega, \mathcal{A}, \mathbb{P})$ ein W.-Raum und $A_1, \ldots, A_n$ n Ereignisse, jeweils $\in \mathcal{A}$. Diese Ereignisse heißen (ℙ-)unabhängig, wenn für jede Menge $I = \{ i_1, i_2, ..., i_k \} \subset \{ 1, ...,n \}$ mit $1 \leq k \leq n$ gilt:

$$\begin{aligned} \mathbb{P}\left(\bigcap_{l=1}^{k} A_{i_l}\right) &= \mathbb{P}(A_{i_1} \cap A_{i_2} \cap .. \cap A_{i_k}) \\ &= \mathbb{P}(A_{i_1}) \cdot \mathbb{P}(A_{i_2}) \cdots \mathbb{P}(A_{i_k}) \end{aligned}$$

Bemerkung 3.3.10. Seien A, B unabhängige Ereignisse und $\mathbb{P}(B) > 0$. Dann erhalten wir:

$$\mathbb{P}(A|B) = \mathbb{P}(A)$$

Beweis.

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

□

Definition 3.3.11. Seien $X_i : (\Omega, \mathcal{A}, \mathbb{P}) \to (M, \mathcal{B})$, für $i = 1, ..., n$, n Experimente bzw. (Zufallsvariable). Dann nennt man $X_1, \ldots, X_n$ unabhängig, falls für beliebige Mengen $B_1, \ldots, B_n$, jeweils $\in \mathcal{B}$, gilt:

$$A_1 := \{X_1 \in B_1\}, A_2 := \{X_2 \in B_2\}, \ldots, A_n := \{X_n \in B_n\}$$

sind unabhängig gemäß Definition 3.3.9.

3.4 Verteilungen und Verteilungsfunktionen

Unter Definition 3.1.5 hatten wir die Verteilung einer Zufallsvariable definiert.

$$\begin{aligned} X &: (\Omega, \mathcal{A}, \mathbb{P}) \to (M, \mathcal{B}) \\ \\ \Rightarrow \mu &: \mathcal{B} \to [0, 1] \\ & \quad B \mapsto \mu(B) := \mathbb{P}\left(\{\omega \in \Omega : X(\omega) \in B\}\right) \end{aligned}$$

I.f. $M = \mathbb{R}$ und $\sigma-$Algebra $\mathcal{B}$ stets so, dass alle Typen von Intervallen in $\mathcal{B}$ liegen.

Um über das Verhalten der Zufallsvariable bzw. des Experimentes Aussagen machen zu können, benötigen wir die Verteilung. Verteilungen sind aber in der Regel recht komplizierte Mengenfunktionen. Um so erstaunlicher ist es, dass sich jede Verteilung durch die zugehörige Verteilungsfunktion, einer Funktion wie wir sie aus der Mathematik kennen, in eindeutiger Weise beschreiben lässt.

Definition 3.4.1. Sei

$$X : (\Omega, \mathcal{A}, \mathbb{P}) \to (\mathbb{R}, \mathcal{B})$$

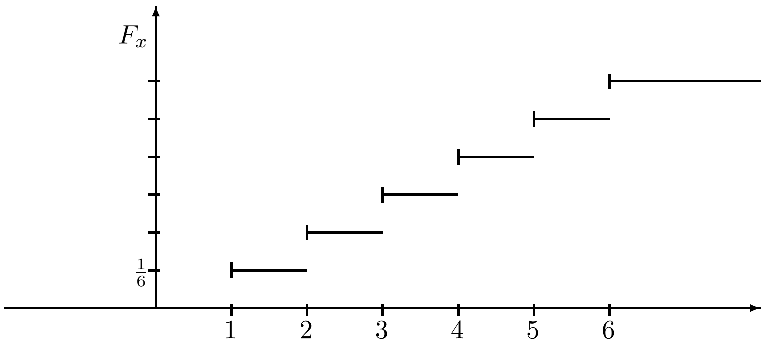
eine Zufallsvariable. Dann wird durch

$$\begin{aligned} F_X &: \mathbb{R} \to [0, 1] \\ & \quad x \mapsto F_X(x) := \mathbb{P}\left(\{w \in \Omega : X(w) \leq x\}\right) = \mathbb{P}(X \leq x) \end{aligned}$$

die Verteilungsfunktion von X definiert.

Beispiel 3.4.2. Betrachte das Experiment "faires Würfeln", das durch die Zufallsvariable $X : (\Omega, \mathcal{A}, \mathbb{P}) \to (\mathbb{R}, \mathcal{B})$ beschrieben werde. Wir erhalten für die Verteilungsfunktion zu diesem Experiment:

$$\begin{aligned} F_X(1) &= \mathbb{P}(X \leq 1) = \frac{1}{6} \\ F_X(2) &= \mathbb{P}(X \leq 2) = \frac{2}{6} \\ F_X(2,5) &= \mathbb{P}(X \leq 2,5) = \frac{2}{6} \\ \vdots \quad & \vdots \quad \vdots \end{aligned}$$



Bemerkung 3.4.3. Eine Verteilungsfunktion F besitzt die folgenden Eigenschaften:

(i) F monoton wachsend
d.h. $x_1 \leq x_2 \Rightarrow F(x_1) \leq F(x_2)$

(ii) $\lim\limits_{x \to \infty} F(x) = 1$ und $\lim\limits_{x \to -\infty} F(x) = 0$

(iii) $\lim\limits_{x \downarrow x_0} F(x) = F(x_0)$
d.h. F ist rechtseitig stetig

(iv) $\lim\limits_{x \uparrow x_0} F(x)$ existiert
d.h. F besitzt linksseitige Limiten

Wir werden im Folgenden zwei Typen von Verteilungen, die hauptsächlich für die Statistik relevant sind, näher studieren. Als erstes betrachten wir die so genannten diskreten Verteilungen.

Definition 3.4.4. Eine Zufallsvariable $X : (\Omega, \mathcal{A}, \mathbb{P}) \to (\mathbb{R}, \mathcal{B})$, die nur abzählbar viele unterschiedliche Werte annehmen kann, wird diskret genannt. Die zugehörige Verteilung nennt man dann eine diskrete Verteilung. (Abzählbar bedeutet entweder endlich viele oder aber unendlich viele unterschiedliche Werte, die man, wie die natürlichen Zahlen, "aufzählen" kann.)

Wir listen jetzt die wesentlichen diskreten Verteilungen mit entsprechenden Anwendungen auf.

Beispiel 3.4.5. Schrauben werden bis zu einer vorgegebenen Tiefe eingeschraubt. Danach wird an der Schraube gezogen, bis sie ausreißt. Dabei treten die beiden Fälle "Gewinderiss" bzw. "Bolzenriss" auf.

$$X(\omega) := \begin{cases} 1 & \text{Bolzenriss} \\ 0 & \text{Gewinderiss} \end{cases}$$

Als Verteilung von X erhalten wir:

$$\begin{aligned}\mu(\{1\}) &= \mathbb{P}(X=1) = p \\ \mu(\{0\}) &= 1-\mu(\{1\}) = 1-p\end{aligned}$$

Die Verteilungsfunktion hat die Gestalt:

$$F_X(x) \equiv F(x) = \begin{cases} 0 & x < 0 \\ 1-p & 0 \leq x < 1 \\ 1 & x \geq 1 \end{cases}$$

Experimente mit nur zwei Ausgängen (0 bzw. 1) werden Bernoulli-Experimente genannt. Die zugehörige Verteilung nennt man entsprechend Bernoulli-Verteilung. p heißt die Erfolgswahrscheinlichkeit von X. Um anzudeuten, dass ein Experiment bzw. eine Zufallsvariable X eine Bernoulli-Verteilung mit der Erfolgswahrscheinlichkeit p besitzt, schreiben wir kurz:

$$X \sim BERN(p)$$

Beispiel 3.4.6. Wiederholt man das Schraubenexperiment $n-$mal unabhängig voneinander, so erhält man $x_1, ..., x_n$ unabhängige Zufallsvariablen, die identisch verteilt sind nach $BERN(p)$. Betrachte nun

$$Y := \sum_{i=1}^{n} X_i$$

Y kann nur die Werte $k = 0, 1, ..., n$ annehmen, ist also auch eine diskrete Zufallsvariable. Zur Berechnung der Verteilung von Y betrachte zunächst:

$$\begin{aligned}&\mathbb{P}\left(\{X_1 = 1\} \cap \ldots \cap \{X_k = 1\} \cap \{X_{k+1} = 0\} \cap \ldots \cap \{X_n = 0\}\right) \\ &\qquad = \mathbb{P}(X_1 = 1) \cdots \mathbb{P}(X_k = 1)\mathbb{P}(X_{k+1} = 0 \cdots \mathbb{P}(X_n = 0) \\ &\qquad = \underbrace{p \cdots p}_{k} \cdot \underbrace{(1-p) \cdots (1-p)}_{n-k} \\ &\qquad = p^k (1-p)^{n-k}\end{aligned}$$

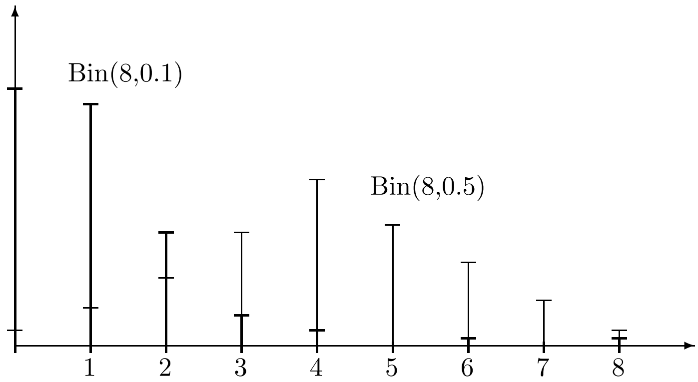
Für $Y = k$ ist es nicht wichtig, welche der Bernoulli-Experimtente den Ausgang 1 haben, solagne es nur genau $k-$viele sind. Es gibt $\binom{n}{k}$ unterschiedliche Möglichkeiten von den n Experimenten k auszuwählen.

$$\Rightarrow \mathbb{P}(Y = k) = \binom{n}{k} p^k (1-p)^{n-k} \quad \text{, k=0,1,...,n}$$

Die Verteilungsfunktion von Y hat damit die Gestalt:

$$\mathbb{P}(Y \leq x) = \sum_{k=0,\, k \leq x}^{n} \binom{n}{k} p^k (1-p)^{n-k} \quad \text{, x} \in \mathbb{R}$$

Experimente, die als Summe von n unabhängigen identischen $BERN(p)$ Experimenten definiert sind, heißen Binomial-Experimente zu den Parametern n und p. Analog nennt man die zugehörige Verteilung Binomial-Verteilung. Wir schreiben abgekürzt $Y \sim BIN(n, p)$, um anzudeuten, dass das Experiment Y eine Binomial-Verteilung zu den Parametern n und p besitzt.



Die Berechnung der Binomial-Koeffizienten, $\binom{n}{k}$, ist insbesondere dann mit großem Rechenaufwand verbunden, wenn n groß wird. Aus diesem Grund ist man an einer Approximation (Näherung) interessiert. Diese wird durch den Poissonschen Grenzwertsatz (PGS) geliefert.

Satz 3.4.7. Poissonscher Grenzwertsatz. Gegeben sei $(X_n)_{n\in\mathbb{N}}$ eine Folge von binomial-verteilten Zufallsvariablen mit $X_n \sim BIN(n, p_n)$. Ferner setzen wir voraus, dass

$$np_n \underset{n\to\infty}{\longrightarrow} \lambda \in \mathbb{R}^+$$

erfüllt ist. Dann gilt für $k = 0, 1, \ldots$

$$\mathbb{P}(X_n = k) \underset{n\to\infty}{\longrightarrow} \exp(-\lambda)\frac{\lambda^k}{k!}$$

Beweis.

$$\begin{aligned}
\mathbb{P}(X_n = k) &= \binom{n}{k} p_n^k (1-p_n)^{n-k} \\
&= \frac{n!}{k!(n-k)!} p_n^k (1-p_n)^n (1-p_n)^{-k} \\
&= \underbrace{\frac{(np_n)^k}{k!}}_{\to \frac{\lambda^k}{k!}} \underbrace{\frac{n!}{n^k(n-k)!}}_{\to 1} \underbrace{\left(1 - \frac{np_n}{n}\right)^n}_{\to e^{-\lambda}} \underbrace{(1-p_n)^{-k}}_{\to 1}
\end{aligned}$$

denn

$$\begin{aligned}
\frac{n!}{n^k(n-k)!} &= \frac{n(n-1)\cdots(n-(k-1))}{nn\cdots n} = 1\frac{n-1}{n}\frac{n-2}{n}\cdots\frac{n-(k-1)}{n} \\
&= \left(1-\frac{1}{n}\right)\left(1-\frac{2}{n}\right)\cdots\left(1-\frac{k-1}{n}\right) \underset{n\to\infty}{\longrightarrow} 1
\end{aligned}$$

□

Bemerkung 3.4.8. Ist p von $X \sim BIN(n,p)$ klein und n groß, so besagt PGS, dass die (lästige) $BIN(n,p)-$Verteilung durch

$$\mathbb{P}(X = k) \approx \exp(-np)\frac{(np)^k}{k!}$$

approximiert werden kann.

Beispiel 3.4.9. Man beachte, dass durch

$$\mathbb{P}(X = k) \equiv p_k := \exp(-\lambda)\frac{\lambda^k}{k!}, \quad k = 0, 1, \ldots$$

eine diskrete Verteilung auf den Zahlen $0, 1, \ldots$ definiert wird, denn $p_k > 0$ und $\sum_{k\geq 0} p_k = 1$. Diese diskrete Verteilung wird Poisson-Verteilung zum Parameter $\lambda > 0$ genannt. Besitzt eine Zufallsvariable X eine solche Verteilung, dann schreiben wir kurz $X \sim POIS(\lambda)$.

Eine Zufallsvariable, die zählt, wie oft ein relativ seltenes Ereignis auftritt, kann oft als Poisson-verteilt zum Parameter $\lambda > 0$ angenommen werden. Z.B. Tote durch Hufschlag in 10 preußischen Kavallerieregimenten während 20 Jahren. Insgesamt haben wir hier $10 \cdot 20 = 200$ Beobachtungen, wobei eine Beobachtung die Anzahl der Toten in einem Regiment in einem Jahr darstellt. Die Zufallsvariable $X-$Anzahl der Toten pro Jahr und Regiment wird hier also 200-mal beobachtet.

x	Anzahl von Beobachtungen mit x Toten
0	109
1	65
2	22
3	3
4	1
≥ 5	0

Beispiel 3.4.10. Betrachte eine Warenlieferung, die $m-$Stücke enthalte. Nehmen wir an, dass von diesen $k-$fehlerhaft sind. Wir ziehen nun eine Stichprobe ohne Zurücklegen vom Umfang n und zählen, wie viele dieser $n-$Stücke fehlerhaft sind. Die Zufallsvariable X gibt diese Anzahl an. Natürlich muss hier $n \leq m$ sein und X kann nur die ganzzahlige Werte von $\max(n - (m - k), 0)$ bis einschließlich, $\min(n, k)$ annehmen. Sei j solch ein Wert, dann gilt:

$$\mathbb{P}(X = j) = \frac{\binom{k}{j}\binom{m-k}{n-j}}{\binom{m}{n}}$$

Man sagt, die Zufallsvariable X besitzt eine hypergeometrische Verteilung zu den Parametern (m, k, n). Abgekürzt schreiben wir $X \sim HYPER(m, k, n)$.

Bemerkung 3.4.11. Die Bernoulli-, Binomial-, Poisson- und die hypergeometrische Verteilung gehören zur Klasse der diskreten Verteilungen und sind für statistische Anwendungen besonders wichtig. Darüber hinaus ist jede diskrete Verteilung eindeutig durch die zugehörige Wahrscheinlichkeitsfunktion

$$\mathbb{P}(X = x_k) = p_k > 0$$

definiert. $(x_k)_{k\geq 0}$ bezeichnet dabei die Folge der unterschiedlichen Werte, die von der diskreten Zufallsvariable angenommen werden können.

Mit diskreten Zufallsvariablen lassen sich keine kontinuierliche Längenmessungen beschreiben, denn diese liefern beliebige Ergebnisse in den reellen Zahlen. Um solche Experimente beschreiben zu können, benötigen wir einen weiteren Typ von Zufallsvariablen.

Definition 3.4.12. Eine Zufallsvariable X besitzt eine stetige Verteilung genau dann, wenn es eine Funktion

$$f: \quad \mathbb{R} \to \mathbb{R}^+$$

gibt (d.h. $f \geq 0$) mit der Eigenschaft, dass

$$\mathbb{P}(X \leq x) = F(x) = \int_{-\infty}^{x} f(t)dt$$

für alle $x \in \mathbb{R}$. Die Funktion f wird dabei die Dichte von X genannt.

Satz 3.4.13. Sei X eine Zufallsvariable mit Dichte f und Verteilungsfunktion $F \equiv F_X$. Dann gilt:

(i) $\int_{-\infty}^{\infty} f(t)dt = 1$
(ii) $\mathbb{P}(a < X \leq b) = \int_a^b f(t)dt, \;\; a \leq b \in \mathbb{R}$
(iii) $\mathbb{P}(X = a) = 0$ und F ist stetig
(iv) Für alle $x \in \mathbb{R}$ für die f stetig ist, gilt: $F'(x) = f(x)$

Beweis. Zu (i): Folgt aus

$$1 = \lim_{x\to\infty} F(x) = \lim_{x\to\infty} \int_{-\infty}^{x} f(t)dt = \int_{-\infty}^{\infty} f(t)dt$$

Zu (ii): Folgt aus

$$\begin{aligned} \mathbb{P}(a < X \leq b) &= \mathbb{P}(\{X \leq b\} \backslash \{X \leq a\}) = \mathbb{P}(X \leq b) - \mathbb{P}(X \leq a) = F(b) - F(a) \\ &= \int_{-\infty}^{b} f(t)dt - \int_{-\infty}^{a} f(t)dt = \int_{a}^{b} f(t)dt \end{aligned}$$

Zu (iii): Sei $h > 0$ beliebig. Dann gilt

$$\mathbb{P}(X = a) \leq \mathbb{P}(a - h < X \leq a) \underset{vgl.ii}{=} \int_{a-h}^{a} f(t)dt$$

Da

$$\int_{a-h}^{a} f(t)dt \underset{h\downarrow 0}{\longrightarrow} \int_{a}^{a} f(t)dt = 0,$$

folgt $\mathbb{P}(X = a) = 0$. Weiter gilt:

$$F(x-h) = \int_{-\infty}^{x-h} f(t)dt \underset{h\downarrow 0}{\uparrow} \int_{-\infty}^{x} f(t)dt = F(x)$$

und damit ist F auch linksseitig stetig. Zusammen mit der rechtsseitigen Stetigkeit von F (vgl. 3.4.3iii) ist F dann auch stetig.

Zu (iv): Vgl. Hauptsatz der Integralrechnung

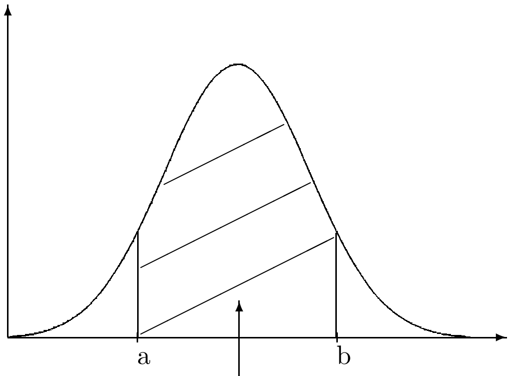
$$\begin{aligned} F'(x) &= \lim_{h\to 0} \frac{F(x+h) - F(x)}{h} = \lim_{h\to 0} \frac{1}{h} \int_{x}^{x+h} f(t)dt \\ &\underset{\substack{\text{MWS der Integralr.} \\ z\in[x,x+h]}}{=} \lim_{h\to 0} \frac{1}{h} f(z(h))((x+h) - x) \\ &= \lim_{h\to 0} f(z(h)) = f(x) \end{aligned}$$

□

Bemerkung 3.4.14. Sei f Dichte einer Zufallsvariable X. Dann ist

$$\mathbb{P}(a < X \leq b) = F(b) - F(a) = \int_{a}^{b} f(t)dt$$

die Fläche unter der Kurve der W.Dichte f.



$\mathbb{P}(a < X \leq b)$

Im Folgenden geben wir einige für die Anwendung relevante stetige Verteilungen an.

Beispiel 3.4.15. Die Funktion

$$f(x) = \begin{cases} \alpha \exp(-\alpha x) & : x \geq 0 \\ 0 & : x < 0 \end{cases}$$

mit $\alpha > 0$ stellt eine W.Dichte dar, denn $f \geq 0$ und

$$\int_{-\infty}^{\infty} f(t)\,dt \quad = \quad \int_{0}^{\infty} \alpha \exp(-\alpha x)\,dx = -\exp(-\alpha x)\Big|_0^{\infty} = 1$$

Ferner gilt für die zugehörige Verteilungsfunktion F für $x \geq 0$

$$F(x) \quad = \quad \int_{-\infty}^{x} f(t)\,dt = \int_{0}^{x} \alpha \exp(-\alpha t)\,dt = -\exp(-\alpha t)\Big|_0^{x} = 1 - \exp(-\alpha x)$$

Also

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

Besitzt eine Zufallsvariable X solch eine Verteilungsfunktion, so sagen wir X ist exponential-verteilt mit Parameter $\alpha \geq 0$. Abgekürzt schreiben wir $X \sim EXP(\alpha)$. Die Exponential-Verteilung gehört zu den klassischen Lebensdauerverteilungen und hat daher für die Qualitätskontrolle große Bedeutung.

Beispiel 3.4.16. Die Funktion

$$f(x) := \begin{cases} \alpha\beta(\alpha x)^{\beta-1} \exp\left(-(\alpha x)^{\beta}\right) & : x \geq 0 \\ 0 & : x < 0 \end{cases}$$

mit $\alpha, \beta > 0$ stellt eine Wahrscheinlichkeitsdichte dar, denn $f \geq 0$ und

$$\int_{-\infty}^{\infty} f(x)\,dx \quad = \quad \int_{0}^{\infty} \alpha\beta(\alpha x)^{\beta-1} \exp(-(\alpha x)^{\beta})\,dx = -\exp\left(-(\alpha x)^{\beta}\right)\Big|_0^{\infty} = 1$$

Ferner gilt für die Verteilungsfunktion F bei $x \geq 0$

$$F(x) \quad = \quad \int_{-\infty}^{x} f(t)\,dt = \int_{0}^{x} \alpha\beta(\alpha t)^{\beta-1} \exp(-(\alpha t)^{\beta})\,dt = -\exp(-(\alpha t)^{\beta})\Big|_0^{x} = 1 - \exp(-(\alpha x)^{\beta})$$

Also

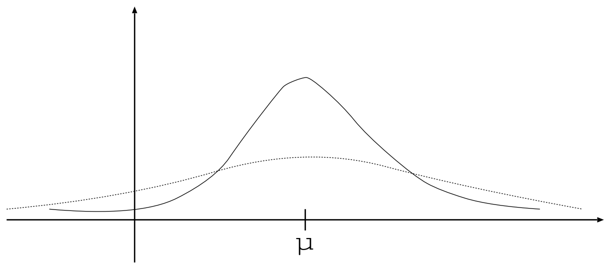
$$F(x) := \begin{cases} 1 - \exp\left(-(\alpha x)^{\beta}\right) & : x \geq 0 \\ 0 & : x < 0 \end{cases}$$

Die so definierte Verteilung wird Weibull-Verteilung mit den Parametern α, β genannt. Abgekürzt schreiben wir dann für eine Zufallsvariable $X \sim WEIB(\alpha, \beta)$, wenn wir ausdrücken wollen, dass X nach einer Weibull-Verteilung zu den Parametern α, β verteilt ist. Die Weibull-Verteilung ist das am häufigsten benutzte Modell bei Zuverlässigkeitsstudien.

Beispiel 3.4.17. Die Klassen der Normalverteilungen hat eine zentrale Stellung in der Statistik. Normalverteilungen treten insbesondere bei der Beurteilung von Messergebnissen auf. Eine Zufallsvariable X mit Dichte

$$f(x) := \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right), \quad x \in \mathbb{R}$$

heißt normalverteilt mit Mittelwert μ und Varianz σ^2. Wir schreiben wieder abgekürzt: $X \sim \mathcal{N}(\mu, \sigma^2)$



Die $\mathcal{N}$(0,1)- Verteilung heißt Standardnormalverteilung und sie ist vertafelt, (vgl. Anhang).

Bemerkung 3.4.18. Sei $X \sim \mathcal{N}(0,1)$ und $\mu \in \mathbb{R}$ sowie $\sigma > 0$. Dann gilt für $Y = \mu + \sigma X$: $Y \sim \mathcal{N}(\mu, \sigma^2)$

Beweis. Zunächst gilt:

$$F_Y(y) = \mathbb{P}(Y \leq y) = \mathbb{P}(\sigma X + \mu \leq y) = \mathbb{P}\left(X \leq \frac{(y-\mu)}{\sigma}\right) = F_X\left(\frac{y-\mu}{\sigma}\right)$$

Daraus folgt:

$$\begin{aligned} f_Y(y) &= F'_Y(y) = \frac{dF_X\left(\frac{(y-\mu)}{\sigma}\right)}{dy} \underset{K.R.}{=} f_X\left(\frac{(y-\mu)}{\sigma}\right)\frac{1}{\sigma} = \frac{1}{\sqrt{2\pi}}\exp\left(-\frac{(y-\mu)^2}{2\sigma^2}\right)\frac{1}{\sigma} \\ &= \frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(-\frac{(y-\mu)^2}{2\sigma^2}\right) \quad \text{- Dichte zu } \mathcal{N}(\mu,\sigma^2) \end{aligned}$$

□

Satz 3.4.19. Sei $X \sim \mathcal{N}(\mu, \sigma^2)$ und bezeichne mit f die zugehörige Dichte. Dann gilt:

(i) f ist achsensymmetrisch zu μ

(ii) f besitzt ein globales Maximum bei μ

(iii) $Y := \frac{(X-\mu)}{\sigma} \sim \mathcal{N}(0,1)$

(iii) wird benutzt, um eine beliebige $\mathcal{N}(\mu, \sigma^2)-$verteilte Zufallsvariable auf die vertafelte $\mathcal{N}(0,1)-$Verteilung zu transformieren.

Bemerkung 3.4.20. Sei $X \sim \mathcal{N}(\mu, \sigma^2)$. Dann gilt:

(i) $\mathbb{P}(\mu - \sigma \leq X \leq \mu + \sigma) \approx 0.68$

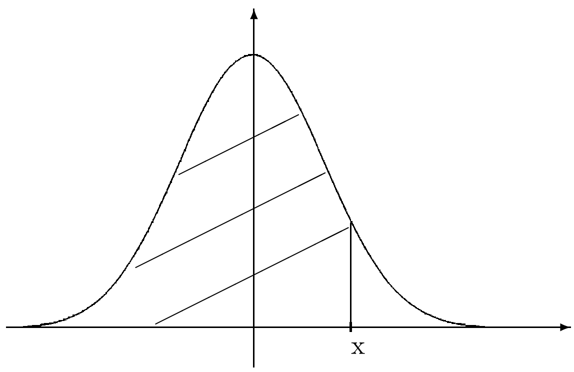
(ii) $\mathbb{P}(\mu - 2\sigma \leq X \leq \mu + 2\sigma) \approx 0.955$

Beweis. Wir zeigen hier nur den ersten Teil, denn (ii) ergibt sich analog. Teil (i) wird in der folgenden Rechnung nachgewiesen.

$$\begin{aligned}
\mathbb{P}(\mu - \sigma \leq X \leq \mu + \sigma) &= \int_{\mu-\sigma}^{\mu+\sigma} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) dx \\
&\underset{\substack{v(x):=\frac{x-\mu}{\sigma} \\ v'(x)=\frac{1}{\sigma}}}{=} \frac{1}{\sqrt{2\pi\sigma^2}} \int_{\mu-\sigma}^{\mu+\sigma} \exp\left(-\frac{v^2(x)}{2}\right) \frac{v'(x)}{v'(x)} dx \\
&= \frac{\sigma}{\sqrt{2\pi\sigma^2}} \int_{v(\mu-\sigma)=-1}^{v(\mu+\sigma)=1} \exp\left(-\frac{v^2}{2}\right) dv = \frac{1}{\sqrt{2\pi}} \int_{-1}^{1} \exp\left(-\frac{v^2}{2}\right) dv \\
&\underset{Y \sim \mathcal{N}(0,1)}{=} F_Y(1) - F_Y(-1) \underset{\text{sym.}}{=} F_Y(1) - (1 - F_Y(1)) = 2F_Y(1) - 1 \\
&\underset{\text{Tabelle}}{\approx} 2 * 0.8413 - 1 = 0.6826
\end{aligned}$$

□

Bemerkung 3.4.21. Im Anhang befindet sich eine ausführliche Tabelle der Werte von $F(x) = F_X(x)$ für die Standardnormalverteilung, d.h. $X \sim \mathcal{N}(0,1)$. Ein Ausschnitt dieser Tabelle wird im Folgenden



aufgelistet.

Verteilungsfunktion

x =	0	0.5	1	1.5	2
F(x) =	0.5	0.6915	0.8413	0.9332	0.9772

Quantil

F(x) =	0.9015	0.9505	0.9750
x =	1.29	1.65	1.96

Zur Berechnung ist folgende Formel hilfreich:

$$\mathbb{P}(-x \leq X \leq x) = \mathbb{P}(X \leq x) - \mathbb{P}(X \leq -x) = F(x) - (1 - F(x))$$

Basierend auf der Normalverteilung lassen sich weitere Verteilungen definieren, die im Rahmen der mathematischen Statistik von Wichtigkeit sind.

Beispiel 3.4.22. Gegeben seien $X_1, ..., X_n$ n unabhängige $\mathcal{N}(0,1)$ verteilte Zufallsvariable. Dann wird durch

$$Y := \sum_{i=1}^{n} X_i^2$$

eine neue Zufallsvariable definiert. Die zu Y gehörende Verteilung wird die Chi-Quadrat-Verteilung mit n Freiheitsgraden genannt. Abgekürzt schreiben wir auch $Y \sim \chi_n^2$. Zur Dichte der der χ_n^2-Verteilung vgl. (Weber S.103). Die χ^2-Verteilung ist ebenfalls für unterschiedliche Freiheitsgrade vertafelt, (vgl. Anhang).

Beispiel 3.4.23. Seien $X \sim \mathcal{N}(0,1)$ und $Y \sim \chi_n^2$ unabhängig. Dann wird durch

$$Z := \frac{X}{\sqrt{\frac{Y}{n}}}$$

eine neue Zufallsvariable definiert. Die zu Z gehörende Verteilung wird $t-$Verteilung, oder manchmal auch Student-Verteilung, mit n-Freiheitsgraden genannt. Abgekürzt schreiben wir $Z \sim t_n$. Zur Dichte der $t-$Verteilung vgl. (Weber S.258). Die $t-$Verteilung ist ebenfalls für unterschiedliche Freiheitsgrade vertafelt, (vgl. Anhang).

Beispiel 3.4.24. Seien $X \sim \chi_m^2$ und $Y \sim \chi_n^2$ unabhängige Zufallsvariable. Dann wird durch

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

eine neue Zufallsvariable definiert. Die zu Z gehörende Verteilung wird $F-$Verteilung mit (m, n) Freiheitsgraden genannt. Abgekürzt schreiben wir $Z \sim F_{m,n}$. Zur Dichte vgl. (Weber S.261). Auch zur $F-$Verteilung liegen Tafeln vor, vgl. Anhang.

Nach den letzten drei Verteilungen, die alle eng mit der Normalverteilung verknüpft sind, noch zum Abschluss eine Verteilung, die entscheidend für die Erzeugung von Zufallszahlen ist und die in fast jeder Programmiersprache implementiert ist.

Beispiel 3.4.25. Hat eine Zufallsvariable X die Dichte

$$f(x) = \frac{1}{b-a} 1_{[a,b]}(x)$$

so besitzt X eine Gleichverteilung auf dem Intervall $[a, b]$.

3.5 Grundbegriffe der Zuverlässigkeitstheorie

Die Zufallsvariable X beschreibe die Lebensdauer eines Bauteils. Im Folgenden sei F die zugehörige Verteilungsfunktion. Offensichtlich gilt dann für F:

$$F(t) = 0 \quad , \text{ für } t < 0.$$

Definition 3.5.1. Für $t > 0$ nennt man

$$F(t) = \mathbb{P}(X \leq t)$$

die Ausfallwahrscheinlichkeit vor dem Zeitpunkt t und

$$S(t) := 1 - F(t) = \mathbb{P}(X > t)$$

die Überlebenswahrscheinlichkeit bis zum Zeitpunkt t.

Wir nehmen im Folgenden weiter an, dass X eine Dichte f hat, also

$$F'(t) = f(t)$$

Definition 3.5.2. Die Größe

$$\int_0^\infty t\, f(t)\, dt$$

wird im Rahmen der Lebensdaueranalyse, die mittlere Lebensdauer MTBF (mean time before failure) genannt.

Bemerkung 3.5.3. Für $h > 0$ gilt:

$$\begin{aligned}
\frac{1}{h}\mathbb{P}(t < X \leq t+h | X > t) &= \frac{1}{h}\frac{\mathbb{P}(t < X \leq t+h\,,\, X > t)}{\mathbb{P}(X > t)} = \frac{1}{h}\frac{\mathbb{P}(t < X \leq t+h)}{S(t)} \\
&= \frac{1}{h}\int_t^{t+h} f(x)\, dx \frac{1}{S(t)} \\
&\underset{h \to 0}{\longrightarrow} \frac{f(t)}{S(t)}
\end{aligned}$$

Der Ausdruck auf der linken Seite in obiger Rechnung gibt die "mittlere W. für den Ausfalls eines Bauteils im Zeitinterval $]t, t+h]$ an, das zum Zeitpunkt t noch intakt war". Lassen wir h gegen 0 gehen, dann entspricht das einer momentanen Ausfallrate.

Definition 3.5.4. Die Größe

$$\lambda(t) := \frac{f(t)}{S(t)}, \quad t > 0$$

wird die Ausfallrate, Hazardrate oder auch momentane Sterbewahrscheinlichkeit genannt.

Beispiel 3.5.5. Sei $X \sim EXP(\alpha)$. Dann besitzt X die Dichte

$$f(t) = \begin{cases} \alpha \exp(-\alpha t) & : t \geq 0 \\ 0 & : t < 0 \end{cases}$$

und wir erhalten für die Überlebensfunktion bei $t \geq 0$

$$S(t) = \exp(-\alpha t)$$

Setzen wir diese Werte in die Gleichung aus Definition 3.5.4 ein, so folgt für die Hazardrate an der Stelle $t \geq 0$

$$\lambda(t) = \frac{f(t)}{S(t)} = \alpha,$$

d.h. exponential-verteilte Zufallsvariable haben eine konstante Hazardrate. Dies besagt, dass das zugehörige Bauteil nicht "altert".

Ferner ergibt sich für die MTBF:

$$\begin{aligned}
\text{MTBF} &= \int_0^\infty t\, f(t)\, dt = \int_0^\infty t\alpha \exp(-\alpha t)\, dt \\
&\underset{P.I.}{=} -t \exp(-\alpha t)\Big|_0^\infty + \int_0^\infty \exp(-\alpha t)\, dt \\
&= 0 + \left(-\frac{1}{\alpha}\exp(-\alpha t)\right)\Big|_0^\infty \\
&= \frac{1}{\alpha}
\end{aligned}$$

Beispiel 3.5.6. Sei $X \sim WEIB(\alpha, \beta)$. Dann hat X die Dichte

$$f(t) = \begin{cases} \alpha\beta(\alpha\, t)^{\beta-1} \exp\left(-(\alpha t)^\beta\right) & : t \geq 0 \\ 0 & : t < 0 \end{cases}$$

und die Überlebensfunktion

$$S(t) = \begin{cases} \exp\left(-(\alpha t)^\beta\right) & : t \geq 0 \\ 0 & : t < 0 \end{cases}$$

Dies liefert als Hazardrate für $t \geq 0$:

$$\lambda(t) = \frac{f(t)}{S(t)} = \alpha\beta(\alpha t)^{\beta-1}$$

Für $\beta > 1$ ist λ(t) monoton wachsend, d.h. "Ausfallrate nimmt mit dem Alter zu". In der Praxis ist dies das am meisten benutzte Modell.

Für $\beta < 1$ ist λ(t) monoton fallend, d.h. "Ausfallrate nimmt mit dem Alter ab". Dies kommt in der Praxis selten vor, wird aber für einige elektronische Bauteile angenommen.

Für $\beta = 1$ ist $\lambda(t) = \alpha =$ const. In diesem Fall entspricht die Weibull-Verteilung der Exponential-Verteilung.

$$MTBF = \int_0^\infty t f(t)\, dt = ?$$

Diese Größe kann für beliebige α, β nur näherungsweise (Numerik!) bestimmt werden.

Zwischen der Hazardrate und der Verteilungsfunktion einer Zufallsvariable gibt es einen engen Zusammenhang, der im folgenden Satz dargestellt ist.

Satz 3.5.7. Unter den in diesem Abschnitt gemachten Bezeichnungen gilt für $x > 0$:

$$-\ln(1 - F(x)) = \int_0^x \lambda(t)\, dt$$

Beweis. Der Beweis ergibt sich aus der folgenden Rechnung:

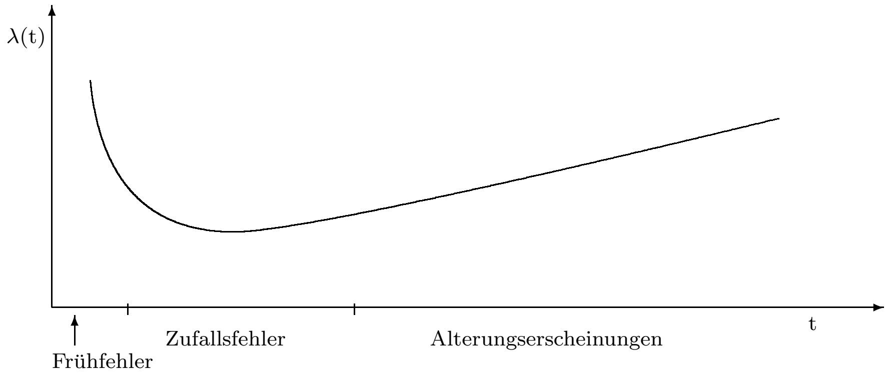
$$\begin{aligned}
\int_0^x \lambda(t)\, dt &= \int_0^x \frac{f(t)}{1 - F(t)}\, dt = \int_0^x \frac{1}{1 - F(t)} F'(t)\, dt \underset{\substack{\text{Substitut.} \\ u := F(t)}}{=} \int_{F(0)}^{F(x)} \frac{1}{1-u}\, du \\
&= -\ln(1-u)\Big|_0^{F(x)} = -\ln(1 - F(x)) + \ln(1 - 0) = -\ln(1 - F(x))
\end{aligned}$$

$\square$

Der letzte Satz besagt, dass man F aus λ und umgekehrt berechnen kann. Da F aber das W.theoretische Verhalten beschreibt, kann man dies auch mit λ. λ ist aber für den Praktiker, der ein W.-Modell für die Lebensdauer eines Objekts sucht, geeigneter, denn

$$\lambda(t) \; - \; \text{Ausfallrate bei Alter t.}$$

Bemerkung 3.5.8. Weitere wichtige Lebensdauerverteilungen haben einen "badewannenförmigen Verlauf" der Hazardrate (z.B. Hjorth-Verteilung). "Frühfehler" können unter Umständen durch eine



gute Endkontrolle ausgemerzt werden. Man erhält dann wieder steigende Hazardraten und kann ein Weibull-Modell benutzen.

3.6 Momente

Beispiel 3.6.1. Zwei Personen, S und T, spielen folgendes Spiel: S würfelt mit einem fairen Würfel und erhält von T

1	DM	bei	1 oder 2
2	DM	bei	3 oder 4
4	DM	bei	5
8	DM	bei	6

Vor jedem Spiel soll S an T einen Betrag zahlen, der gleich seiner durchschnittlichen Gewinnerwartung pro Spiel ist. Diese beträgt:

$$1\left(\frac{1}{6}+\frac{1}{6}\right)+2\left(\frac{1}{6}+\frac{1}{6}\right)+\frac{4}{6}+\frac{8}{6}=\frac{18}{6}=3DM$$

Definition 3.6.2. Sei $g : \mathbb{R} \to \mathbb{R}$ eine vorgegebene Funktion

i) Für eine diskrete Zufallsvariable X, d.h. X kann nur die Werte $x_i \in \mathbb{R}$, $i \in \mathbb{N}$, annehmen, bezeichnet

$$\mathbb{E}(g(X)) := \sum_{i\in\mathbb{N}} g(x_i)\mathbb{P}(X = x_i),$$

falls diese Größe existiert, den <u>Erwartungswert von $g(X)$</u>.

ii) Für eine stetige Zufallsvariable X mit Dichte f bezeichnet

$$\mathbb{E}(g(X)) := \int_{-\infty}^{\infty} g(x)\, f(x)\, dx,$$

falls diese Größe existiert, ebenfalls den Erwartungswert von $g(X)$.

iii) Ist $g(x) = x$, für alle $x \in \mathbb{R}$, so nennt man $\mathbb{E}(X) = \mu$ den Erwartungswert von X.

Beispiel 3.6.3. i) Sei $X \sim BERN(p)$. Dann erhalten wir:

$$\mathbb{E}(X) = 1 \cdot \mathbb{P}(X = 1) + 0 \cdot \mathbb{P}(X = 0) = 1p = p$$

ii) Für $X \sim EXP(\alpha)$ ergibt sich wie wir unter Beispiel 3.5.5 gesehen haben:

$$\mathbb{E}(X) = \mathrm{MTBF} = \frac{1}{\alpha}$$

Bemerkung 3.6.4. Seien X eine Zufallsvariable und $g : \mathbb{R} \to \mathbb{R}$, $h : \mathbb{R} \to \mathbb{R}$ Funktionen mit der Eigenschaft, dass $\mathbb{E}(g(X))$ und $\mathbb{E}(h(X))$ existieren. Dann gilt für $a, b \in \mathbb{R}$ bel.:

$$\mathbb{E}(a\,g(X) + b\,h(X)) = a\,\mathbb{E}(g(X)) + b\,\mathbb{E}(h(X))$$

Beweis. Betrachte hier nur den Fall, dass X eine Dichte f besitzt. Dann gilt:

$$\begin{aligned}
\mathbb{E}(a\,g(X) + b\,h(X)) &= \int_{-\infty}^{\infty} (a\,g(x) + b\,h(x))\,f(x)\,dx \\
&\underset{\substack{\text{Rechenregeln}\\ \text{für Integrale}}}{=} a \int_{-\infty}^{\infty} g(x)\,f(x)\,dx + b \int_{-\infty}^{\infty} h(x)\,f(x)\,dx \\
&= a\,\mathbb{E}(g(X)) + b\,\mathbb{E}(h(X))
\end{aligned}$$

□

Satz 3.6.5. Additionssatz Seien $X_1, ..., X_n$ Zufallsvariable def. über $(\Omega, \mathcal{A}, \mathbb{P})$. Ferner seien $g_i : \mathbb{R} \to \mathbb{R}$, $i = 1, \ldots, n$ Funktionen mit der Eigenschaft, dass $\mathbb{E}(g_i(X_i))$ existiert, für $i = 1, \ldots, n$. Dann gilt:

$$\mathbb{E}\left(\sum_{i=1}^{n} g_i(X_i)\right) = \sum_{i=1}^{n} \mathbb{E}(g_i(X_i))$$

Beweis. Entfällt! □

Beispiel 3.6.6. (i) Sei $X \sim BIN(n.p)$. Dann folgt: $\mathbb{E}(X) = np$.
Zunächst wissen wir, dass

$$X = \sum_{i=1}^{n} Y_i,$$

wobei $Y_1, \ldots, Y_n \sim BERN(p)$ unabhängig voneinander sind. Da $\mathbb{E}(Y_i) = p$, für $i = 1, \ldots, n$, wie wir schon gesehen haben, erhalten wir nach dem Additionssatz:

$$\mathbb{E}(X) = \mathbb{E}\left(\sum_{i=1}^{n} Y_i\right) = \sum_{i=1}^{n} \mathbb{E}(Y_i) = \sum_{i=1}^{n} p = np$$

(ii) Seien $X_1, ..., X_n$ Zufallsvariable mit identischer Verteilungsfunktion und existiere $\mathbb{E}(X_1) = \mu$. Dann folgt ebenfalls aus dem Additionssatz:

$$\mathbb{E}\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n}\mathbb{E}\left(\sum_{i=1}^{n} X_i\right) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}(X_i) = \frac{1}{n}\sum_{i=1}^{n} \mu = \mu$$

Definition 3.6.7. Sei X eine Zufallsvariable und existiere der Erwartungswert von X^2. Ferner bezeichne $\mathbb{E}(X) = \mu$ den Erwartungswert von X. Dann wird durch

$$VAR(X) := \sigma^2 := \begin{cases} \sum\limits_{i \in \mathbb{N}} (x_i - \mu)^2 \mathbb{P}(X = x_i) & : X \text{ diskret} \\ \int\limits_{-\infty}^{\infty} (x - \mu)^2 f(x)\, dx & : X \text{ stetig mit Dichte } f \end{cases}$$

die Varianz der Zufallsvariablen X definiert. Im diskreten Fall werden hierbei von X nur die Werte x_i, mit $i \in \mathbb{N}$ angenommen. Man beachte, in jedem Fall gilt $VAR(X) = \mathbb{E}((X - \mu)^2)$.

Bemerkung 3.6.8. Sei X eine Zufallsvariable mit existierender Varianz. Dann gilt:

$$VAR(X) = \mathbb{E}(X^2) - (\mathbb{E}(X))^2 = \mathbb{E}(X^2) - \mu^2$$

Beweis. Betrachte hier nur den Fall, dass X eine Dichte f besitzt.

$$\begin{aligned} VAR(X) &= \int_{-\infty}^{\infty} \underbrace{(x - \mu)^2}_{x^2 - 2x\mu + \mu^2} f(x)\, dx = \int_{-\infty}^{\infty} x^2 f(x)\, dx - 2\mu \underbrace{\int_{-\infty}^{\infty} x\, f(x)\, dx}_{=\mu} + \mu^2 \underbrace{\int_{-\infty}^{\infty} f(x)\, dx}_{=1} \\ &= \mathbb{E}(X^2) - \mu^2 \end{aligned}$$

□

Beispiel 3.6.9. Wir bestimmen in diesem Beispiel Varianzen von Zufallsvariablen nach der Formel aus Bemerkung 3.6.8.

(i) Sei $X \sim BERN(p)$. Dann hatten wir bereits gesehen, dass $\mathbb{E}(X) = p$. Ferner gilt:

$$\mathbb{E}(X^2) = 1^2 \mathbb{P}(X = 1) + 0^2 \mathbb{P}(X = 0) = p$$

Mit der Formel aus Bemerkung 3.6.8 erhalten wir also:

$$VAR(X) = p - p^2 = p(1 - p)$$

(ii) Sei jetzt $X \sim EXP(\alpha)$. Dann hatten wir schon $\mathbb{E}(X) = 1/\alpha$ gezeigt. Ferner gilt:

$$\begin{aligned} \mathbb{E}(X^2) &= \int_0^{\infty} x^2\, \alpha \exp(-\alpha\,) dx \\ &\underset{P.I.}{=} -x^2 \exp(-\alpha\, x)\Big|_0^{\infty} + \int_0^{\infty} 2x \exp(-\alpha\, x) dx \\ &= \frac{2}{\alpha} \int_0^{\infty} x\, \alpha \exp(-\alpha\, x) dx \\ &\underset{(vgl. 3.6.3ii)}{=} \frac{2}{\alpha} \frac{1}{\alpha} = \frac{2}{\alpha^2} \end{aligned}$$

Insgesamt erhalten wir dann mit Bemerkung 3.6.8

$$VAR(X) = \mathbb{E}(X^2) - \mu^2 = \frac{2}{\alpha^2} - \frac{1}{\alpha^2} = \frac{1}{\alpha^2}$$

Satz 3.6.10. Multiplikationssatz. Seien $X_1, ..., X_n$ unabhängige Zufallsvariable definiert über $(\Omega, \mathcal{A}, \mathbb{P})$. Ferner seien $g_i : \mathbb{R} \to \mathbb{R}$ Funktionen derart, dass $\mathbb{E}(g_i(X_i))$ existiert für $i = 1, ..., n$. Dann gilt:

$$\mathbb{E}(g_1(X_1)\, g_2(X_2) \ldots g_n(X_n)) = \mathbb{E}(g_1(X_1))\, \mathbb{E}(g_2(X_2)) \ldots \mathbb{E}(g_n(X_n))$$

Beweis. Entfällt. □

Satz 3.6.11. Seien $X_1, ..., X_n$ unabhängige Zufallsvariable mit $VAR(X_i) = \sigma_i^2$, für $i = 1, \ldots, n$. Dann gilt:

$$VAR\left(\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n} VAR(X_i)$$

Beweis. Bezeichne $\mu_i := \mathbb{E}(X_i)$, für $i = 1, ..., n$. Dann gilt nach dem Additionssatz:

$$\mathbb{E}\left(\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n} \mathbb{E}(X_i) = \sum_{i=1}^{n} \mu_i$$

Damit erhalten wir nun aus dem Multiplikationssatz, wenn man beachtet, dass $\mathbb{E}(X_i - \mu_i) = 0$:

$$\begin{aligned}
VAR\left(\sum_{i=1}^{n} X_i\right) &= \mathbb{E}\left(\left(\sum_{i=1}^{n} X_i - \sum_{i=1}^{n} \mu_i\right)^2\right) = \mathbb{E}\left(\left(\sum_{i=1}^{n} (X_i - \mu_i)\right)^2\right) \\
&= \mathbb{E}\left(\sum_{i=1}^{n}\sum_{j=1}^{n} (X_i - \mu_i)(X_j - \mu_j)\right) \\
&= \sum_{i=1}^{n}\sum_{j=1}^{n} \mathbb{E}\left((X_i - \mu_i)(X_j - \mu_j)\right) \\
&= \sum_{i=1}^{n}\sum_{\substack{j=1 \\ j \neq i}}^{n} \mathbb{E}\left((X_i - \mu_i)\right) \mathbb{E}\left((X_j - \mu_j)\right) + \sum_{i=1}^{n} \mathbb{E}((X_i - \mu_i)(X_i - \mu_i)) \\
&= \sum_{i=1}^{n} \mathbb{E}\left((X_i - \mu_i)^2\right) = \sum_{i=1}^{n} VAR(X_i)
\end{aligned}$$

□

Bemerkung 3.6.12. Analog ergibt sich aus dem letzten Satz für $X_1, ..., X_n$ unabhängige Zufallsvariable mit identischer Verteilungsfunktion und existierender $VAR(X_1) = \sigma^2$:

$$VAR\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n}\sigma^2$$

Beispiel 3.6.13. Falls $X \sim BIN(n, p)$, dann ist $X = \sum_{i=1}^{n} Y_i$, wobei $Y_1, \ldots, Y_n$ unabhängig und identisch $BERN(p)$ verteilt sind. Da $VAR(Y_i) = p(1-p)$, folgt aus dem Multiplikationssatz:

$$VAR(X) = VAR(\sum_{i=1}^{n} Y_i) = np(1-p)$$

Bemerkung 3.6.14. Seien $X_1, ..., X_n$ unabhängige Zufallsvariable mit $\mathbb{E}(X_i) = \mu$ und $VAR(X_i) = \sigma^2$, für $i = 1, \ldots, n$. Bezeichne jetzt

$$s_n^2 := \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \overline{X})^2$$

mit $\overline{X} := \frac{1}{n} \sum_{i=1}^{n} X_i$. Dann folgt:

$$\mathbb{E}(s_n^2) = \sigma^2$$

Beweis. Wende Multiplikationssatz an auf

$$s_n^2 \quad = \quad \frac{n}{n-1} \frac{1}{n} \sum_{i=1}^{n} (X_i - \overline{X})^2 = \frac{n}{n-1} \left(\frac{1}{n} \sum_{i=1}^{n} X_i^2 - \overline{X}^2 \right).$$

□

Bemerkung 3.6.15. Im Folgenden geben wir Erwartungswert und Varianz einiger wichtiger Verteilungen an.

Diskret:

$X \sim$	$BERN(p)$	$BIN(n,p)$	$POIS(\lambda)$	$Hyper(m,k,n)$
$\mathbb{E}(X)$	p	np	λ	n $\frac{k}{m}$
$VAR(X)$	p(1-p)	np(1-p)	λ	n $\frac{k}{m}(1-\frac{k}{m})\frac{m-n}{m-1}$

Stetig:

$X \sim$	$EXP(\alpha)$	$WEIB(\alpha,\beta)$
$\mathbb{E}(X)$	$\frac{1}{\alpha}$	num
$VAR(X)$	$\frac{1}{\alpha^2}$	num

Stetig:

$X \sim$	$\mathcal{N}(\mu,\sigma^2)$	χ_n^2	t_n	$F_{m,n}$
$\mathbb{E}(X)$	μ	n	$(n \geq 2)\, 0$	$(n \geq 2) \frac{n}{n-2}$
$VAR(X)$	σ^2	2n	$(n \geq 3)\, \frac{n}{n-2}$	$(n \geq 5) \frac{2(n+m-2)}{m(n-4)(n-2)}$

3.7 Tschebyscheff-Ungleichung

Satz 3.7.1. Sei X eine Zufallsvariable mit $\mathbb{E}(X^2) < \infty$. Dann gilt für jedes $\varepsilon > 0$ und für jedes $c \in \mathbb{R}$:

$$\mathbb{P}(|X - c| > \varepsilon) \leq \frac{1}{\varepsilon^2} \mathbb{E}\left((X-c)^2\right)$$

Beweis. Zeige nur den Fall, dass X eine Dichte f besitzt.

$$
\begin{aligned}
\mathbb{P}(|X-c|>\varepsilon) &= \mathbb{P}(X-c>\varepsilon \quad \text{oder} \quad -(X-c)>\varepsilon) \\
&= \mathbb{P}(X>c+\varepsilon)+\mathbb{P}(X<c-\varepsilon) \\
&= \int_{c+\varepsilon}^{\infty} f(x)\,dx + \int_{-\infty}^{c-\varepsilon} f(x)\,dx \\
&\underset{\oplus}{\leq} \int_{c+\varepsilon}^{\infty} \frac{(x-c)^2}{\varepsilon^2} f(x)\,dx + \int_{-\infty}^{c-\varepsilon} \frac{(c-x)^2}{\varepsilon^2} f(x)\,dx \\
&\leq \frac{1}{\varepsilon^2}\left(\int_{c+\varepsilon}^{\infty} (x-c)^2 f(x)\,dx + \int_{-\infty}^{c-\varepsilon} (x-c)^2 f(x)\,dx + \int_{c-\varepsilon}^{c+\varepsilon} (x-c)^2 f(x)\,dx\right) \\
&= \frac{1}{\varepsilon^2}\int_{-\infty}^{\infty} (x-c)^2 f(x)\,dx \\
&= \frac{1}{\varepsilon^2}\mathbb{E}\left((X-c)^2\right)
\end{aligned}
$$

Es bleibt somit noch $\oplus$ nachzuweisen. Wir zeigen hier nur den Fall:

$$\int_{c+\varepsilon}^{\infty} f(x)\,dx \leq \int_{c+\varepsilon}^{\infty} \frac{(x-c)^2}{\varepsilon^2} f(x)\,dx$$

Dies ergibt sich aber aus

$$x \geq c+\varepsilon \Rightarrow \frac{x-c}{\varepsilon} \geq 1 \Rightarrow \frac{(x-c)^2}{\varepsilon^2} \geq 1^2 = 1$$

$\square$

Bemerkung 3.7.2. Setzt man in 3.7.1 für $c = \mathbb{E}(X)$, dann erhält man

$$\mathbb{P}(|X-\mathbb{E}(X)|>\varepsilon) \leq \frac{1}{\varepsilon^2}\mathbb{E}\left((X-\mathbb{E}(X))^2\right) = \frac{1}{\varepsilon^2} VAR(X)$$

Diese Beziehung wird Tschebyscheff-Ungleichung genannt.

Beispiel 3.7.3. Anwendung der Tschebyscheff-Ungleichung.

i) Seien $X_1, ..., X_n$ n unabhängige Experimente, die alle die gleiche Verteilung besitzen. Bezeichne

$$\mathbb{E}(X_1) = \mu \;, VAR(X_1) = \sigma^2$$

Man ist an μ interessiert und schätzt μ durch

$$\overline{X} := \frac{1}{n}\sum_{i=1}^{n} X_i$$

$\overline{X}$ ist jetzt eine neue Zufallsvariable mit

$$\mathbb{E}(\overline{X}) = \mu \quad \text{und} \quad VAR(\overline{X}) = \frac{1}{n}\sigma^2$$

Die Tschebyscheff-Ungleichung besagt dann:

$$\mathbb{P}(|\overline{X} - \mu| > \varepsilon) \leq \frac{1}{\varepsilon^2} VAR(\overline{X}) = \frac{1}{\varepsilon^2}\frac{\sigma^2}{n} \underset{n\to\infty}{\longrightarrow} 0$$

d.h.

$$\mathbb{P}(|\overline{X} - \mu| \leq \varepsilon) = 1 - \mathbb{P}(|\overline{x} - \mu| > \varepsilon) \underset{n\to\infty}{\longrightarrow} 1$$

In Worten "Mit zunehmendem Stichprobenumfang n nähert sich die Schätzung $\overline{X}$ dem wahren Wert μ an."

ii) Sei jetzt $X \sim EXP(1)$, d.h.

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

Wie wir schon gesehen haben gilt dann:

$$\mathbb{E}(X) = 1 \quad \text{und} \quad VAR(X) = \frac{1}{1^2} = 1$$

Wenden wir jetzt die Tschebyscheff-Ungleichung an, dann erhalten wir

$$\mathbb{P}(|X - 1| > 3) \leq \frac{1}{3^2} VAR(X) = \frac{1}{9}$$

iii) Die 3σ - Regel:
Sei X eine Zufallsvariable mit $\mathbb{E}(X) = \mu$ und $VAR(X) = \sigma^2$. Dann liefert die Tschebyscheff-Ungleichung:

$$\mathbb{P}(|X - \mu| > 3\sigma) \leq \frac{1}{(3\sigma)^2}\,\sigma^2 = \frac{1}{9}$$

Äquivalent hierzu ist:

$$\mathbb{P}(|X - \mu| \leq 3\sigma) \geq \frac{8}{9}$$

Letzte Ungleichung besagt:

$$\mathbb{P}(\mu - 3\sigma \leq X \leq \mu + 3\sigma) \geq \frac{8}{9}$$

d.h. mit $8/9 \approx 89\%$ Wahrscheinlichkeit liefert das Experiment ein Ergebnis, dass im Intervall $[\mu - 3\sigma, \mu + 3\sigma]$ liegt.

3.8 Zentraler Grenzwertsatz

Der hohe Stellenwert der Normalverteilung beruht auf dem Zentralen Grenzwertsatz (ZGWS). Dieser besagt, dass für Zufallsvariablen $X_1, ..., X_n$, die alle die gleiche Verteilung besitzen und unabhängig sind, gilt:

$$\mathbb{P}\left(n^{\frac{1}{2}}\,\frac{\overline{X} - \mu}{\sqrt{s_n^2}} \leq x\right) \underset{n\to\infty}{\longrightarrow} \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} \exp\left(-\frac{t^2}{2}\right) dt$$

Vorausgesetzt wird allerdings, dass $\mathbb{E}(X) = \mu$ und $VAR(X) = \sigma^2 < \infty$. Diese Konvergenz gilt für jedes $x \in \mathbb{R}$. Dabei bezeichnen

$$\overline{X} := \frac{1}{n}\sum_{i=1}^{n} X_i \quad \text{und} \quad s_n^2 := \frac{1}{n}\sum_{i=1}^{n} (X_i - \overline{X})^2$$

Der ZGWS besagt, dass bei “großen” n die unbekannte Wahrscheinlichkeit

$$\mathbb{P}\left(n^{\frac{1}{2}}\frac{\overline{X}-\mu}{\sqrt{s_n^2}}\leq x\right)$$

durch die Verteilungsfunktion der Standardnormalverteilung an der Stelle x, d.h. durch

$$\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{x}\exp\left(-\frac{t^2}{2}\right)\,dt$$

“approximiert” werden kann.

Der Anwendungsbereich des ZGWS ist sehr groß, wie wir in den folgenden Kapiteln zur Statistik sehen werden.

Bemerkung 3.8.1. Statt $\sqrt{s_n^2}$ sollte $\sqrt{\sigma^2}$ in obiger Formel gewählt werden, falls σ^2 bekannt ist.

Kapitel 4

Schätztheorie

4.1 Allgemeine Grundbegriffe

Bemerkung 4.1.1. Gegeben ist eine Grundgesamtheit. Wir interessierten uns für einen Parameter p der Verteilung eines Merkmals dieser Grundgesamtheit. Z.B. kann ein Merkmal eine $\mathcal{N}(\mu, \sigma^2)$-Verteilung besitzen, dann sind die interessanten Parameter $p = \mu$ und $p = \sigma^2$. Zur "Schätzung" dieser Parameter ziehen wir eine Stichprobe $X_1, ..., X_n$, d.h. wir machen n Experiment, die in der Regel unabhängig und identisch verteilt sind, und suchen uns eine Statistik

$$p_n := g(X_1, ..., X_n)$$

als Schätzer eines der Parameter. Dabei treten offensichtlich die folgenden beiden Probleme auf:

- Wie findet man solch einen Schätzer?
- Welche Qualiät bzw. Eigenschaften hat der Schätzer?

Beispiel 4.1.2. Ein Merkmal sei $\mathcal{N}(\mu, \sigma^2)$ verteilt, wobei μ und σ^2 nicht bekannt sind. Als Schätzer für μ bzw. σ^2 wählen wir

$$\mu \approx \overline{X} \quad \text{bzw.} \quad \sigma^2 \approx s_n^2$$

Im ersten Fall ist

$$g(X_1, ..., X_n) := \frac{1}{n} \sum_{i=1}^{n} X_i$$

und im zweiten Fall

$$g(X_1, ..., X_n) := \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \overline{X})^2.$$

Beide Schätzer sind <u>erwartungstreu</u>, d.h.

$$\mathbb{E}(\overline{X}) = \mu, \quad \mathbb{E}(s_n^2) = \sigma^2$$

Neben der Erwartungstreue sollte man von einem Schätzer p_n erwarten, dass er sich mit zunehmendem Stichprobenumfang n dem wahren Wert p nähert.

Beispiel 4.1.3. Seien $X_1, ..., X_n$ unabhängig und identisch verteilt mit Verteilungsfunktion F und $\mu := \mathbb{E}(X)$ bzw. $\sigma^2 := VAR(X)$. Dann hatten wir unter Beispiel 3.7.3 (i) gesehen:

$$\text{Für } \varepsilon > 0 \text{ beliebig:} \quad \mathbb{P}(|\overline{X} - \mu| \leq \varepsilon) \to 1$$

Definition 4.1.4. Gegeben sei eine Verteilungsfunktion F und ein Schätzer p_n für einen Parameter p von F. Falls dann für alle $\varepsilon > 0$

$$\mathbb{P}(|p_n - p| > \varepsilon) \underset{n\to\infty}{\to} 0,$$

so nennt man p_n einen <u>konsistenten Schätzer</u> für p.

Bemerkung 4.1.5. $\overline{X}$ und s_n^2 sind konsistente und erwartungstreue Schätzer für μ bzw. σ^2.

4.2 Maximum Likelihood Methode

Gehen wir nun der eingangs gestellten Frage nach, wie man geeignete Schätzer finden kann. Dazu betrachten wir zunächst ein Beispiel.

Beispiel 4.2.1. Seien $X_1, \ldots, X_n$ unabhängige und identisch nach $POIS(\lambda)$ verteilte Zufallsvariable. D.h.

$$\mathbb{P}(X = k) = \frac{\lambda^k}{k!} e^{-\lambda}, k = 0, 1, \ldots$$

λ ist der unbekannte Parameter. Gesucht wird eine Schätzung λ_n von λ, die auf den Beobachtungen $x_1, \ldots, x_n$ der Zufallsvariable $X_1, \ldots, X_n$ beruht. Betrachte

$$\begin{aligned} l(\lambda) &:= \mathbb{P}(X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n) \\ &\underset{\text{unabh.}}{=} \mathbb{P}(X_1 = x_1) \cdot \mathbb{P}(X_2 = x_2) \cdots \mathbb{P}(X_n = x_n) \\ &= \frac{\lambda^{x_1}}{x_1!} e^{-\lambda} \frac{\lambda^{x_2}}{x_2!} e^{-\lambda} \cdots \frac{\lambda^{x_n}}{x_n!} e^{-\lambda} \end{aligned}$$

Logarithmieren liefert

$$L(\lambda) := \ln(l(\lambda)) = \sum_{i=1}^{n} \ln\left(\frac{\lambda^{x_i}}{x_i!} e^{-\lambda}\right)$$

$L(\lambda)$ kann nicht berechnet werden, da wir λ nicht kennen. Da die Daten aber zu dieser Verteilung ($POIS(\lambda)$ gehören, sollte man λ_n so wählen, dass die Funktion $L(\lambda)$ maximal wird, d.h. λ_n passt am besten zu den Daten oder λ_n ist das wahrscheinlichste λ, das diese Stichprobe $x_1, \ldots, x_n$ hat entstehen lassen können.

Maximierung von $L(\lambda)$ ist typisches Min/Max Problem aus der Analysis Grundvorlesung.

$$\begin{aligned} L(\lambda) &= \sum_{i=1}^{n} \ln\left(\frac{\lambda^{x_i}}{x_i!} e^{-\lambda}\right) = \sum_{i=1}^{n} (x_i \ln(\lambda) - \ln(x_i!) - \lambda) \\ &= \ln(\lambda) \sum_{i=1}^{n} x_i - \sum_{i=1}^{n} \ln(x_i!) - n\lambda \end{aligned}$$

Damit erhalten wir für die Ableitung:

$$\begin{aligned} L'(\lambda) &= \frac{1}{\lambda} \sum_{i=1}^{n} x_i - n = 0 \\ &\Rightarrow \lambda = \frac{1}{n} \sum_{i=1}^{n} x_i = \overline{x} \end{aligned}$$

Daraus folgt schließlich:

$$\left.\begin{array}{r} 0 < \lambda < \overline{x} \Rightarrow L'(\lambda) > 0 \\ \lambda > \overline{x} \Rightarrow L'(\lambda) < 0 \end{array}\right\} \text{ Also } \lambda_n := \overline{x} \text{ ist eine globale Maximalstelle von } L.$$

Als Schätzer für λ wählt man $\lambda_n := \overline{x}$.

Beispiel 4.2.2. Seien $X_1, ..., X_n$ unabhängige und identisch nach $EXP(\lambda)$ verteilte Zufallsvariable. D.h. es gibt eine Dichte

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

Gesucht wird wieder eine Schätzung λ_n von λ, die auf den Beobachtungen $x_1, \dots, x_n$ der Z.V. $X_1, \dots, X_n$ beruht. Betrachte jetzt

$$l(\lambda) := f(x_1) \cdots f(x_n) = \lambda e^{-\lambda x_1} \cdots \lambda e^{-\lambda x_n}$$

Logarithmieren liefert

$$L(\lambda) := \ln(l(\lambda)) = \sum_{i=1}^{n} \ln(\lambda e^{-\lambda x_i})$$

Als Schätzer für λ wählen wir wieder die Maximalstelle von $L(\lambda)$.

$$L(\lambda) \quad = \quad \sum_{i=1}^{n} \ln(\lambda) - \sum_{i=1}^{n} \lambda x_i = n \ln(\lambda) - \lambda \sum_{i=1}^{n} x_i$$

Für die Ableitung ergibt sich:

$$L'(\lambda) \quad = \quad \frac{n}{\lambda} - \sum_{i=1}^{n} x_i = 0 \qquad \Rightarrow \lambda = \frac{n}{\sum\limits_{i=1}^{n} x_i} = \frac{1}{\overline{x}}$$

Eine analoge Betrachtung wie unter Beispiel 4.2.1 zeigt, dass es sich hierbei wieder um eine globale Maximalstelle handelt. Als Schätzer für λ wählt man also $\lambda_n = \frac{1}{\overline{x}}$.

Beispiel 4.2.3. Seien $X_1, ..., X_n$ unabhängige und identisch nach $\mathcal{N}(\mu, \sigma^2)$ verteilte Zufallsvariable. D.h. die Dichte hat die Form

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

Gesucht werden Schätzer μ_n für μ und σ_n^2 für σ^2, die auf den Beobachtungen $x_1, ..., x_n$ beruhen. Analog zum Beispiel 4.2.2 liefert eine Maximierung von

$$\begin{aligned} L(\mu, \sigma) \quad &:= \quad \sum_{i=1}^{n} \ln(f(x_i)) = \sum_{i=1}^{n} \ln\left(\frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x_i-\mu)^2}{2\sigma^2}\right)\right) \\ &= \quad -n \ln\left(\sqrt{2\pi\sigma^2}\right) - \sum_{i=1}^{n} \frac{(x_i-\mu)^2}{2\sigma^2} \end{aligned}$$

dann die Schätzer

$$\mu_n = \overline{x} \quad \text{und} \quad \sigma_n^2 = \frac{1}{n} \sum_{i=1}^{n} (x_i - \overline{x})^2$$

Beachte, in diesem Fall "$\frac{1}{n}$ statt $\frac{1}{n-1}$"!

Definition 4.2.4. Maximum-Likelihood Methode. Seien $X_1, \dots, X_n$ unabhängige und indentisch verteilte Zufallsvariable, deren Verteilung sich aus einem (oder mehreren) Parameter bestimmt. D.h. man nimmt an, dass die Verteilung der Zufallsvariable zu einer bekannten vorgegebenen Klasse von Verteilungen gehört, wie etwa die Klasse der Normalverteilungen oder die Klasse der Exponentialverteilungen, und man muss lediglich den zugehörigen Parameter bestimmen, der die Verteilung innerhalb der Klasse exakt identifiziert (z.B. μ, σ für $\mathcal{N}(\mu, \sigma^2)$; λ für $EXP(\lambda)$ oder für $POIS(\lambda)$. Gesucht wird auf der Basis der Beobachtungen $x_1, \ldots, x_n$ eine Schätzung dieses Parameters. Bezeichne dann mit

$$L(\lambda) := \begin{cases} \sum\limits_{i=1}^{n} \ln(\mathbb{P}_\lambda(X = x_i)) & : \text{falls diskrete Verteilung} \\ \sum\limits_{i=1}^{n} \ln(f_\lambda(x_i)) & : \text{falls stetige Verteilung} \end{cases}$$

die Log-Likelihood Funktion. X ist dabei eine Zufallsvariable, deren Verteilung aus der zugehörigen Klasse kommt und die den Parameter λ hat. $\mathbb{P}_\lambda$ bezeichnet die Wahrscheinlichkeitsfunktion bzw. f_λ die Dichte zum Parameter λ. Der Maximum-Likelihood Schätzer (MLE), λ_n, ist derjenige Wert, der $L(\cdot)$ maximiert.

Beispiel 4.2.5. Schätzung des Parameters p bei Bernoulli-verteilten Zufallsvariablen mit MLE.

Für die Log-Likelihood Funktion erhalten wir

$$L(p) = \sum_{i=1}^{n} \ln(\mathbb{P}_p(X = x_i))$$

wobei

$$\mathbb{P}_p(X = x_i) = \begin{cases} p & : x_i = 1 \\ 1-p & : x_i = 0 \end{cases}$$

Angenommen, $k-$mal wurde eine 1 und $(n-k)-$mal eine 0 beobachtet. Dann ergibt sich

$$L(p) = k \ln(p) + (n-k)\ln(1-p)$$

und für die Ableitung

$$L'(p) = \frac{k}{p} - \frac{n-k}{1-p} = \frac{n}{p(1-p)}\left(\frac{k}{n} - p\right) = 0$$

Offensichtlich ist $p = k/n = \overline{x}$ und wir erhalten weiter aus obiger Darstellung:

$$\left.\begin{array}{ll} 0 < p < \overline{x}: & L'(p) > 0 \\ \overline{x} < p < 1: & L'(p) < 0 \end{array}\right\} \Rightarrow p_n = \frac{k}{n} = \overline{x} \text{ ist der MLE.}$$

4.3 Konfidenzintervall

Bemerkung 4.3.1. Eine Schätzung ist eine Näherung. Wie bei jedem Näherungsverfahren ist man am Fehler interessiert. Vgl. z.B. Numerische Integration:

$$\int_a^b f(x)\,dx \approx T \quad \text{Trapezformel}$$

Benutzen wir die Trapezformel zur numerischen Integration, dann berechnen wir das Integral nicht exakt. Wir begehen einen Fehler

$$R := \left| \int_a^b f(x)\,dx - T \right| \le \varepsilon$$

R, der durch ε beschränkt ist. Äquivalent hierzu ist die Darstellung

$$T - \varepsilon \leq \int_a^b f(x)\,dx \leq T + \varepsilon$$

aus der sich anschließend

$$\int_a^b f(x)\,dx \in [T - \varepsilon, T + \varepsilon] \text{ zu } 100\%$$

ergibt.

Das gleiche Konzept verfolgt man auch in der Statistik. Basierend auf einer Stichprobe $X_1, \ldots, X_n$ wird ein Intervall

$$I_n := [a_n, b_n],$$

das abhängig von von der Stichprobe ist, gesucht. Dieses Intervall soll den gesuchten, unbekannten Parameter λ mit einer vorgegebenen Sicherheit, z.B. $x\%$, enthält. Da I_n vom Zufall abängt, ist $x = 100\%$ unmöglich.

Definition 4.3.2. Sei $X_1, ..., X_n$ eine Stichprobe zu einer Verteilungsklasse mit Parameter λ und $0 < \alpha < 1$. Ist dann $I_n := [a_n, b_n]$ ein Intervall, das von der Stichprobe abhängt (also zufällig ist) und gilt

$$\mathbb{P}(\lambda \in I_n) \geq 1 - \alpha$$

dann nennt man I_n ein $100 \cdot (1 - \alpha)$ %-tiges Konfidenzintervall (KI) für λ und $1 - \alpha$ das Konfidenzniveau.

4.4 Konfidenzintervalle bei normalverteilter Grundgesamtheit

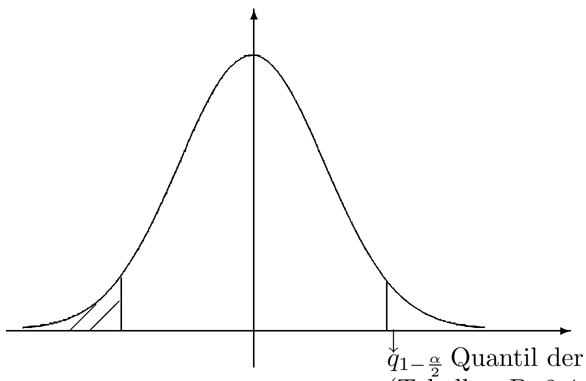
I.F. seien $X_1, \ldots, X_n$ unabhängig und identisch nach $\mathcal{N}(\mu, \sigma^2)$ verteilte Zufallsvariable. Wenn nichts anderes gesagt wird, bedeuten

$$\overline{X} := \frac{1}{n} \sum_{i=1}^n X_i, \quad s_n^2 := \frac{1}{n-1} \sum_{i=1}^n (X_i - \overline{X})^2$$

und $1 - \alpha$ bezeichnet das vorgegebene Konfidenzniveau, wobei $0 < \alpha < 1$.

Satz 4.4.1. KI von μ bei bekanntem σ^2. Man kann zeigen, dass

$$\sqrt{n}\, \frac{\overline{X} - \mu}{\sigma} \sim \mathcal{N}(0, 1)$$



$q_{1-\frac{\alpha}{2}}$ Quantil der $\mathcal{N}$(0,1)
(Tabelle z.B. 3.4.5.6.4)

Daraus ergibt sich

$$\mathbb{P}\left(-q_{1-\frac{\alpha}{2}} \leq \sqrt{n}\,\frac{\overline{X}-\mu}{\sigma} \leq q_{1-\frac{\alpha}{2}}\right) = 1-\alpha$$

Das gesuchte KI I_n erhält man dann durch folgende Umrechnung:

$$\begin{aligned} -q_{1-\frac{\alpha}{2}} \leq \sqrt{n}\,\frac{\overline{X}-\mu}{\sigma} \leq q_{1-\frac{\alpha}{2}} & \\ & \Leftrightarrow \quad -\frac{\sigma}{\sqrt{n}}\,q_{1-\frac{\alpha}{2}} \leq \overline{X}-\mu \leq \frac{\sigma}{\sqrt{n}}\,q_{1-\frac{\alpha}{2}} \\ & \Leftrightarrow \quad \overline{X}-\frac{\sigma}{\sqrt{n}}\,q_{1-\frac{\alpha}{2}} \leq \mu \leq \overline{X}+\frac{\sigma}{\sqrt{n}}\,q_{1-\frac{\alpha}{2}} \end{aligned}$$

Also:

$$I_n := \left[\overline{X}-\frac{\sigma}{\sqrt{n}}q_{1-\frac{\alpha}{2}}, \overline{X}+\frac{\sigma}{\sqrt{n}}q_{1-\frac{\alpha}{2}}\right]$$

Beispiel 4.4.2. Bestimme ein 95% - KI für den Erwartungswert μ einer $\mathcal{N}(\mu,\sigma^2)$ Verteilung, wobei $\sigma^2 = 9$ als bekannt vorausgesetzt wird. Die Stichprobe hatte den Umfang n= 100 und den Mittelwert $\overline{x} = 5$ ergeben.

$$\alpha = 0.05 \Rightarrow q_{1-0.025} = q_{0,975} = 1.96$$

Zur Bestimmung des Quantils wird die Tafel im Anhang benutzt.

$$I_{100} \quad = \quad \left[5-\frac{3}{10}\,1.96,\ 5+\frac{3}{10}\,1.96\right] = 5 \pm 0.588 = [4.412;\ 5.588]$$

In einigen Situationen möchte man vor der Datenerhebung wissen, wie groß der Stichprobenumfang n sein muss, damit man anschließend ein 95% KI einer bestimmten vorgegebenen Länge L, sagen wir der Länge $L = 0.4$, erhält. Beachte hierzu:

$$I_n = \left[\overline{x}-\frac{3}{\sqrt{n}}\,1.96,\ \overline{x}+\frac{3}{\sqrt{n}}\,1.96\right]$$

Damit erhalten wir für die Länge des Intervalls die Darstellung:

$$\begin{aligned} L = 2\,\frac{3}{\sqrt{n}}\,1.96 &= 0.4 \\ \Leftrightarrow n &= \left(\frac{2*3*1.96}{0.4}\right)^2 \\ &= 864.36 \end{aligned}$$

D.h. als Stichprobenumfang müssen wir $n = 865$ wählen.

Satz 4.4.3. KI von μ bei unbekanntem σ^2. Man kann zeigen, dass

$$\begin{aligned} A &= \sqrt{n}\frac{\overline{X}-\mu}{\sigma} \sim \mathcal{N}(0,1) \\ B &= \frac{1}{\sigma^2}\sum_{i=1}^{n}(X_i-\overline{X})^2 \sim \chi^2_{n-1} \quad \text{Chi-Quadrat mit } n-1 \text{ Freiheitsgraden} \end{aligned}$$

Darüber hinaus lässt sich nachweisen, dass A und B unabhängig sind. Demnach besitzt

$$T := \frac{A}{\sqrt{\frac{B}{n-1}}} \sim t_{n-1}$$

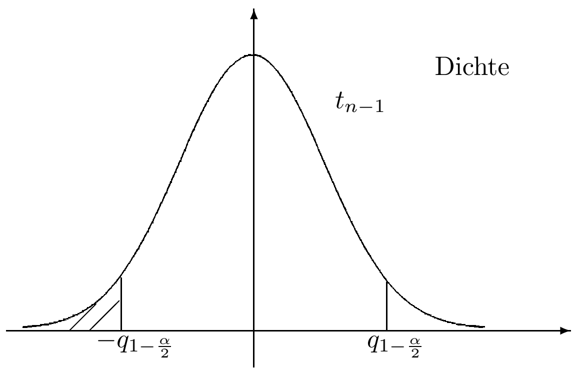
eine $t-$Verteilung mit $n-1$ Freiheitsgraden. Da aber

$$T = \frac{\sqrt{n}\,\frac{\overline{X}-\mu}{\sigma}}{\sqrt{\frac{1}{\sigma^2}\,\frac{1}{n-1}\,\sum\limits_{i=1}^{n}(X_i-\overline{X})^2}} = \sqrt{n}\frac{\overline{X}-\mu}{s_n}$$

erhalten wir

$$\sqrt{n}\frac{\overline{X}-\mu}{s_n} \sim t_{n-1}$$

und wir können jetzt ein entsprechendes KI bestimmen.



Bei geeigneter Wahl der Quantile (vgl. Tafel der $t-$Verteilung im Anhang) folgt damit:

$$\mathbb{P}\left(-q_{1-\frac{\alpha}{2}} \leq \sqrt{n}\frac{\overline{X}-\mu}{s_n} \leq q_{1-\frac{\alpha}{2}}\right) = 1-\alpha$$

Wie unter Satz 4.4.1 konstruiert man

$$I_n = \left[\overline{X} - \frac{s_n}{\sqrt{n}}\,q_{1-\frac{\alpha}{2}}, \overline{X} + \frac{s_n}{\sqrt{n}}\,q_{1-\frac{\alpha}{2}}\right]$$

Beispiel 4.4.4. Eine Stichprobe aus einer $\mathcal{N}(\mu,\sigma^2)$ Grundgesamtheit vom Umfang $n = 20$ hat die Werte

$$\overline{x} = 1.944 \quad \text{und} \quad s_n^2 = 0.00282 (s_n \approx 0.053)$$

ergeben. Gesucht wird ein 99% KI für μ, d.h. $\alpha = 0.01$. σ^2 ist nicht bekannt.

$$q_{1-\frac{\alpha}{2}} = q_{0.995} = 2.86 \quad (t_{19}\text{-Verteilung})$$

Also

$$\begin{aligned} I_{20} &= \left[\overline{x} - \frac{s_n}{\sqrt{20}}\,2.86;\, \overline{x} + \frac{s_n}{\sqrt{20}}\,2.86\right] = \left[1.944 - \frac{0.053}{4.47}\,2.86;\, 1.944 + \frac{0.053}{4.47}\,2.86\right] \\ &= [1.944 - 0.034;\, 1.944 + 0.034] \\ &= 1.944 \pm 0.034 \\ &= [1.91;\, 1.978] \end{aligned}$$

Wir stellen uns auch hier die Frage, wie groß muss n gewählt werden, damit man ein 99% KI einer festen Länge L, sagen wir $L = 0.05$ erhält?

$$I_n = \left[\overline{x} - \frac{s_n}{\sqrt{n}}\,q_{0.995}^{(n)};\ \overline{x} + \frac{s_n}{\sqrt{n}}\,q_{0.995}^{(n)}\right] \quad \Rightarrow \ldots n = \left(\frac{2\,s_n\,q_{0.995}^{(n)}}{0.05}\right)^2$$

Aber s_n ist jetzt vom Zufall abhähngig und $q_{0.995}(n)$ ist wegen t_{n-1}-Verteilung auch von n abhängig. Man kann daher nicht vor der Durchführung der Experimente sagen, wie viele man benötigt. Nach jedem Experiment muss man L berechnen. Ist L dann ≈ 0.05, so stoppt man mit der Datenerhebung (sequentielles Verfahren!)

Bemerkung 4.4.5. Für große Stichproben zeigt die Tabelle ein eigentümliches Verhalten:

$n =$	30	50	100	200	∞
$q_{0.95} =$	1.7	1.68	1.66	1.65	1.65

∞ steht dabei für die $\mathcal{N}$(0,1)-Verteilung, denn

$$t_n \underset{n\to\infty}{\longrightarrow} \mathcal{N}(0,1)$$

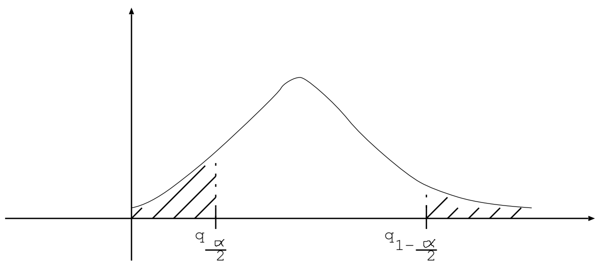
D.h. bei großem n kann man die $\mathcal{N}$(0,1) - Tafel zur Bestimmung der Quantile nutzen. Bei kleinem n muss die $t-$Tafel benutzt werden.

Satz 4.4.6. KI für σ^2. Man kann zeigen, dass

$$\frac{1}{\sigma^2}\sum_{i=1}^{n}(X_i-\overline{X})^2 \sim \chi^2_{n-1}$$

eine Chi-Quadrat Verteilung mit $n-1$ Freiheitsgraden besitzt. Also

$$\frac{n-1}{\sigma^2}\,s_n^2 \sim \chi^2_{n-1}$$



Ermittle nun $q_{\frac{\alpha}{2}}$ und $q_{1-\frac{\alpha}{2}}$ aus der Tabelle der χ^2_{n-1}-Verteilung (vgl. Anhang). Dann erhält man:

$$\mathbb{P}\left(q_{\frac{\alpha}{2}} \le \frac{n-1}{\sigma^2}\,s_n^2 \le q_{1-\frac{\alpha}{2}}\right) = \left(1-\frac{\alpha}{2}\right)-\frac{\alpha}{2} = 1-\alpha$$

Das KI für σ^2 ergibt sich dann aus der folgenden Umrechnung:

$$\begin{aligned} q_{\frac{\alpha}{2}} \le \frac{n-1}{\sigma^2}\,s_n^2 \le q_{1-\frac{\alpha}{2}} & \\ \Leftrightarrow \quad & \frac{q_{\frac{\alpha}{2}}}{(n-1)s_n^2} \le \frac{1}{\sigma^2} \le \frac{q_{1-\frac{\alpha}{2}}}{(n-1)s_n^2} \\ \Leftrightarrow \quad & \frac{(n-1)s_n^2}{q_{\frac{\alpha}{2}}} \ge \sigma^2 \ge \frac{(n-1)s_n^2}{q_{1-\frac{\alpha}{2}}} \end{aligned}$$

Also

$$I_n = \left[\frac{(n-1)s_n^2}{q_{1-\frac{\alpha}{2}}}, \frac{(n-1)s_n^2}{q_{\frac{\alpha}{2}}}\right]$$

Beispiel 4.4.7. Eine Stichprobe aus einer nomalverteilten Grundgesamtheit vom Umfang $n = 20$ hat den Wert

$$s_n^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \overline{x})^2 = 0.00282$$

ergeben. Gesucht wird ein 95% KI für σ^2. Also:

$$\alpha = 0.05; \; \frac{\alpha}{2} = 0.025; \; q_{0.025} \underset{\chi^2_{19}}{=} 8.91; q_{0.975} \underset{\chi^2_{19}}{=} 32.85$$

Damit ergibt sich für das KI:

$$I_{20} = \left[\frac{19 \cdot 0.00282}{32.85}; \frac{19 \cdot 0.00828}{8.91} \right] \approx [0.00163; \, 0.006]$$

4.5 Asymptotische Konfidenzintervalle

Der zentrale Grenzwertsatz (vgl. Kapitel 3.8) besagt

$$(+) \quad \mathbb{P}\left(\sqrt{n} \, \frac{\overline{X} - \mu}{\sqrt{s_n^2}} \leq z \right) \underset{n \to \infty}{\longrightarrow} \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z} \exp\left(-\frac{t^2}{2} \right) \mathrm{dt}$$

falls $X_1, ..., X_n$ unabhängig und identisch verteilte Zufallsvariable sind mit $\mathbb{E}(X) = \mu$ und $VAR(x) = \sigma^2 < \infty$. Diese Konvergenz gilt dann für jedes $z \in \mathbb{R}$. Ersetzt man in (+) $\sqrt{s_n^2}$ durch $\sqrt{\sigma^2}$, so konvergiert die linke Seite ebenfalls gegen den Wert auf der rechten Seite. (+) besagt, dass

$$\sqrt{n} \, \frac{\overline{X} - \mu}{\sqrt{s_n^2}} \quad \text{und} \quad \sqrt{n} \, \frac{\overline{X} - \mu}{\sqrt{\sigma^2}}$$

asymptotisch normalverteilt sind.

Definition 4.5.1. Seien $X_1, \ldots, X_n$ Zufallsvariable, die zu einer Verteilungsklasse gehören, die durch einen Parameter λ eindeutig bestimmt ist. Ferner sei $0 < \alpha < 1$. Ist dann $I_n := [a_n, b_n]$ ein Intervall, das von der Stichprobe abhängt und gilt

$$\lim_{n \to \infty} \mathbb{P}(\lambda \in I_n) \geq 1 - \alpha$$

dann nennt man I_n ein $100 \cdot (1 - \alpha)$%-tiges asymptotisches Konfidenzintervall (AKI). zum Niveau 1-α.

Beispiel 4.5.2. In Graz waren 1962 unter den ersten 3000 Lebendgeburten (Mehrlingsgeburten nicht mitgezählt) 1578 Jungen. Gesucht wird ein 99% AKI für die W. p des Ereignisses "Geburt eines Jungen".

Wir benutzen den ZGWS zur Konstruktion des AKI. Betrachte dazu die Zufallsvarialen $X_1, ..., X_{3000}$ mit

$$X_i := \begin{cases} 1 & : \text{i-te Geburt ist Junge} \\ 0 & : \text{i-te Geburt ist Mädchen} \end{cases}$$

Dann sind $X_1, \ldots, X_{3000}$ sind unabhängig und identisch nach $BERN(p)$ verteilt. Nach dem ZGWS ist dann

$$\sqrt{3000} \frac{\overline{X} - p}{\sqrt{s_{3000}^2}}$$

asymptotisch $\mathcal{N}(0,1)$ verteilt, da $\mathbb{E}(X) = p$. Aus der Tabelle für die Normalverteilung im Anhang entnehmen wir:

$$q_{1-\frac{\alpha}{2}} = q_{0.995} = 2.58 \quad q_{\frac{\alpha}{2}} = -2.58$$

und damit gilt:

$$\mathbb{P}\left(-2.58 \leq \sqrt{3000}\,\frac{\overline{X}-p}{\sqrt{s^2_{3000}}} \leq 2.58\right) \underset{\text{asympt.}}{\approx} 0.99$$

Das Intervall I_{3000} berechnet sich dann zu

$$I_{3000} = \left[\overline{x} - 2.58\,\frac{\sqrt{s^2_{3000}}}{\sqrt{3000}};\, \overline{x} + 2.58\,\frac{\sqrt{s^2_{3000}}}{\sqrt{3000}}\right]$$

Da

$$\begin{aligned} s^2_{3000} &= \frac{1}{3000}\sum_{i=1}^{3000} x_i^2 - (\overline{x})^2 \\ &= \frac{1578}{3000} - \left(\frac{1578}{3000}\right)^2 \\ &\approx 0.249 \end{aligned}$$

(denn $\overline{x} = \frac{1578}{3000} \approx 0.526$) ergibt sich für das Intervall

$$\begin{aligned} I_{3000} &= \left[0.526 - 2.58\,\frac{\sqrt{0.249}}{\sqrt{3000}};\, 0.526 + 2.58\frac{\sqrt{0.249}}{\sqrt{3000}}\right] \\ &\approx [0.526 - 0.024;\, 0.526 + 0.024] \\ &= [0.502;\, 0.55] \end{aligned}$$

Man beachte, $p = 0.5 \notin I_{3000}$!

Bemerkung 4.5.3. Diese Konstruktion eines AKI lässt sich für den Erwartungswert $\mu = \mathbb{E}(\mathrm{x})$ über den ZGWS für jede Folge von unabhängig und identisch verteilten Zufallsvariable $X_1, \ldots, X_n$ durchführen, falls $VAR(X) = \sigma^2 < \infty$ ist. Falls die Stichprobe eine symmetrische Verteilung vermuten lässt, dann liefert $n \geq 30$ schon gute Ergebnisse für das AKI des Erwartungswertes.

Kapitel 5

Testtheorie

5.1 Allgemeines zur Testtheorie

Beispiel 5.1.1. In der Literatur wird behauptet, dass mehr Jungen als Mädchen geboren werden. Als Begruündung wird angegeben, dass Neugeborene Jungen eine höhere Sterberate haben. Allgemein sollte man annehmen:

$$H_0\text{:}\quad \text{Jungen-Mädchengeburten sind gleichhäufig.}$$

Wie kann man die Behauptung:

$$H_1\text{:}\quad \text{Es werden mehr Jungen als Mädchen geboren}$$

begründen?

Betrachtet man die Stichprobe: Graz 1962 von 3000 Geburten 1578 Jungen. Dann stellt sich die Frage: Sind die 1578 ein Indiz für H_1, oder ist diese Zahl im Rahmen der zufälligen Schwankungen? Man wird die Nullhypothese H_0 zugunsten der Alternative H_1 ablehnen, falls

$$X > c$$

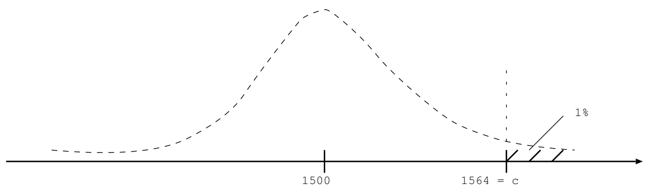
wobei X = Anzahl der Jungen. Es bleibt die Frage offen: Wie ist der kritische Wert, d.h. die Konstante c zu wählen?

Natürlich können wir keine 100% sichere Auswahl treffen, denn X kann alle Werte $0, \ldots, 3000$ annehmen, egal ob H_0 oder H_1 richtig ist. Wir müssen also wieder einen kontrollierten Fehler α, Fehler 1.Art, zulassen (Üblicherweise $\alpha = 1\%$ oder 5%)

Unter dem Fehler 1.Art verstehen wir die Wahrscheinlichkeit, dass H_0 verworfen wird, obwohl H_0 korrekt ist. Nehmen wir an, $\alpha = 1\%$. Dann können wir die Nullhypothese bzw. die Alternative formal durch

$$\begin{aligned} H_0\text{:}&\quad p = 0.5 = p_0 \\ H_1\text{:}&\quad p > 0.5 = p_0 \end{aligned}$$

ersetzen. Ferner ist $X \sim BIN(3000, p)$, d.h. p ist der wahre Parameter.



Unter H_0, d.h. $p = 0.5$, gilt $X \sim BIN(3000; 0.5)$ Gemäß ZGWS folgt für $n \to \infty$

$$\sqrt{3000}\,\frac{\frac{X}{3000} - 0.5}{\sqrt{0.5(1-0.5)}} \sim \mathcal{N}(0,1)$$

Also erhalten wir

$$\mathbb{P}\left(\sqrt{3000}\,\frac{\frac{X}{3000} - 0.5}{0.5} > c_1\right) \approx 0.01$$

aus der Tabelle der $\mathcal{N}(0,1)-$ Verteilung für c_1 den Wert 2.33.

$$\begin{aligned}
\Rightarrow \quad \sqrt{3000}\,\frac{\frac{X}{3000}-0.5}{0.5} &> c_1 \\
\Leftrightarrow \quad \frac{X}{3000} &> \frac{0.5\,c_1}{\sqrt{3000}} + 0.5 \\
\Leftrightarrow \quad X &> c_1\,0.5\,\sqrt{3000} + 1500 \approx 1564 =: c
\end{aligned}$$

D.h., falls $X > 1564$, können wir H_0 ablehnen. Der Fehler, den wir dabei begehen, ist kleiner als 1%. Damit sind die $1578 = X$ neugeborenen Jungen ein Indiz dafür, dass mehr Jungen als Mädchen geborgen werden. Sie sprechen eher für die Alternative H_1 als für H_0.

5.2 Grundlagen und Definitionen zur Testtheorie

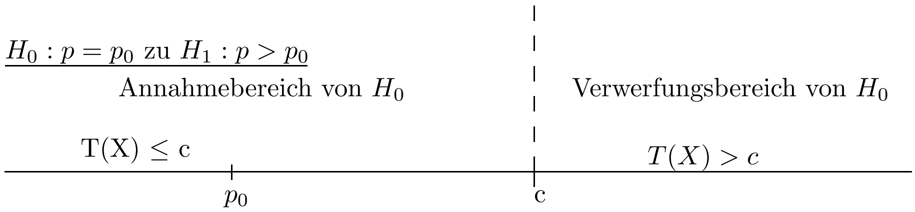
Gegeben sei eine Zufallsvariable X, deren Verteilung aus einer fest vorgegebenen Klasse von Verteilungen stammt und die eindeutig durch einen unbekannten Parameter p bestimmt ist. Über diesen Prameter hat man aufgrund langjähriger Erfahrung oder Sicherheitsbestimmungen eine Vorstellung, genannt die Nullhypothese.

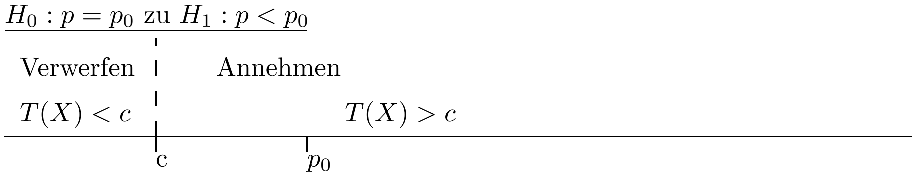
$$H_0: \qquad p = p_0$$

Aus der Problemstellung ergibt sich dann eine Alternative

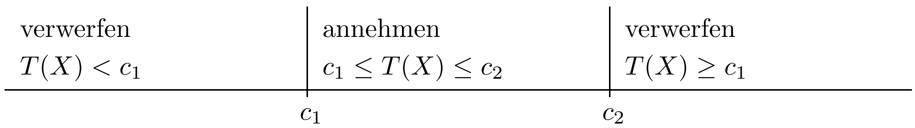
$$H_1: \qquad p > p_0 \quad \text{oder} \quad H_1: \qquad p < p_0$$

als einseitige Alternative bzw. $H_1: p \neq p_0$ als zweiseitige Alternative. p ist dabei der wahre und p_0 der vermutete Parameter. Ferner ist der Fehler 1.Art $0 < \alpha < 1$ vorgegeben. α ist die Wahrscheinlichkeit H_0 abzulehnen, obwohl H_0 korrekt ist. Handelt es sich um einen einseitigen Test, so wird zu α ein kritischer Wert c bestimmt, der den Verwerfungsbereich von H_0 definiert.





Im zweiseitigen Test müssen zwei kritische Werte c_1, c_2 bestimmt werden. $H_0 : p = p_0$ gegen $H_1 : p \neq p_0$ Die Testgröße bezeichnen wir mit $T(X)$, eigentlich mit

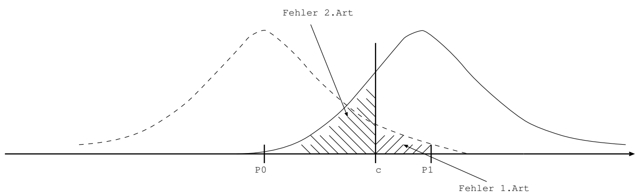


$$T(X) = \underbrace{T(X_1, ..., X_n)}_{n \text{ Experimente}} .$$

Neben dem Fehler 1.Art α gibt es auch einen Fehler 2.Art β. Unter diesem Fehler verstehen wir die Wahrscheinlichkeit H_0 anzunehmen, obwohl H_1 richtig ist.

		Unbekannte Wirklichkeit	
$H_0 : p = p_0$ $H_1 : p = p_1$		H_0 korrekt $p = p_0$	H_1 korrekt $p = p_1$
Entsch.	$p = p_0$	richtige Entsch. $\mathbb{P} = 1 - \alpha$	falsche Entsch. $\mathbb{P} = \beta$ Fehler 2.Art
	$p = p_1$	falsche Entsch. $\mathbb{P} = \alpha$ Fehler 1.Art	richtige Entsch. $\mathbb{P} = 1 - \beta$

Der Wert $1 - \beta$ wird die Macht des Test $T(X)$ genannt. $1 - \beta(p)$ entspricht der Wahrscheinlichkeit H_0 abzulehnen wenn p korrekt ist. Diese Funktion (in p) nennt man die Gütefunktion des Tests. Fehler 1. und 2. Art werden beim Testen unterschiedlich behandelt. Der kritische Wert c muss so gewählt werden, dass α und β möglichst klein werden. Wie die Skizze zeigt, ist eine Verkleinerung von α, d.h. $c \to$ rechts, stets durch eine Vergrößerung von β begleitet und umgekehrt. Bei der Wahl von c muss unbedingt der α-Fehler gehalten werden, egal wie groß der β-Fehler wird. Unter Beibehaltung des α-Fehlers kann man den β-Fehler durch Erhöhung des Stichprobenumfangs verkleinern.



5.3 Test bei Normalverteilung

Allgemeine Voraussetzungen in diesem Abschnitt:

$$X_1, ..., X_n \sim \mathcal{N}(\mu, \sigma^2)$$

unabhängige Zufallsvariable und bezeichne mit α bzw. β den Fehler 1. bzw. 2. Art. Ferner sei

$$\overline{X} := \frac{1}{n}\sum_{i=1}^{n} X_i \quad , \quad s_n^2 = s_x^2 := \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2$$

Satz 5.3.1. Test auf theoretischen Mittelwert.

Vor.:	Daten $\mathcal{N}(\mu, \sigma^2)$ verteilt
Frage:	Ist $\mu = \mu_t$, wobei μ_t theoretischer Mittelwert?
Prüfstatistik:	$t = \sqrt{n}\,\frac{\overline{X}-\mu_t}{s_n}$
Test:	$H_0 : \mu = \mu_t$ gegen $H_1^1 : \mu > \mu_t$ oder $H_1^2 : \mu < \mu_t$ oder (einseitig) $H_1^3 : \mu \neq \mu_t$ (zweiseitig) Der Test soll zum Niveau α durchgeführt werden.
Kritischer Bereich:	H_0 gegen $H_1^1 : \mu > \mu_t \Rightarrow H_0$ wird abgelehnt, falls $t > K = q_{1-\alpha}$ H_0 gegen $H_1^2 : \mu < \mu_t \Rightarrow H_0$ wird abgelehnt, falls $t < K = q_\alpha = q_{1-\alpha}$ H_0 gegen $H_1^3 : \mu \neq \mu_t \Rightarrow H_0$ wird abgelehnt, falls $t > K = q_{1-\frac{\alpha}{2}}$ oder $t < K = -q_{1-\frac{\alpha}{2}} \Leftrightarrow \lvert t\rvert > K = q_{1-\frac{\alpha}{2}}$ $q_{1-\alpha}$ ist das $1-\alpha$ Quantil der t-Verteilung mit $n-1$ Freiheitsgraden (kurz t_{n-1} - Verteilung)

Beispiel 5.3.2. Eine Stichprobe aus einer $\mathcal{N}(\mu, \sigma^2)$ verteilten Grundgesamtheit ergab den Mittelwert $\overline{x} = 47.0$. Aus den 20 Messwerten wurde $s_n^2 = 25$ berechnet. Aus theoretischen Gründen wird $\mu_t = 45.0$ vermutet.
Es soll getestet werden: $H_0 : \mu = 45$ gegen $H_1 : \mu > 45$, $\alpha = 0.05$.

Als Prüfgröße wählen wir:

$$t = \sqrt{20}\,\frac{47-45}{5} = 2.68$$

Den kritischen Bereich ermitteln wir aus der Tabelle zu

$$K = q_{1-\alpha} = q_{0.95} = 1.729$$

(95% Quantil der t_{19}-Verteilung)

Da $2.68 > 1.729$, muss H_0 abgelehnt werden, d.h. μ ist signifikant größer als 45.

Bemerkung 5.3.3. Ist die tatsächliche Varianz σ^2 bekannt, dann wird unter Satz 5.3.1 in t s_n durch σ ersetzt. Der kritische Wert K wird dann über das entsprechende Quantil $q_{1-\alpha}$ der Standardnormalverteilung $\mathcal{N}(0,1)$ und nicht mehr über das Quantil der t-Verteilung bestimmt.

Satz 5.3.4. Test zweier (unabhängiger) normalverteilter Stichproben auf Gleichheit der Mittelwerte, gleiche Varianz vorausgesetzt.

Vor.: Gegeben seien zwei Stichproben

$$X_1, ..., Xn_1 \sim \mathcal{N}(\mu_x, \sigma^2), \quad Y_1, ..., Yn_2 \sim \mathcal{N}(\mu_y, \sigma^2)$$

d.h. die Varianzen müssen identisch sein!

Fragestellung: Ist $\mu_x = \mu_y$?

Prüfstatistik:

$$t = \sqrt{\frac{n_1 n_2}{n_1 + n_2}} \frac{\overline{X} - \overline{Y}}{s_D}$$

wobei

$$s_D := \sqrt{\frac{1}{n_1 + n_2 - 2} \left(\sum_{i=1}^{n_1} (X_i - \overline{X})^2 + \sum_{j=1}^{n_2} (Y_j - \overline{Y})^2 \right)}$$

Test: H_0: $\mu_x = \mu_y$ gegen
$H_1^1 : \mu_x > \mu_t$ oder $H_1^2 : \mu_x < \mu_y$ (einseitig) oder
$H_1^3 : \mu_x \neq \mu_y$ (zweiseitig)
Der Test soll zum Niveau α durchgeführt werden.

Kritischer Bereich: Beachte, unter H_0 besitzt t eine t-Verteilung mit $n_1 + n_2 - 2$ Freiheitsgraden.
H_0 gegen H_1^1: $\mu_x > \mu_y \Rightarrow H_0$ wird abgelehnt, falls t > K = $q_{1-\alpha}$
H_0 gegen H_1^2: $\mu_x > \mu_y \Rightarrow H_0$ wird abgelehnt, falls t < K = q_α = -$q_{1-\alpha}$
H_0 gegen H_1^3: $\mu_x \neq \mu_y \Rightarrow H_0$ wird abgelehnt, falls $|t|$ > K = $q_{1-\frac{\alpha}{2}}$

$q_{1-\alpha}$ ist das 1-α Quantil der t-Verteilung mit $n_1 + n_2 - 2$ Freiheitsgraden.

Beispiel 5.3.5. Gegeben seien die folgenden Werte zu einem Zweistichprobenproblem: $n_1 = 16$, $\overline{x} = 14.5$, $s_x^2 = 4$ und $n_2 = 14$, $\overline{y} = 13.0$, $s_y^2 = 3$. Es ist bekannt, dass die Daten aus $\mathcal{N}(\mu_x, \sigma^2)$ bzw. $\mathcal{N}(\mu_y, \sigma^2)$ stammen (Varianzen identisch!). Zum Niveau $\alpha = 0.05$ soll $H_0 : \mu_x = \mu_y$ gegen $H_1 : \mu_x \neq \mu_y$ getestet werden.

Es handelt sich hier um eine zweiseitige Fragestellung! $\Rightarrow q_{1-\frac{\alpha}{2}} = q_{0.975} = 2.048$ - 97.5% Quantil der $t_{n_1+n_2-2} = t_{28}$ Verteilung (28 Freiheitsgrade).
Berechnung der Prüfgröße:

$$\begin{aligned} s_D^2 &= \frac{1}{n_1 + n_2 - 2} \left((n_1 - 1)s_x^2 + (n_2 - 1)s_y^2 \right) \\ &= \frac{15 * 4 + 13 * 3}{28} = \frac{99}{28} \\ \Rightarrow \quad & s_D = \sqrt{\frac{99}{28}} = 1.88 \end{aligned}$$

Also

$$t = \sqrt{\frac{16 * 14}{16 + 14}} \frac{14.5 - 13.0}{1.88} = 2.180$$

Da $|t| = 2.180 > 2.048 = q_{0.975}$ muss $H_0 : \mu_x = \mu_y$ zugunsten von $\mu_x \neq \mu_y$ verworfen werden. μ_x ist signifikant von μ_y verschieden.

Satz 5.3.6. Test zweier (verbundener) Stichproben auf Gleichheit der Mittelwerte.

Vor.:	Gegeben seien $(X_1, Y_1), ..., (X_n, Y_n)$ derart, dass $d_1, ..., d_n \sim \mathcal{N}(\mu, \sigma^2)$. Dabei bezeichnet $$d_i := X_i - Y_i$$ die jeweilige Differenz des Beobachtungspaares. Bezeichne weiter μ_x bzw. μ_y den theoretischen Mittelwert der $X-$ (bzw. $Y-$)Daten.
Fragestellung:	Ist $\mu_x = \mu_y$, d.h. $\mu_d = 0$?
Prüfstatistik:	$$t = \sqrt{n}\,\frac{\overline{d}}{s_d}$$ wobei $\overline{d}$ das arithmetische Mittel der d_i's und s_d^2 die zugehörige Stichprobenvarianz bezeichnen.
Test:	$H_0 : \mu_x = \mu_y (\mu_d = 0)$ gegen $H_1^1 : \mu_x > \mu_y (\mu_d > 0)$ oder $H_1^2 : \mu_x < \mu_y (\mu_d < 0)$ $H_1^3 : \mu_x \neq \mu_y (\mu_d \neq 0)$
Kritischer Bereich:	Hier muss für $q_{1-\alpha}$ das $(1-\alpha)$ Quantil der t_{n-1}-Verteilung benutzt werden.

Beispiel 5.3.7. In der folgenden Tabelle ist der Ertrag von 8 Kirschbäumen in zwei Jahren, die sich hinsichtlich der Witterung stark unterschieden haben, aufgelistet.

Baum	1	2	3	4	5	6	7	8
Jahr x	36	31.5	34	32.5	35	31.5	31	35.5
Jahr y	34	35.5	33.5	36	39	35	33	39.5
Diff d	2	-4	0.5	-3.5	-4	-3.5	-2	-4

Die Stichprobe ist verbunden, da die Erträge (X, Y) am selben Baum gemessen wurden! Es soll zum Signifikanzniveau $\alpha = 0.05$ H_0: $\mu_d = 0$ gegen $H_1 : \mu_d \neq 0$ getestet werden.

$$\begin{aligned}
\overline{d} &= \frac{1}{8}(2 - 4 \ldots - 4) = -2.3125 \approx -2.31 \\
s_d &= \sqrt{\frac{1}{7}\sum_{i=1}^{8}(d_i - \overline{d})^2} \approx 2.33 \\
t &= \sqrt{8}\,\frac{-2.31}{2.33} \approx -2.8
\end{aligned}$$

Es handelt sich hier um einen zweiseitige Fragestellung, also erhalten wir als Quantil:

$$q_{1-\frac{\alpha}{2}} = q_{0.975} = 2.365, \quad 97.5\% \text{ Quantil der } t_7 \text{ - Verteilung}$$

Da $|t| > 2.365$ muss H_0 abgelehnt werden. Also führen die Witterungseinflüsse zu signifikanten Ertragsunterschieden.

Satz 5.3.8. Test zweier Stichproben auf Gleichheit der Varianzen.

Vor.: Gegeben seien zwei unabhängige Stichproben; $X_1, ..., X_{n_1}$ Stichprobe zur ersten Grundgesamtheit $\sim \mathcal{N}(\mu_x, \sigma_x^2)$ und $Y_1, ..., Y_{n_2}$ Stichprobe zur zweiten Grundgesamtheit $\sim \mathcal{N}(\mu_y, \sigma_y^2)$

Frage: Ist $\sigma_x^2 = \sigma_y^2$?

Ptüfstatistik:

$$F := \begin{cases} s_x^2/s_y^2 & \text{falls} \quad s_x^2 \geq s_y^2 \\ s_y^2/s_x^2 & \text{falls} \quad s_y^2 > s_x^2 \end{cases}$$

Test: $H_0 : \sigma_x^2 = \sigma_y^2$
$H_1 : \sigma_x^2 \neq \sigma_y^2$ zweiseitig
α - Signifikanzniveau

Kritischer Bereich: Zur Verteilung von F beachte man:

$$\frac{n_1 - 1}{\sigma_x^2} s_x^2 \sim \chi^2 \quad \text{mit} \quad n_1 - 1 \quad \text{FG;} \qquad \frac{n_2 - 1}{\sigma_y^2} s_y^2 \sim \chi^2 \quad \text{mit} \quad n_2 - 1 \quad \text{FG}$$

Auf Grund der Unabhängigkeit der Stichproben erhalten wir dann:

$$\frac{\frac{\left(\frac{n_1 - 1}{\sigma_x^2} s_x^2\right)}{(n_1 - 1)}}{\frac{\left(\frac{n_2 - 1}{\sigma_y^2} s_y^2\right)}{(n_2 - 1)}} = \frac{\sigma_y^2}{\sigma_x^2} \frac{s_x^2}{s_y^2} \underset{H_0!}{=} \frac{s_x^2}{s_y^2}$$

besitzt eine F_{n_1-1,n_2-1} Verteilung, d.h. F-Verteilung mit (n_1-1, n_2-1) FG. Zur Bestimmung des kritischen Bereichs muss hier $q_{1-\frac{\alpha}{2}}$, als $1 - \frac{\alpha}{2}$ Quantil der F-Verteilung benutzt werden. Falls $s_x^2 \geq s_y^2$, dann wird die F_{n_1-1,n_2-1}-Verteilung benutzt sonst die F_{n_2-1,n_1-1}-Verteilung.
H_0 gegen H_1: $\sigma_x^2 \neq \sigma_y^2 \Rightarrow H_0$ ablehnen, falls $F > q_{1-\frac{\alpha}{2}}$.

Beispiel 5.3.9. (Vgl. Beispiel 5.3.5). $n_1 = 16$, $s_x^2 = 4$, $n_2 = 14$ und $s_y^2 = 3$.

Teste zum Niveau $\alpha = 0.05$: H_0: $\sigma_x^2 = \sigma_y^2$ gegen H_1: $\sigma_x^2 \neq \sigma_y^2$

$$s_x^2 = 4 > 3 = s_y^2 \Rightarrow F = \frac{s_x^2}{s_y^2} = \frac{4}{3}$$

$$q_{1-\frac{\alpha}{2}} = q_{0.975}(F_{n_1-1,n_2-1}) = q_{0.975}(F_{15,13}) = 3.05$$

Da $F = \frac{4}{3} < 3.05 = q_{0.975}$, besitzen die Stichproben keine signifikant verschiedenen Varianzen.

Bemerkung 5.3.10. Will man im Zweistichprobenproblem $\mu_x = \mu_y$ testen, so sollte man zuvor die Stichproben auf Gleichheit der Varianzen untersuchen, denn ersterer Test setzt gleiche Varianzen voraus!

Satz 5.3.11. Test des Maßkorrelationskoeffizienten.

Vor.:	$(X_1, Y_1), ..., (X_n, Y_n)$ seien normalverteilte Wertepaare.
Frage:	Ist der Korrelationskoeffizient

$$r_{X,Y} = \frac{\sum\limits_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})}{\sqrt{\sum\limits_{i=1}^{n}(X_i - \overline{X})\sum\limits_{j=1}^{n}(Y_j - \overline{Y})}}$$

signifikant von 0 verschieden?

Prüfstatistik:

$$t = \frac{|r_{X,Y}|}{\sqrt{1 - r_{X,Y}^2}}\sqrt{n-2}$$

Test:	H_0: Korrelation $= 0$ H_1: Korrelation $\neq 0$ (zweiseitig) α vorgegebenes Signifikanzniveau
Kritischer Bereich:	Hier muss $q_{1-\frac{\alpha}{2}}$, das $1 - \frac{\alpha}{2}$ Quantil der t-Verteilung mit $n - 2$ FG benutzt werden. H_0 wird abgelehnt, falls $t > q_{1-\frac{\alpha}{2}}$ ist.

Beispiel 5.3.12. (Vgl. Beispiel 2.4.1 und 2.5.3 "Größe und Gewicht bei 20 Schülern"). Teste zum Niveau $\alpha = 0.05$ die Korrelation. Wie schon in den angeführten Beispielen ermittelt, gilt $r_{x,y} = 0.73$ und $n = 20$. Also erhalten wir:

$$t = \frac{|r_{x,y}|}{\sqrt{1 - r_{x,y}^2}}\sqrt{n-2} = \sqrt{18}\,\frac{0.73}{\sqrt{1 - 0.73^2}} \approx 4.53$$

Ferner ist nach der $t_{18}-$Tabelle im Anhang $q_{1-\frac{\alpha}{2}} = q_{0.975} = 2.101$. Da $t = 4.53 > 2.101 = q_{0.975}$ ist die Korrelation zwischen den Merkmalen Größe-Gewicht signifikant von 0 verschieden, d.h. es besteht ein Zusammenhang zwischen diesen Merkmalen.

5.4 Der χ^2-Test

Liegen nominalskalierte Daten vor, kann man lediglich mit Häufigkeiten arbeiten. Mit dem sogenannten χ^2-Anpassungstest lässt sich dann prüfen, ob diese Häufigkeiten sich so verteilen wie eine von uns erwartete Verteilung.

Satz 5.4.1. Ein Merkmal habe n Ausprägungen. B_i sei die absolute Häufigkeit des $i-$ten Merkmals und E_i die erwartete Häufigkeit des Merkmals i. Ein Maß für den Grad der Übereinstimmung wird durch

$$\chi^2 = \sum_{i=1}^{n}\frac{(B_i - E_i)^2}{E_i} = \sum_{i=1}^{n}\frac{B_i^2}{E_i} - N$$

gegeben. Hier bezeichnet N den Stichprobenumfang. Falls die erwarteten Häufigkeiten den tatsächlichen entsprechen, dann ist die Größe χ^2 näherungsweise χ^2-verteilt mit $n-1-g$ Freiheitsgraden. Dies kann überprüft werden, indem man in der χ^2-Tabelle den Wert $\chi^2_{Tab}(n-1-g, \alpha) = (1-\alpha)$ Quantil der $\chi^2_{n-1-g}-$Verteilung abliest und ihn mit dem berechneten Wert χ^2 vergleicht. Ist der berechnete Wert größer als der Wert aus der Tabelle, dann wird H_0 abgelehnt, d.h. die Daten zeigen signifikante Unterschiede zu den erwarteten Häufigkeiten. Bei der Berechnung der Freiheitsgrade bezeichnet

g - Anzahl aus den Daten geschätzter Parameter.

Beispiel 5.4.2. Kein Parameter wird aus den Daten gechätzt (g = 0)! Nach den Mendelschen Gesetzen erwartet man bei einem Kreuzungsversuch zwischen Tieren mit normalen und braunen Augen in der

2.ten Generation ein Verhältnis von 3:1. Weichen die folgenden Daten signifikant ($\alpha = 5\%$) von diesem Verhältnis ab?

i	Phänotyp	Häufigkeiten beob. B_i	erw. E_i	$\frac{B_i^2}{E_i}$
1	braun	273	$1010\,\frac{1}{4} = 252.5$	295.16
2	normal	737	$1010\,\frac{3}{4} = 757.5$	717.05
		N = 1010		1012.21

$$\begin{aligned}
\chi^2 &= 1012.21 - 1010 = 2.21 \\
\text{FG} &= n - 1 - g = 2 - 1 - 0 = 1 \\
\chi^2_{Tab}(1; 5\%) &= 3.84 (= q_{0.95}\chi^2 - \text{Vert.})
\end{aligned}$$

Da $\chi^2 = 2.21 < 3.84 = \chi^2_{Tab}(1;\ 5\%)$ kann H_0 nicht zurückgewiesen werden.

Beispiel 5.4.3. Ein Parameter wird geschätzt (g = 1)! Man hat ein Arzneimittel in bestimmter Konzentration Blutkulturen zugefügt. Die induzierten Chromosomenbrüche pro Zelle sollen auf Poisson-Verteilung geprüft werden. ($\alpha = 5\%$)

Brüche pro Zelle k	0	1	2	3	4	5	6	7-9	10	$\sum$
Anzahl der Zellen	14	28	26	18	10	2	1	0	1	100

$N = 100$ untersuchte Zelle.

$$p(k) = \frac{\lambda^k}{k!} e^{-\lambda} \quad - \quad \text{Pois}(\lambda)\text{- Verteilung}$$

gibt den Anteil mit k C.Br. an. λ muss geschätzt werden. MLE liefert $\hat{\lambda} = \overline{X}$ Mittelwert der Brüche pro Zelle.

$$\overline{X} = \frac{(0 * 14 + 1 * 28 + 2 * 26 + ... + 10 * 1)}{100} = \frac{200}{100} = 2$$

Damit erhalten wir:

k =	0	1	2	3	4	5	≥ 6
B_k =	14	28	26	18	10	2	2
$N * p(k) = E_k$ =	13.53	27.07	27.04	18.04	9.02	3.61	1.65
$\frac{B_k^2}{E_k}$	14.49	28.96	24.97	17.96	11.09	1.11	2.42

Zusammengefasst folgt:

$$\begin{aligned}
\chi^2 &= \sum_{k=0}^{6} \frac{B_k^2}{E_k} - 100 = 101 - 100 = 1 \\
FG &= n - 1 - g = 7 - 1 - 1 = 5
\end{aligned}$$

Für den Wert aus der Tabelle ergibt sich

$$\chi^2_{Tab}(FG; 5\%) = \chi^2_{Tab}(5; 5\%) = 11.07$$

Also folgt H_0: *Anzahl der C. Brüche gehorchen einer Pois-Vert.* muss beibehalten werden.

Kapitel 6

Lineare Regression

6.1 Allgemeines

In diesem Kapitel gehen wir allgemein davon aus, dass wir

$$(X_1, Y_1), \ldots, (X_n, Y_n);$$

unabhängige Zufallsvektoren vorliegen haben, für die zusätzlich noch

$$Y = a + bX + \varepsilon$$

gilt. Dabei ist $\varepsilon \sim \mathcal{N}(0, \sigma^2)$ und unabhängig von X. D.h., messen wir für X den Wert x, dann folgt:

$$Y \sim \mathcal{N}(a + bx, \sigma^2)$$

Bei diesem (linearen) Modell nehmen wir ferner an, dass x exakt gemessen wird!

Als Schätzer für a und b haben wir nach der Methode der kleinsten Quadrate

$$\widehat{a} = \overline{Y} - \widehat{b}\overline{X}, \qquad \widehat{b} = \frac{s_{xy}}{s_x^2}.$$

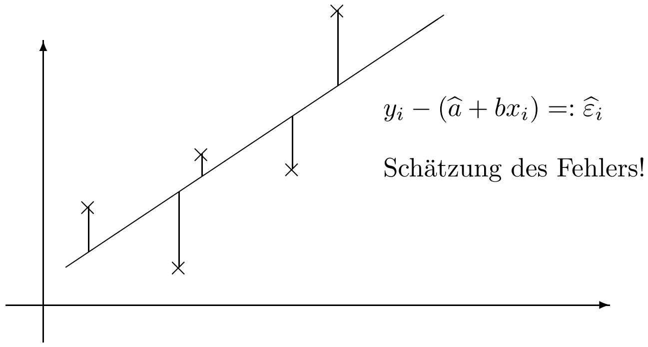
mit

$$s_{xy} := \frac{1}{n-1} \sum_{i=1}^{n} (Y_i - \overline{Y})(X_i - \overline{X})$$

errechnet.

Für die Summe der (geschätzten) Fehlerquadrate, hier mit R bezeichnet, gilt dann

$$\begin{aligned}
R &= \sum_{i=1}^{n} \Big(Y_i - (\widehat{a} + \widehat{b}X_i)\Big)^2 \\
&= \sum_{i=1}^{n} \Big(Y_i - (\overline{Y} - \widehat{b}\overline{X} + \widehat{b}X_i)\Big)^2 = \sum_{i=1}^{n} \Big((Y_i - \overline{Y}) - (\widehat{b}X_i - \widehat{b}\overline{X})\Big)^2 \\
&= (n-1)s_y^2 - 2\widehat{b}(n-1)s_{x,y} + \widehat{b}^2(n-1)s_x^2 \\
&= (n-1)s_y^2 - 2\widehat{b}(n-1)\widehat{b}\,s_x^2 + \widehat{b}^2(n-1)s_x^2 \\
&= (n-1)(s_y^2 - \widehat{b}^2 s_x^2)
\end{aligned}$$



Satz 6.1.1. Sei $g(x) := a + bx$ die "wahre" Gerade und $\widehat{g}(x) := \widehat{a} - \widehat{b}(x)$ die geschätzte Gerade. Dann gilt:

i) Die geschätzte Gerade $\widehat{g}$ verläuft durch den Punkt $(\overline{x}, \overline{y})$, denn

$$\widehat{g}(\overline{x}) = \widehat{a} - \widehat{b\overline{x}} = (\overline{y} - \widehat{b\overline{x}}) + \widehat{b\overline{x}} = \overline{y}$$

ii) Unter den allgemeinen Voraussetzungen besitzt

$$T = s_x \sqrt{(n-1)(n-2)}\,\frac{\widehat{b} - b}{\sqrt{R}}$$

eine t_{n-2}-Verteilung (t-Vert. mit n-2 FG).

6.2 Testtheorie

Der Titel dieses Abschnitts ist insofern etwas übertrieben, als dass wir uns hier nicht mit der allgemeinen Testtheorie beschäftigen werden, sondern nur einen Test bzgl. der Steigung b diskutieren werden.

Nehmen wir an, dass die Steigung der wahren Geraden b sei und wir möchten testen, ob $b > b_0$ ist, wobei b_0 ein theoretischer Wert ist. D.h. in der Notation des letzten Kapitels:

$$H_0\colon b = b_0 \text{ gegen } H_1\colon b > b_0 \text{ zum Niveau } \alpha$$

Wir berechnen für dieses Testproblem den Wert

$$T = s_x \sqrt{(n-1)(n-2)}\,\frac{\widehat{b} - b_0}{\sqrt{R}}$$

und bestimmen anschließend

$$c := q_{1-\alpha}(1-\alpha) \text{ Quantil der } t_{n-2}\text{-Vert.}$$

aus der entsprechenden Tabelle im Anhang. Ist $T > c$ wird H_0 abgelehnt, sonst beibehalten. Dieser Test lässt sich natürlich auch entsprechenden anderen Alternativen anpassen, z.B. H_1: $b < b_0$ oder H_1: $b \neq b_0$ (zweiseitig!).

Beispiel 6.2.1. Erzeugung einer Quecksilberemulsion in Natriumzitratlösung durch Ultraschall. $X-$Beschallungsdauer in s, $Y-$dispergierte Phase in g/l

x =	15	30	60	120	300	600
y =	4	6.3	5.2	6.4	5.5	6.4

Es darf angenommen werden, dass die allgemeinen Voraussetzungen erfüllt sind. Getestet werden soll $b = 0$ gegen $b > 0$ bei $\alpha = 5\%$

Wir erhalten dann: $\overline{x} = 187.5$; $\overline{y} = 5.63$; $s_x^2 \approx 51637.5$; $s_y^2 \approx 0.9$; $s_{x,y} \approx 96.3$; $\widehat{b} = \frac{s_{xy}}{s_x^2} = 0.0019$; R = $5(s_y^2 - \widehat{b}^2 s_x^2) \approx 3.6$ Also gilt:

$$T = s_x \sqrt{20} \frac{b}{\sqrt{R}} \approx 1$$

Da $q_{1-0.05} = q_{0.95} = 2.13$ $(t_{n-2} = t_4)$ Quantil. D.h. H_0 kann nicht verworfen werden! Anscheinend hat die Beschallungsdauer keinen Einfluss auf die Emulsion.

6.3 Schätztheorie

Aus Satz 6.1.1 erhält man durch

$$\widehat{b} \mp \frac{\sqrt{R}}{s_x \sqrt{(n-1)(n-2)}} q_{1-\frac{\alpha}{2}}$$

mit $q_{1-\frac{\alpha}{2}}$, $1 - \frac{\alpha}{2}$Quantil der t_{n-2}- Verteilung, ein 100(1-α)% KI für b.

Beispiel 6.3.1. Die folgende Tabelle zeigt Messungen zur"Brinellhärte von Kaltarbeitsstählen in Abhängigkeit von der Einsenktiefe."

Einsenktiefe	6	9	11	13	22	26	28	33	35
B.-Härte	68	67	65	53	44	40	37	34	32

Es soll ein 95% KI für b bestimmt werden.

Die Berechnung ergibt $\overline{x} = 20.33$: $\overline{y}= 48.39$; $s_x^2 = 118$; $s_{xy} =$ -155.71; $\widehat{b} = \frac{-155.71}{118.00}$ = -1.32; $\widehat{a} = \overline{y}$ - $\widehat{b}\overline{x}$ = 75.73. Wir erhalten damit die geschätzte Gerade:

$$\widehat{g}(x) = 75.73 - 1.32x$$

Weiter ergibt sich

$$R = \sum_{i=1}^{n} (y_i - \widehat{g}(x_i))^2 = 77.15 \quad \Rightarrow \frac{\sqrt{R}}{s_x \sqrt{8 * 7}} = \frac{\sqrt{77.15}}{\sqrt{118}\sqrt{56}} \approx 0.108$$

Mit $q_{1-\frac{\alpha}{2}} = q_{0.975} = 2.365$ (t_7-Quantil) erhalten wir abschließend das KI:

$$[-1.32 - 0.255; -1.32 + 0.255] = [-1.575; -1.065]$$

Neben Konfidenzintervallen ist man auch an so genannten Konfidenzbändern für die gesamte Gerade interessiert. D.h. wir suchen zwei Funktionen

$$\begin{array}{rcl} oKB & - & \text{oberes Konfidenzband} \\ uKB & - & \text{unteres Konfidenzband} \end{array}$$

die die Eigenschaft besitzen, dass die wahre Gerade

$$g(x) := a + bx$$

mit Wahrscheinlichkeit $1-\alpha$ zwischen ihnen liegt. Also

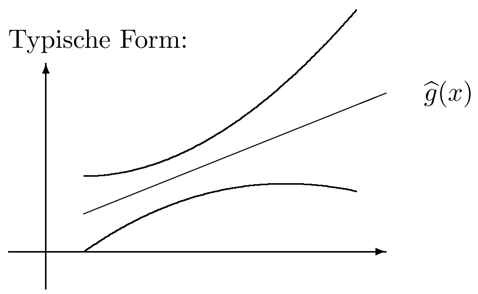
$$\mathbb{P}(\forall\, x \in D_g : uKB(x) \leq g(x) \leq oKB(x)) = 1-\alpha$$

Die beiden Funktionen uKB und oKB bestimmen einen so genannten simultanen Konfidenzstreifen oder auch simultanes Konfidenzband genannt.

Bezeichne $\widehat{g}$(x) $:= \widehat{a} - \widehat{b}x$ die geschätzte Regressionsgerade. Dann wird durch

$$\begin{aligned} uKb(x) &= \widehat{g}(x) - \sqrt{k(x)} \\ oKb(x) &= \widehat{g}(x) + \sqrt{k(x)} \\ k(x) &:= q * 2\,\frac{R}{n-2}\left(\frac{1}{n} + \frac{(x-\overline{x})^2}{(n-1)s_x^2}\right) \end{aligned}$$

wobei $q = q_{1-\alpha}$ - (1-α)Quantil der $F_{2,n-2}$-Verteilung darstellt, solch ein Konfidenzstreifen bestimmt. Üblicherweise haben Konfidenzstreifen die in der Abbildung dargestellte Form.



Im Folgenden interessieren wir uns wieder für zwei Funktionen

$$\begin{aligned} uPB(\mathrm{x}) &\quad - \quad \text{oberes Prognoseband} \\ oPB(\mathrm{x}) &\quad - \quad \text{unteres Prognoseband} \end{aligned}$$

mit der Eigenschaft, dass eine bel. Realisierung von Y, wenn man den Wert $X = x$ (Input) kennt, mit Wahrscheinlichkeit $1-\alpha$ zwischen $uPB(x)$ und $oPB(x)$ liegt, d.h. für alle $x \in D_g$ gilt:

$$\mathbb{P}(uPB(x) \leq \underbrace{g(x)+\varepsilon}_{Y} \leq oPB(x)) = 1-\alpha$$

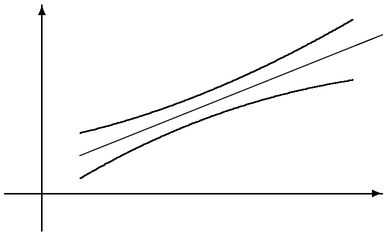
Die beiden Funktionen uPB und oPB bestimmen einen so genannten Prognosestreifen oder auch Prognoseband genannt. Es handelt sich hierbei nicht um ein simultanes Band!

Bezeichne $\widehat{g}(x) := \widehat{a} + \widehat{b}x$ wieder die geschätzte Regressionsgerade. Dann erhalten wir durch

$$\begin{aligned} uPB(x) &= \widehat{g}(x) - p(x) \\ oPB(x) &= \widehat{g}(x) + p(x) \\ p(x) &:= q\sqrt{\frac{R}{n-2}}\sqrt{1 + \frac{1}{n} + \frac{(x-\overline{x})^2}{(n-1)s_x^2}} \end{aligned}$$

wobei $q = q_{1-\frac{\alpha}{2}} - \left(1-\frac{\alpha}{2}\right)$ Quantil der t_{n-2}-Verteilung darstellt, solch einen Prognosestreifen.

Typische Form:



Kapitel 7

Anhang: Tabellen

7.1 Normalverteilung

Die folgende Tabelle zeigt die Verteilungsfunktion der Standard-Normalverteilung $\mathcal{N}(0,1)$.

$F(t)$	t	$F(t)$	t	$F(t)$	t	$F(t)$	t
0,9999	3,719	0,9975	2,807	0,965	1,812	0,83	0,954
0,9998	3,540	0,9970	2,748	0,960	1,751	0,82	0,915
0,9997	3,432	0,9965	2,697	0,955	1,696	0,81	0,878
0,9996	3,353	0,9960	2,652	0,950	1,645	0,80	0,841
0,9995	3,291	0,9955	2,612	0,945	1,599	0,79	0,806
0,9994	3,239	0,9950	2,576	0,940	1,555	0,78	0,772
0,9993	3,195	0,9945	2,543	0,935	1,514	0,76	0,706
0,9992	3,156	0,9940	2,513	0,930	1,476	0,74	0,643
0,9991	3,122	0,9935	2,484	0,925	1,440	0,72	0,582
0,9990	3,091	0,9930	2,458	0,920	1,405	0,70	0,524
0,9989	3,062	0,9925	2,433	0,915	1,372	0,68	0,467
0,9988	3,036	0,9920	2,409	0,910	1,341	0,66	0,412
0,9987	3,012	0,9915	2,387	0,905	1,311	0,64	0,358
0,9986	2,989	0,9910	2,366	0,900	1,282	0,62	0,305
0,9985	2,968	0,9905	2,346	0,890	1,227	0,60	0,253
0,9984	2,948	0,9900	2,327	0,880	1,175	0,58	0,202
0,9983	2,929	0,9850	2,171	0,870	1,126	0,56	0,151
0,9982	2,912	0,9800	2,054	0,860	1,080	0,54	0,100
0,9981	2,895	0,9750	1,960	0,850	1,036	0,52	0,050
0,9980	2,879	0,9700	1,881	0,840	0,994	0,50	0,000

Tabelle 7.1: Verteilungsfunktion und Quantile der Standard-Normalverteilung $\mathcal{N}(0,1)$.

7.2 Student's t-Verteilung

Die folgende Tabelle zeigt Teile der t-Verteilungsfunktion für einige ausgewählte Freiheitsgrade.

dgf=	$F(t) =$			
n	0,99	0,975	0,95	0,9
1	31,821	12,706	6,314	3,078
2	6,965	4,303	2,920	1,886
3	4,541	3,182	2,353	1,638
4	3,747	2,776	2,132	1,533
5	3,365	2,571	2,015	1,476
6	3,143	2,447	1,943	1,440
7	2,998	2,365	1,895	1,415
8	2,896	2,306	1,860	1,397
9	2,821	2,262	1,833	1,383
10	2,764	2,228	1,812	1,372
11	2,718	2,201	1,796	1,363
12	2,681	2,179	1,782	1,356
13	2,650	2,160	1,771	1,350
14	2,624	2,145	1,761	1,345
15	2,602	2,131	1,753	1,341
16	2,583	2,120	1,746	1,337
17	2,567	2,110	1,740	1,333
18	2,552	2,101	1,734	1,330
19	2,539	2,093	1,729	1,328
20	2,528	2,086	1,725	1,325

Tabelle 7.2: Teile der t-Verteilungsfunktion für einige ausgewählte Freiheitsgrade. (dgf).

7.3 Chi-Square Verteilung

Die folgende Tabelle zeigt Teile der χ^2-Verteilungsfunktion für einige ausgewählte Freiheitsgrade.

Tabelle 7.3: Teile der χ^2_n-Verteilungsfunktion für einige ausgewählte Freiheitsgrade n (dgf).

dgf=	$F(t) =$									
n	0,995	0,990	0,975	0,950	0,900	0,100	0,050	0,025	0,010	0,005
1	7,879	6,635	5,024	3,841	2,706	0,016	0,004	0,001	0,000	0,000
2	10,597	9,210	7,378	5,991	4,605	0,211	0,103	0,051	0,020	0,010
3	12,838	11,345	9,348	7,815	6,251	0,584	0,352	0,216	0,115	0,072
4	14,860	13,277	11,143	9,488	7,779	1,064	0,711	0,484	0,297	0,207
5	16,750	15,086	12,833	11,07	9,236	1,610	1,145	0,831	0,554	0,412
6	18,548	16,812	14,449	12,592	10,645	2,204	1,635	1,237	0,872	0,676
7	20,278	18,475	16,013	14,067	12,017	2,833	2,167	1,690	1,239	0,989
8	21,955	20,090	17,535	15,507	13,362	3,490	2,733	2,180	1,646	1,344
9	23,589	21,666	19,023	16,919	14,684	4,168	3,325	2,700	2,088	1,735
10	25,188	23,209	20,483	18,307	15,987	4,865	3,940	3,247	2,558	2,156
11	26,757	24,725	21,920	19,675	17,275	5,578	4,575	3,816	3,053	2,603
12	28,300	26,217	23,337	21,026	18,549	6,304	5,226	4,404	3,571	3,074
13	29,819	27,688	24,736	22,362	19,812	7,042	5,892	5,009	4,107	3,565
14	31,319	29,141	26,119	23,685	21,064	7,790	6,571	5,629	4,660	4,075
15	32,801	30,578	27,488	24,996	22,307	8,547	7,261	6,262	5,229	4,601
16	34,267	32,000	28,845	26,296	23,542	9,312	7,962	6,908	5,812	5,142
17	35,718	33,409	30,191	27,587	24,769	10,085	8,672	7,564	6,408	5,697
18	37,156	34,805	31,526	28,869	25,989	10,865	9,390	8,231	7,015	6,265
19	38,582	36,191	32,852	30,144	27,204	11,651	10,117	8,907	7,633	6,844
20	39,997	37,566	34,170	31,410	28,412	12,443	10,851	9,591	8,260	7,434
21	41,401	38,932	35,479	32,671	29,615	13,240	11,591	10,283	8,897	8,034
22	42,796	40,289	36,781	33,924	30,813	14,041	12,338	10,982	9,542	8,643
23	44,181	41,638	38,076	35,172	32,007	14,848	13,091	11,689	10,196	9,260
24	45,559	42,980	39,364	36,415	33,196	15,659	13,848	12,401	10,856	9,886
25	46,928	44,314	40,646	37,652	34,382	16,473	14,611	13,120	11,524	10,520
26	48,290	45,642	41,923	38,885	35,563	17,292	15,379	13,844	12,198	11,160
27	49,645	46,963	43,195	40,113	36,741	18,114	16,151	14,573	12,879	11,808
28	50,993	48,278	44,461	41,337	37,916	18,939	16,928	15,308	13,565	12,461
29	52,336	49,588	45,722	42,557	39,087	19,768	17,708	16,047	14,256	13,121
30	53,672	50,892	46,979	43,773	40,256	20,599	18,493	16,791	14,953	13,787
31	55,003	52,191	48,232	44,985	41,422	21,434	19,281	17,539	15,655	14,458
32	56,328	53,486	49,480	46,194	42,585	22,271	20,072	18,291	16,362	15,134
33	57,648	54,776	50,725	47,400	43,745	23,110	20,867	19,047	17,074	15,815
34	58,964	56,061	51,966	48,602	44,903	23,952	21,664	19,806	17,789	16,501
35	60,275	57,342	53,203	49,802	46,059	24,797	22,465	20,569	18,509	17,192
36	61,581	58,619	54,437	50,998	47,212	25,643	23,269	21,336	19,233	17,887
37	62,883	59,893	55,668	52,192	48,363	26,492	24,075	22,106	19,960	18,586
38	64,181	61,162	56,896	53,384	49,513	27,343	24,884	22,878	20,691	19,289
39	65,476	62,428	58,120	54,572	50,660	28,196	25,695	23,654	21,426	19,996
40	66,766	63,691	59,342	55,758	51,805	29,051	26,509	24,433	22,164	20,707
41	68,053	64,950	60,561	56,942	52,949	29,907	27,326	25,215	22,906	21,421
42	69,336	66,206	61,777	58,124	54,090	30,765	28,144	25,999	23,650	22,138
43	70,616	67,459	62,990	59,304	55,230	31,625	28,965	26,785	24,398	22,859

dgf=	$F(t) =$									
n	0,995	0,990	0,975	0,950	0,900	0,100	0,050	0,025	0,010	0,005
44	71,893	68,710	64,201	60,481	56,369	32,487	29,787	27,575	25,148	23,584
45	73,166	69,957	65,410	61,656	57,505	33,350	30,612	28,366	25,901	24,311
46	74,437	71,201	66,617	62,830	58,641	34,215	31,439	29,160	26,657	25,041
47	75,704	72,443	67,821	64,001	59,774	35,081	32,268	29,956	27,416	25,775
48	76,969	73,683	69,023	65,171	60,907	35,949	33,098	30,755	28,177	26,511
49	78,231	74,919	70,222	66,339	62,038	36,818	33,930	31,555	28,941	27,249
50	79,490	76,154	71,420	67,505	63,167	37,689	34,764	32,357	29,707	27,991

7.4 F-Verteilung

Die folgende Tabelle zeigt Teile der F-Verteilungsfunktion für einige ausgewählte Freiheitsgrade. Neben den gelisteten Werten der Quantile kann man durch

$$F^{-1}_{n_1,n_2}(1-\gamma) = \frac{1}{F^{-1}_{n_2,n_1}(\gamma)}$$

die Tabelle erweitern.

Tabelle 7.4: Teile der F_{n_1,n_2}-Verteilungsfunktion für einige ausgewählte Freiheitsgrade (n_1, n_2).

		n_1									
n_2	$F_{n_1,n_2}(t) =$	1	2	3	4	5	6	7	8	9	10
1	0,99	4052,18	4999,5	5403,35	5624,58	5763,65	5858,99	5928,36	5981,07	6022,47	6055,85
1	0,975	647,79	799,5	864,16	899,58	921,85	937,11	948,22	956,66	963,28	968,63
1	0,95	161,45	199,5	215,71	224,58	230,16	233,99	236,77	238,88	240,54	241,88
1	0,9	39,86	49,5	53,59	55,83	57,24	58,2	58,91	59,44	59,86	60,19
2	0,99	98,5	99	99,17	99,25	99,3	99,33	99,36	99,37	99,39	99,4
2	0,975	38,51	39	39,17	39,25	39,3	39,33	39,36	39,37	39,39	39,4
2	0,95	18,51	19	19,16	19,25	19,3	19,33	19,35	19,37	19,38	19,4
2	0,9	8,53	9	9,16	9,24	9,29	9,33	9,35	9,37	9,38	9,39
3	0,99	34,12	30,82	29,46	28,71	28,24	27,91	27,67	27,49	27,35	27,23
3	0,975	17,44	16,04	15,44	15,1	14,88	14,73	14,62	14,54	14,47	14,42
3	0,95	10,13	9,55	9,28	9,12	9,01	8,94	8,89	8,85	8,81	8,79
3	0,9	5,54	5,46	5,39	5,34	5,31	5,28	5,27	5,25	5,24	5,23
4	0,99	21,2	18	16,69	15,98	15,52	15,21	14,98	14,8	14,66	14,55
4	0,975	12,22	10,65	9,98	9,6	9,36	9,2	9,07	8,98	8,9	8,84
4	0,95	7,71	6,94	6,59	6,39	6,26	6,16	6,09	6,04	6	5,96
4	0,9	4,54	4,32	4,19	4,11	4,05	4,01	3,98	3,95	3,94	3,92
5	0,99	16,26	13,27	12,06	11,39	10,97	10,67	10,46	10,29	10,16	10,05
5	0,975	10,01	8,43	7,76	7,39	7,15	6,98	6,85	6,76	6,68	6,62
5	0,95	6,61	5,79	5,41	5,19	5,05	4,95	4,88	4,82	4,77	4,74
5	0,9	4,06	3,78	3,62	3,52	3,45	3,4	3,37	3,34	3,32	3,3
6	0,99	13,75	10,92	9,78	9,15	8,75	8,47	8,26	8,1	7,98	7,87
6	0,975	8,81	7,26	6,6	6,23	5,99	5,82	5,7	5,6	5,52	5,46
6	0,95	5,99	5,14	4,76	4,53	4,39	4,28	4,21	4,15	4,1	4,06
6	0,9	3,78	3,46	3,29	3,18	3,11	3,05	3,01	2,98	2,96	2,94
7	0,99	12,25	9,55	8,45	7,85	7,46	7,19	6,99	6,84	6,72	6,62
7	0,975	8,07	6,54	5,89	5,52	5,29	5,12	4,99	4,9	4,82	4,76
7	0,95	5,59	4,74	4,35	4,12	3,97	3,87	3,79	3,73	3,68	3,64
7	0,9	3,59	3,26	3,07	2,96	2,88	2,83	2,78	2,75	2,72	2,7
8	0,99	11,26	8,65	7,59	7,01	6,63	6,37	6,18	6,03	5,91	5,81
8	0,975	7,57	6,06	5,42	5,05	4,82	4,65	4,53	4,43	4,36	4,3
8	0,95	5,32	4,46	4,07	3,84	3,69	3,58	3,5	3,44	3,39	3,35
8	0,9	3,46	3,11	2,92	2,81	2,73	2,67	2,62	2,59	2,56	2,54
9	0,99	10,56	8,02	6,99	6,42	6,06	5,8	5,61	5,47	5,35	5,26
9	0,975	7,21	5,71	5,08	4,72	4,48	4,32	4,2	4,1	4,03	3,96
9	0,95	5,12	4,26	3,86	3,63	3,48	3,37	3,29	3,23	3,18	3,14
9	0,9	3,36	3,01	2,81	2,69	2,61	2,55	2,51	2,47	2,44	2,42
10	0,99	10,04	7,56	6,55	5,99	5,64	5,39	5,2	5,06	4,94	4,85
10	0,975	6,94	5,46	4,83	4,47	4,24	4,07	3,95	3,85	3,78	3,72
10	0,95	4,96	4,1	3,71	3,48	3,33	3,22	3,14	3,07	3,02	2,98
10	0,9	3,29	2,92	2,73	2,61	2,52	2,46	2,41	2,38	2,35	2,32
11	0,99	9,65	7,21	6,22	5,67	5,32	5,07	4,89	4,74	4,63	4,54
11	0,975	6,72	5,26	4,63	4,28	4,04	3,88	3,76	3,66	3,59	3,53
11	0,95	4,84	3,98	3,59	3,36	3,2	3,09	3,01	2,95	2,9	2,85
11	0,9	3,23	2,86	2,66	2,54	2,45	2,39	2,34	2,3	2,27	2,25
12	0,99	9,33	6,93	5,95	5,41	5,06	4,82	4,64	4,5	4,39	4,3
12	0,975	6,55	5,1	4,47	4,12	3,89	3,73	3,61	3,51	3,44	3,37
12	0,95	4,75	3,89	3,49	3,26	3,11	3	2,91	2,85	2,8	2,75
12	0,9	3,18	2,81	2,61	2,48	2,39	2,33	2,28	2,24	2,21	2,19
13	0,99	9,07	6,7	5,74	5,21	4,86	4,62	4,44	4,3	4,19	4,1
13	0,975	6,41	4,97	4,35	4	3,77	3,6	3,48	3,39	3,31	3,25
13	0,95	4,67	3,81	3,41	3,18	3,03	2,92	2,83	2,77	2,71	2,67
13	0,9	3,14	2,76	2,56	2,43	2,35	2,28	2,23	2,2	2,16	2,14
14	0,99	8,86	6,51	5,56	5,04	4,69	4,46	4,28	4,14	4,03	3,94
14	0,975	6,3	4,86	4,24	3,89	3,66	3,5	3,38	3,29	3,21	3,15
14	0,95	4,6	3,74	3,34	3,11	2,96	2,85	2,76	2,7	2,65	2,6
14	0,9	3,1	2,73	2,52	2,39	2,31	2,24	2,19	2,15	2,12	2,1
15	0,99	8,68	6,36	5,42	4,89	4,56	4,32	4,14	4	3,89	3,8
15	0,975	6,2	4,77	4,15	3,8	3,58	3,41	3,29	3,2	3,12	3,06
15	0,95	4,54	3,68	3,29	3,06	2,9	2,79	2,71	2,64	2,59	2,54
15	0,9	3,07	2,7	2,49	2,36	2,27	2,21	2,16	2,12	2,09	2,06
16	0,99	8,53	6,23	5,29	4,77	4,44	4,2	4,03	3,89	3,78	3,69
16	0,975	6,12	4,69	4,08	3,73	3,5	3,34	3,22	3,12	3,05	2,99
16	0,95	4,49	3,63	3,24	3,01	2,85	2,74	2,66	2,59	2,54	2,49
16	0,9	3,05	2,67	2,46	2,33	2,24	2,18	2,13	2,09	2,06	2,03
17	0,99	8,4	6,11	5,18	4,67	4,34	4,1	3,93	3,79	3,68	3,59
17	0,975	6,04	4,62	4,01	3,66	3,44	3,28	3,16	3,06	2,98	2,92
17	0,95	4,45	3,59	3,2	2,96	2,81	2,7	2,61	2,55	2,49	2,45
17	0,9	3,03	2,64	2,44	2,31	2,22	2,15	2,1	2,06	2,03	2
18	0,99	8,29	6,01	5,09	4,58	4,25	4,01	3,84	3,71	3,6	3,51
18	0,975	5,98	4,56	3,95	3,61	3,38	3,22	3,1	3,01	2,93	2,87
18	0,95	4,41	3,55	3,16	2,93	2,77	2,66	2,58	2,51	2,46	2,41
18	0,9	3,01	2,62	2,42	2,29	2,2	2,13	2,08	2,04	2	1,98
19	0,99	8,18	5,93	5,01	4,5	4,17	3,94	3,77	3,63	3,52	3,43
19	0,975	5,92	4,51	3,9	3,56	3,33	3,17	3,05	2,96	2,88	2,82
19	0,95	4,38	3,52	3,13	2,9	2,74	2,63	2,54	2,48	2,42	2,38
19	0,9	2,99	2,61	2,4	2,27	2,18	2,11	2,06	2,02	1,98	1,96
20	0,99	8,1	5,85	4,94	4,43	4,1	3,87	3,7	3,56	3,46	3,37
20	0,975	5,87	4,46	3,86	3,51	3,29	3,13	3,01	2,91	2,84	2,77
20	0,95	4,35	3,49	3,1	2,87	2,71	2,6	2,51	2,45	2,39	2,35
20	0,9	2,97	2,59	2,38	2,25	2,16	2,09	2,04	2	1,96	1,94

Tabelle 7.5: Teile der F_{n_1,n_2}-Verteilungsfunktion für einige ausgewählte Freiheitsgrade (n_1, n_2).

		n_1									
n_2	$F_{n_1,n_2}(t) =$	11	12	13	14	15	16	17	18	19	20
1	0,99	6083,32	6106,32	6125,86	6142,67	6157,28	6170,1	6181,43	6191,53	6200,58	6208,73
1	0,975	973,03	976,71	979,84	982,53	984,87	986,92	988,73	990,35	991,8	993,1
1	0,95	242,98	243,91	244,69	245,36	245,95	246,46	246,92	247,32	247,69	248,01
1	0,9	60,47	60,71	60,9	61,07	61,22	61,35	61,46	61,57	61,66	61,74
2	0,99	99,41	99,42	99,42	99,43	99,43	99,44	99,44	99,44	99,45	99,45
2	0,975	39,41	39,41	39,42	39,43	39,43	39,44	39,44	39,44	39,45	39,45
2	0,95	19,4	19,41	19,42	19,42	19,43	19,43	19,44	19,44	19,44	19,45
2	0,9	9,4	9,41	9,41	9,42	9,42	9,43	9,43	9,44	9,44	9,44
3	0,99	27,13	27,05	26,98	26,92	26,87	26,83	26,79	26,75	26,72	26,69
3	0,975	14,37	14,34	14,3	14,28	14,25	14,23	14,21	14,2	14,18	14,17
3	0,95	8,76	8,74	8,73	8,71	8,7	8,69	8,68	8,67	8,67	8,66
3	0,9	5,22	5,22	5,21	5,2	5,2	5,2	5,19	5,19	5,19	5,18
4	0,99	14,45	14,37	14,31	14,25	14,2	14,15	14,11	14,08	14,05	14,02
4	0,975	8,79	8,75	8,71	8,68	8,66	8,63	8,61	8,59	8,58	8,56
4	0,95	5,94	5,91	5,89	5,87	5,86	5,84	5,83	5,82	5,81	5,8
4	0,9	3,91	3,9	3,89	3,88	3,87	3,86	3,86	3,85	3,85	3,84
5	0,99	9,96	9,89	9,82	9,77	9,72	9,68	9,64	9,61	9,58	9,55
5	0,975	6,57	6,52	6,49	6,46	6,43	6,4	6,38	6,36	6,34	6,33
5	0,95	4,7	4,68	4,66	4,64	4,62	4,6	4,59	4,58	4,57	4,56
5	0,9	3,28	3,27	3,26	3,25	3,24	3,23	3,22	3,22	3,21	3,21
6	0,99	7,79	7,72	7,66	7,6	7,56	7,52	7,48	7,45	7,42	7,4
6	0,975	5,41	5,37	5,33	5,3	5,27	5,24	5,22	5,2	5,18	5,17
6	0,95	4,03	4	3,98	3,96	3,94	3,92	3,91	3,9	3,88	3,87
6	0,9	2,92	2,9	2,89	2,88	2,87	2,86	2,85	2,85	2,84	2,84
7	0,99	6,54	6,47	6,41	6,36	6,31	6,28	6,24	6,21	6,18	6,16
7	0,975	4,71	4,67	4,63	4,6	4,57	4,54	4,52	4,5	4,48	4,47
7	0,95	3,6	3,57	3,55	3,53	3,51	3,49	3,48	3,47	3,46	3,44
7	0,9	2,68	2,67	2,65	2,64	2,63	2,62	2,61	2,61	2,6	2,59
8	0,99	5,73	5,67	5,61	5,56	5,52	5,48	5,44	5,41	5,38	5,36
8	0,975	4,24	4,2	4,16	4,13	4,1	4,08	4,05	4,03	4,02	4
8	0,95	3,31	3,28	3,26	3,24	3,22	3,2	3,19	3,17	3,16	3,15
8	0,9	2,52	2,5	2,49	2,48	2,46	2,45	2,45	2,44	2,43	2,42
9	0,99	5,18	5,11	5,05	5,01	4,96	4,92	4,89	4,86	4,83	4,81
9	0,975	3,91	3,87	3,83	3,8	3,77	3,74	3,72	3,7	3,68	3,67
9	0,95	3,1	3,07	3,05	3,03	3,01	2,99	2,97	2,96	2,95	2,94
9	0,9	2,4	2,38	2,36	2,35	2,34	2,33	2,32	2,31	2,3	2,3
10	0,99	4,77	4,71	4,65	4,6	4,56	4,52	4,49	4,46	4,43	4,41
10	0,975	3,66	3,62	3,58	3,55	3,52	3,5	3,47	3,45	3,44	3,42
10	0,95	2,94	2,91	2,89	2,86	2,85	2,83	2,81	2,8	2,79	2,77
10	0,9	2,3	2,28	2,27	2,26	2,24	2,23	2,22	2,22	2,21	2,2
11	0,99	4,46	4,4	4,34	4,29	4,25	4,21	4,18	4,15	4,12	4,1
11	0,975	3,47	3,43	3,39	3,36	3,33	3,3	3,28	3,26	3,24	3,23
11	0,95	2,82	2,79	2,76	2,74	2,72	2,7	2,69	2,67	2,66	2,65
11	0,9	2,23	2,21	2,19	2,18	2,17	2,16	2,15	2,14	2,13	2,12
12	0,99	4,22	4,16	4,1	4,05	4,01	3,97	3,94	3,91	3,88	3,86
12	0,975	3,32	3,28	3,24	3,21	3,18	3,15	3,13	3,11	3,09	3,07
12	0,95	2,72	2,69	2,66	2,64	2,62	2,6	2,58	2,57	2,56	2,54
12	0,9	2,17	2,15	2,13	2,12	2,1	2,09	2,08	2,08	2,07	2,06
13	0,99	4,02	3,96	3,91	3,86	3,82	3,78	3,75	3,72	3,69	3,67
13	0,975	3,2	3,15	3,12	3,08	3,05	3,03	3	2,98	2,96	2,95
13	0,95	2,63	2,6	2,58	2,55	2,53	2,51	2,5	2,48	2,47	2,46
13	0,9	2,12	2,1	2,08	2,07	2,05	2,04	2,03	2,02	2,01	2,01
14	0,99	3,86	3,8	3,75	3,7	3,66	3,62	3,59	3,56	3,53	3,51
14	0,975	3,09	3,05	3,01	2,98	2,95	2,92	2,9	2,88	2,86	2,84
14	0,95	2,57	2,53	2,51	2,48	2,46	2,44	2,43	2,41	2,4	2,39
14	0,9	2,07	2,05	2,04	2,02	2,01	2	1,99	1,98	1,97	1,96
15	0,99	3,73	3,67	3,61	3,56	3,52	3,49	3,45	3,42	3,4	3,37
15	0,975	3,01	2,96	2,92	2,89	2,86	2,84	2,81	2,79	2,77	2,76
15	0,95	2,51	2,48	2,45	2,42	2,4	2,38	2,37	2,35	2,34	2,33
15	0,9	2,04	2,02	2	1,99	1,97	1,96	1,95	1,94	1,93	1,92
16	0,99	3,62	3,55	3,5	3,45	3,41	3,37	3,34	3,31	3,28	3,26
16	0,975	2,93	2,89	2,85	2,82	2,79	2,76	2,74	2,72	2,7	2,68
16	0,95	2,46	2,42	2,4	2,37	2,35	2,33	2,32	2,3	2,29	2,28
16	0,9	2,01	1,99	1,97	1,95	1,94	1,93	1,92	1,91	1,9	1,89
17	0,99	3,52	3,46	3,4	3,35	3,31	3,27	3,24	3,21	3,19	3,16
17	0,975	2,87	2,82	2,79	2,75	2,72	2,7	2,67	2,65	2,63	2,62
17	0,95	2,41	2,38	2,35	2,33	2,31	2,29	2,27	2,26	2,24	2,23
17	0,9	1,98	1,96	1,94	1,93	1,91	1,9	1,89	1,88	1,87	1,86
18	0,99	3,43	3,37	3,32	3,27	3,23	3,19	3,16	3,13	3,1	3,08
18	0,975	2,81	2,77	2,73	2,7	2,67	2,64	2,62	2,6	2,58	2,56
18	0,95	2,37	2,34	2,31	2,29	2,27	2,25	2,23	2,22	2,2	2,19
18	0,9	1,95	1,93	1,92	1,9	1,89	1,87	1,86	1,85	1,84	1,84
19	0,99	3,36	3,3	3,24	3,19	3,15	3,12	3,08	3,05	3,03	3
19	0,975	2,76	2,72	2,68	2,65	2,62	2,59	2,57	2,55	2,53	2,51
19	0,95	2,34	2,31	2,28	2,26	2,23	2,21	2,2	2,18	2,17	2,16
19	0,9	1,93	1,91	1,89	1,88	1,86	1,85	1,84	1,83	1,82	1,81
20	0,99	3,29	3,23	3,18	3,13	3,09	3,05	3,02	2,99	2,96	2,94
20	0,975	2,72	2,68	2,64	2,6	2,57	2,55	2,52	2,5	2,48	2,46
20	0,95	2,31	2,28	2,25	2,22	2,2	2,18	2,17	2,15	2,14	2,12
20	0,9	1,91	1,89	1,87	1,86	1,84	1,83	1,82	1,81	1,8	1,79

Literaturverzeichnis

[1] Hartung J.
Statistik, Lehr- und Handbuch der angewandten Statistik
Oldenbourg Verlag

[2] Krickeberg K. Ziezold H.
Stochastische Methoden
Springer Verlag

[3] Weber H.
Einführung in die Wahrscheinlichkeitsrechnung und Statistik für Ingenieure
Teubner Stuttgart

[4] Graf, Henning, Stange, Wilrich
Formeln und Tabellen der angewandten mathematischen Statistik
Springer Verlag

Index