ulm university universität uulm

Statistik I

Vorlesungsskript

Prof. Dr. Evgeny Spodarev

Ulm
2008

Vorwort

Dieses Skript entstand aus dem Zyklus der Vorlesungen über Statistik, die ich in den Jahren 2005–2008 an der Universität Ulm gehalten habe. Dabei handelt es sich um die erste Einführung in die Statistik, die durch die aufbauende Vorlesung Statistik II ergänzt wird.

Dieses Skript gibt eine Übersicht über die typischen Fragestellungen und Methoden der mathematischen Statistik. Es stellt einen Versuch dar, einen Mittelweg zwischen praktisch orientierten (aber mathematisch oft sehr dürftigen) Statistik-Monographien einerseits und trockenen Büchern über die mathematische Statistik andererseits einzuschlagen. Ob es mir gelungen ist, soll der Leser beurteilen.

Ich möchte gerne meinen Kollegen aus dem Institut für Stochastik, Herrn Prof. Volker Schmidt und Herrn Dipl.-Math. Malte Spiess, für ihre Unterstützung und anregenden Diskussionen während der Entstehung des Skriptes danken. Herr Tobias Brosch hat eine hervorragende Arbeit beim Tippen des Skriptes und bei der Erstellung zahlreicher Abbildungen, die den Text begleiten, geleistet. Dafür gilt ihm mein herzlicher Dank.

Ulm, den 19.08.2008 Evgeny Spodarev

Inhaltsverzeichnis

1 Einführung **1**
1.1 Typische Fragestellungen, Aufgaben und Ziele der Statistik . . . 1
1.2 Statistische Merkmale und ihre Typen . . . 4
1.3 Statistische Daten und Stichproben . . . 5
1.4 Stichprobenfunktionen . . . 6

2 Beschreibende Statistik **7**
2.1 Verteilungen und ihre Darstellungen . . . 7
2.1.1 Häufigkeiten und Diagramme . . . 7
2.1.2 Empirische Verteilungsfunktion . . . 10
2.2 Beschreibung von Verteilungen . . . 12
2.2.1 Lagemaße . . . 12
2.2.2 Streuungsmaße . . . 15
2.2.3 Konzentrationsmaße . . . 17
2.2.4 Maße für Schiefe und Wölbung . . . 20
2.3 Quantilplots (Quantil-Grafiken) . . . 21
2.4 Dichteschätzung . . . 25
2.5 Beschreibung und Exploration von bivariaten Datensätzen . . . 27
2.5.1 Grafische Darstellung von bivariaten Datensätzen . . . 27
2.5.2 Zusammenhangsmaße . . . 29
2.5.3 Einfache lineare Regression . . . 33

3 Punktschätzer **40**
3.1 Parametrisches Modell . . . 40
3.2 Parametrische Familien von statistischen Prüfverteilungen . . . 41
3.2.1 Gamma-Verteilung . . . 41
3.2.2 Student-Verteilung (t-Verteilung) . . . 44
3.2.3 Fisher-Snedecor-Verteilung (F-Verteilung) . . . 47
3.3 Punktschätzer und ihre Grundeigenschaften . . . 49
3.3.1 Eigenschaften von Punktschätzern . . . 49
3.3.2 Schätzer des Erwartungswertes und empirische Momente . . . 51
3.3.3 Schätzer der Varianz . . . 53
3.3.4 Eigenschaften der Ordnungsstatistiken . . . 61
3.3.5 Empirische Verteilungsfunktion . . . 63
3.4 Methoden zur Gewinnung von Punktschätzern . . . 71
3.4.1 Momentenschätzer . . . 71
3.4.2 Maximum-Likelihood-Schätzer . . . 73
3.4.3 Bayes-Schätzer . . . 82
3.4.4 Resampling-Methoden zur Gewinnung von Punktschätzern . . . 84

3.5 Weitere Güteeigenschaften von Punktschätzern 88
3.5.1 Ungleichung von Cramér-Rao 88
3.5.2 Bedingte Erwartung 93
3.5.3 Suffizienz 96
3.5.4 Vollständigkeit 100
3.5.5 Bester erwartungstreuer Schätzer 102

Literatur **105**

Index **107**

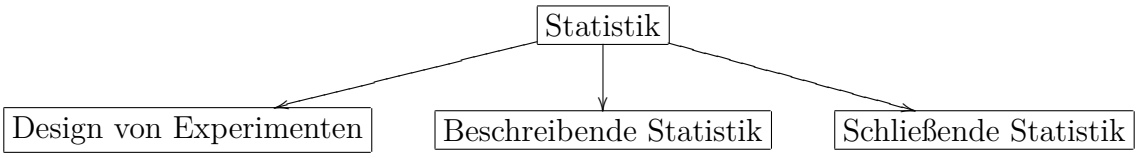
1 Einführung

1.1 Typische Fragestellungen, Aufgaben und Ziele der Statistik

Im alltäglichen Sprachgebrauch versteht man unter „Statistik“ eine Darstellung von Ergebnissen des Zusammenzählens von Daten und Fakten jeglicher Art, wie z.B. ökonomischen Kenngrößen, politischen Umfragen, Daten der Marktforschung, klinischen Studien in der Biologie und Medizin, usw.

Die *mathematische Statistik* jedoch kann viel mehr. Sie arbeitet mit *Daten-Stichproben,* die nach einem bestimmten Zufallsmechanismus aus der *Grundgesamtheit* aller Daten, die in Folge von Beobachtung, Experimenten (reale Daten) oder Computersimulation (synthetische Daten) erhoben wurden. Dabei beschäftigt sich die mathematische Statistik mit folgenden Fragestellungen:

1. Wie sollen die Daten gewonnen werden? (Design von Experimenten)
2. Wie sollen (insbesondere riesengroße) Datensätze beschrieben werden, um die Gesetzmäßigkeiten und Strukturen in ihnen entecken zu können? (Beschreibende (deskriptive) und explorative Statistik)
3. Welche Schlüsse kann man aus den Daten ziehen? (Schließende oder induktive Statistik)



In dieser einführenden Vorlesung werden wir Teile der beschreibenden und schließenden Statistik kennelernen, wobei die Datenerhebung aus Platzgründen ausgelassen wird. Die *Arbeitsweise eines Statistikers* sieht folgendermaßen aus:

1. *Datenerhebung*
2. *Visualisierung und beschreibende Datenanalyse*
3. *Datenbereinigung* (z.B. Erkennung fehlerhafter Messungen, Ausreißern, usw.)
4. *Explorative Datenanalyse* (Suche nach Gesetzmäßigkeiten)
5. *Modellierung der Daten* mit Methoden der Stochastik
6. *Modellanpassung* (Schätzung der Modellparameter)
7. *Modellvalidierung* (wie gut war die Modellanpassung?)

Pflanze	1	2	3	4	5	6	7	8	9	10
rund	45	27	24	19	32	26	88	22	28	25
kantig	12	8	7	10	11	6	24	10	6	7
Verhältnis . . . : 1	3,8	3,4	3,4	1,9	2,9	4,3	3,7	2,2	4,7	3,6

Tab. 1.1: Ergebnisse für die 10 Pflanzen des ersten Versuchs von Mendel

8. *Schließende Datenanalyse:*
   - Konstruktion von *Vertrauensintervallen* (Konfidenzintervallen) für Modellparameter und deren Funktionen,
   - Tests statistischer Hypothesen,
   - Vorhersage von Zielgrößen (z.B. auf Basis modellbezogener Computersimulation).

Uns werden in diesem Vorlesungsskript vor allem die Arbeitspunkte 2), 4)–6) und 8) beschäftigen.

Beispiel 1.1.1
Nachfolgend geben wir einige typische Fragestellungen der Statistik an Beispielen von Datensätzen:

1. *Statistische Herleitung von Grundsätzen der biologischen Evolution (Mendel, 1865):*

   Es wurden Nachkommen von zwei Erbsensorten, die sich in der Samenform unterscheiden, gezüchtet: die erste Sorte hat runde, die zweite kantige Erbsen. Johann Gregor Mendel hat festgestellt, dass sich runde Samen dominant vererben. Dabei werden bei einer Bestäubung von Pflanzen der einen Sorte mit Pollen der anderen alle Nachkommen runde Samen zeigen, die genetisch bevorzugt sind, d.h., beide Merkmale aufweisen. Kreuzt man diese hybriden Pflanzen, so zeigen sie runde und kantige Samen im Verhältnis $3:1$ (Spaltungs- und Dominanzregeln von Mendel). Bei der statistischen Überprüfung seiner Vermutungen erhielt Mendel 5475 runde und 1850 kantige Samen, die somit im Verhältnis $2,96:1$ stehen. In der Tabelle 1.1 sind Ergebnisse für die ersten 10 Pflanzen gezeigt. Man sieht, dass das oben genannte Verhältnis zufällig um $3:1$ schwankt. Durch die Bildung des Mittels über das Gesamtkollektiv der Daten wird die Gesetzmäßigkeit $3:1$ gefunden (explorative Statistik).

2. *Kreditwürdigkeit bei Kreditvergabe*

   Die Banken sind offensichtlich daran interessiert, Bankkredite an Kunden zu vergeben, die in der Zukunft solvent bleiben, also die Kreditraten regelmäßig zurückzahlen können. Um die Kreditwürdigkeit zu überprfen, werden Umfragen gemacht, wobei die Antworten unter anderem in folgenden Variablen kodiert werden:
   - X_1 Laufendes Konto bei der Bank ($1 =$ nein, $2 =$ ja und durchschnittlich geführt, $3 =$ ja und gut geführt)
   - X_2 Laufzeit des Kredits in Monaten
   - X_3 Kredithöhe in €
   - X_4 Rückzahlung früherer Kredite (gut/ schlecht)

	Y	
X_1: laufendes Konto	1	0
nein	$45,0$	$19,9$
gut	$15,3$	$49,7$
mittel	$39,7$	$30,2$
X_3: Kredithöhe in €	1	0
$0 < \ldots \leq 500$	$1,00$	$2,14$
$500 < \ldots \leq 1000$	$11,33$	$9,14$
$1000 < \ldots \leq 1500$	$17,00$	$19,86$
$1500 < \ldots \leq 2500$	$19,67$	$24,57$
$2500 < \ldots \leq 5000$	$25,00$	$28,57$
$5000 < \ldots \leq 7500$	$11,33$	$9,71$
$7500 < \ldots \leq 10000$	$6,67$	$3,71$
$10000 < \ldots \leq 15000$	$7,00$	$2,00$
$15000 < \ldots \leq 20000$	$1,00$	$0,29$
X_4: Frühere Kredite	1	0
gut	$82,33$	$94,85$
schlecht	$17,66$	$5,15$
X_5: Verwendungszweck	1	0
privat	$57,53$	$69,29$
beruflich	$42,47$	$30,71$

Tab. 1.2: Lernstichprobe zur Vergabe von Krediten

- X_5 Verwendungszweck (privat / geschäftlich)
- X_6 Geschlecht (weiblich / männlich)

Um an Hand eines ausgefüllten Fragebogens wie diesem eine Entscheidung über die Vergabe des Kredits treffen zu können, werden *Lernstichproben* herangezogen, bei denen das Ergebnis Y der erfolgten Kreditvergabe bekannt ist. Dabei bedeutet $Y = 0$ gut und $Y = 1$ schlecht. Betrachten wir eine solche Stichprobe einer süddeutschen Bank, die 1000 Umfragebögen umfasst. Dabei sind 700 kreditwürdig und 300 davon nicht kreditwürdig gewesen. Die Tabelle 1.2 zeigt Prozentzahlen dieses Datensatzes für ausgewählte Merkmale X_i. Dabei ist es möglich, mit Hilfe statistischer Methoden (Regression) eine Kreditentscheidung bei einem Kunden an Hand dieser Lernprobe automatisch treffen zu können. Dieser Vorgang wird manchmal auch „statistisches Lernen“ genannt. Fragestellungen wie diese werden erst in Statistik II (verallgemeinerte lineare Modelle) behandelt.

3. *Korrosion von Legierungen*

   In diesem Beispiel wurde der Korrosionsgrad einer Kupfer-Nickel-Legierung in Abhängigkeit ihres Eisengehalts untersucht. Dazu wurden 13 verschiedene Räder mit dieser Legierung beschichtet und 60 Tage lang in Meerwasser gedreht. Danach wurde der Gewichtsverlust in mg pro dm^2 und Tag bestimmt. Aus dem Bild 1.1 ist zu sehen, dass die Korrosion in Abhängigkeit vom Eisengehalt linear abnimmt. Mit statistischen Methoden

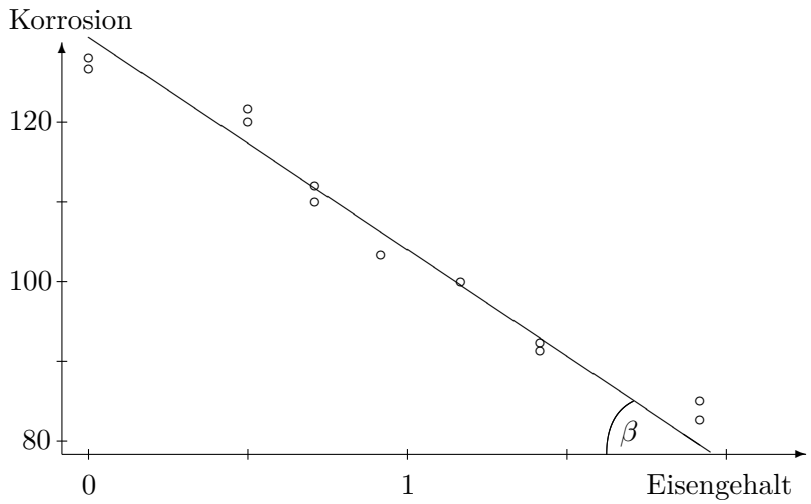
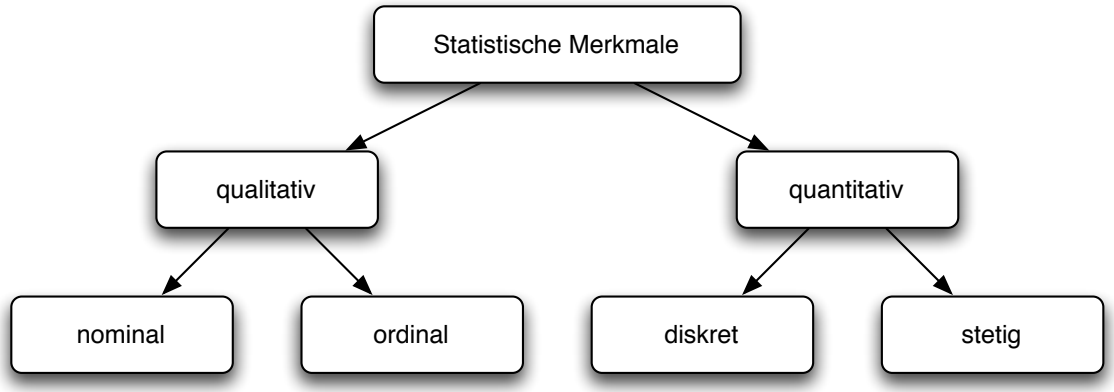


Abb. 1.1: Korrosion von Kupfer-Nickel-Legierung

(einfache lineare Regression) kann die Geschwindigkeit dieser Abnahme geschätzt werden.

1.2 Statistische Merkmale und ihre Typen

Die Daten, die zur statistischen Analyse vorliegen, können eine oder mehrere interessierende Größen (die auch *Variablen* oder *Merkmale* genannt werden) umfassen. Ihre Werte werden *Merkmalsausprägungen* genannt. In dem nachfolgenden Diagramm werden mögliche Typen der statistischen Merkmale gegeben.



Diese Typen entstehen in Folge der Klassifikation von Wertebereichen (Skalen) der Merkmale. Dennoch ist diese Einteilung nicht vollständig und kann bei Bedarf erweitert werden. Man unterscheidet *qualitative* und *quantitative* Merkmale. *Quantitative Merkmale* lassen sich inhaltlich gut durch Zahlen darstellen (z.B. Kredithöhe in €, Körpergewicht und Körpergröße, Blutdruck usw.). Sie können *diskrete* oder *stetige* Wertebereiche haben, wobei diskrete Merkmale isolierte Werte annehmen können (z.B. Anzahl der Schäden eines Versicherers pro Jahr). Stetige Wertebereiche hingegen sind überabzählbar. Dennoch liegen in der Praxis stetige Merkmale in gerundeter Form vor (z.B. Körpergröße auf cm gerundet, Geldbeträge auf € gerundet usw.).

Im Gegensatz zu den quantitativen Merkmalen sind die Inhalte der *qualitativen Merkmale*, wie z.B. Blutgruppe (0, A, B und AB) oder Familienstand (ledig, verheiratet, verwitwet), nicht sinnvoll durch Zahlen darzustellen. Sie können zwar formell mit Zahlen kodiert werden (z.B. bei Blutgruppen $0 = 0$, $A = 1$, $B = 2$, $AB = 3$), aber solche Kodierungen stellen keinen inhaltlichen Zusammenhang zwischen Ausprägungen und Zahlen-Codes dar sondern dienen lediglich der besseren Identifikation der Merkmale auf einem Rechner. Es ist insbesondere unsinnig, Mittelwerte und ähnliches von solchen Codes zu bilden.

Ein qualitatives Merkmal mit nur 2 Ausprägungen (z.B. männlich / weiblich, Raucher / Nichtraucher) heißen *alternativ*. Ein qualitatives Merkmal kann *ordinal* (wenn sich eine natürliche lineare Ordnung in den Merkmalsausprägungen finden lässt, wie z.B. gut / mittel / schlecht bei Qualitätsbewertung in Umfragen oder sehr gut / gut / befriedigend / ausreichend / mangelhaft / ungenügend bei Schulnoten) oder *nominal* (wenn eine solche Ordnung nicht vorhanden ist) sein. Beispiele von nominalen Merkmalen sind Fahrzeugmarken in der KFZ-Versicherung (z.B. BMW, Peugeot, Volvo, usw.) oder Führerscheinklassen (A, B, C, ...). Datenmerkmale können auch mehrdimensionale Ausprägungen haben. In dieser Vorlesung behandeln wir jedoch hauptsächlich eindimensionale Merkmale.

1.3 Statistische Daten und Stichproben

Aus den obigen Beispielen wird klar, dass ein Statistiker mit Datensätzen der Form $(x_1, \ldots, x_n)$ arbeitet, wobei die Einzeleinträge x_i aus einer Grundgesamtheit $G \subset \mathbb{R}^k$ stammen, die hypothetisch unendlich groß ist. Der vorliegende Datensatz $(x_1, \ldots, x_n)$ wird auch *(konkrete) Stichprobe* von Umfang n genannt. Die Menge B aller potentiell möglichen Stichproben bezeichnen wir als *Stichprobenraum* und setzen zur Vereinfachung der Notation $B = \mathbb{R}^k$. In diesem Skript werden wir meistens die univariate statistische Analyse (also $k = 1$ ein eindimensionales Merkmal) betreiben. In der beschreibenden Statistik arbeitet man mit Stichproben $(x_1, \ldots, x_n)$ und ihren Funktionen, um diese Daten visualisieren zu können. Für die Aufgabe der schließenden Statistik jedoch reicht diese Datenebene nicht mehr aus. Daher wird die zweite Ebene der Betrachtung eingeführt, die sogenannte *Modellebene*. Dabei wird angenommen, dass die konkrete Stichprobe $(x_1, \ldots, x_n)$ eine *Realisierung* eines stochastischen Modells $(X_1, \ldots, X_n)$ darstellt, wobei $X_1, \ldots, X_n$ (meistens unabhängige identisch verteilte) Zufallsvariablen auf einem (nicht näher spezifiziertem) Wahrscheinlichkeitsraum $(\Omega, \mathcal{F}, \mathbb{P})$ sind. Diese Zufallsvariablen $X_i, \quad i = 1, \ldots, n$ können als konsequente Beobachtungen eines Merkmals interpretiert werden. In Bsp. 1.1.1, 1) z.B. die Erbsenform mit

$$X_i = \begin{cases} 0, & \text{falls Erbse } i \text{ rund,} \\ 1, & \text{falls Erbse } i \text{ eckig,} \end{cases} \qquad i = 1, \ldots, n\,.$$

Der Vektor $(X_1, \ldots, X_n)$ wird dabei *Zufallsstichprobe* genannt. Man setzt weiter voraus, dass $EX_i^2 < \infty \; \forall i = 1, \ldots, n$, damit man von der Varianz $\operatorname{Var} X_i$ der Einzeleinträge sprechen kann. Es wird außerdem angenommen, dass ein $\omega \in \Omega$ existiert, sodass $X_i(\omega) = x_i \quad \forall i = 1, \ldots, n$. Sei F die Verteilungsfunktion der Zufallsvariablen X_i. Eine der wichtigsten Aufgaben der Statistik ist die Bestimmung von F (man sagt, „Schätzung von F“) aus den konkreten Daten $(x_1, \ldots, x_n)$. Dabei können auch Momente von F und ihre Funktionen (Erwartungswert, Varianz, Schiefe, usw.) von Interesse sein.

1.4 Stichprobenfunktionen

Um die obigen Aufgaben erfüllen zu können, braucht man gewisse Funktionen $\varphi : \mathbb{R}^n \to \mathbb{R}^m, \quad m \in \mathbb{N}$ auf dem Stichprobenraum, die diese Stichprobe bewerten.

Definition 1.4.1
Eine Borel-messbare Abbildung $\varphi : \mathbb{R}^n \to \mathbb{R}^m$ heißt *Stichprobenfunktion*. Wenn man auf der Modellebene mit einer Zufallsstichprobe $(X_1, \ldots, X_n)$ arbeitet, so heißt die Zufallsvariable

$$\varphi(X_1, \ldots, X_n)$$

eine *Statistik*. In der Schätztheorie spricht man dabei von *Schätzern* und bei statistischen Tests wird $\varphi(X_1, \ldots, X_n)$ *Teststatistik* genannt.

Beispiele für Stichprobenfunktionen sind unter anderen das *Stichprobenmittel*

$$\bar{x}_n = \frac{1}{n} \sum_{i=1}^{n} x_i \,,$$

die *Stichprobenvarianz*

$$s_n^2 = \frac{1}{n-1} \sum_{i=1}^{n} \left(x_i - \bar{x}_n\right)^2$$

und die *Ordnungsstatistiken*

$$x_{(1)} \leq x_{(2)} \leq \ldots \leq x_{(n)} \,,$$

die entstehen, wenn man eine Stichprobe, die aus quantitativen Merkmalen besteht, linear ordnet ($x_{(1)} = \min_{i=1,\ldots,n} x_i, \ldots, x_{(n)} = \max_{i=1\ldots n} x_i$). Weitere Beispiele und ihre Charakteristiken werden in Kapitel 2 gegeben.

2 Beschreibende Statistik

Sei eine konkrete Stichprobe $(x_1, \ldots, x_n)\,, \quad x_i \in \mathbb{R}$ gegeben, wobei die x_i als Realisierungen der Zufallsvariablen $X_i \stackrel{d}{=} X$ mit Verteilunfsfunktion F interpretiert werden können.

2.1 Verteilungen und ihre Darstellungen

In diesem Abschnitt werden wir Methoden zur statistischen Beschreibung und grafischen Darstellung der (unbekannten) Verteilung F betrachten.

2.1.1 Häufigkeiten und Diagramme

Falls das quantitative Merkmal X eine endliche Anzahl von Ausprägungen $\{a_1, \ldots, a_k\}$, $a_1 < a_2 < \ldots < a_k$, besitzt, also

$$\mathbb{P}\big(X \in \{a_1, \ldots, a_k\}\big) = 1\,,$$

dann kann eine Schätzung der Zähldichte $p_i = P(X = a_i)$ von X aus den Daten $(x_1, \ldots, x_n)$ grafisch dargestellt werden. Ähnliche Darstellungen sind für die Dichte $f(x)$ von absolut stetigen Merkmalen X möglich, wobei ihr Wertebereich C sich in k Klassen aufteilen lässt: $(c_{i-1}, c_i]\,, \quad i = 1, \ldots, k$, wobei $c_0 = -\infty$, $c_1 < \ldots < c_{k-1}$, $c_k = \infty$ ist. Dann kann die Zähldichte $p_i = \mathbb{P}\big(X \in (c_{k-1}, c_k]\big)$ als

$$p_i = \int_{c_{i-1}}^{c_i} f(x)\, dx_i\,, \qquad i = 0, \ldots, k$$

betrachtet werden.

Definition 2.1.1

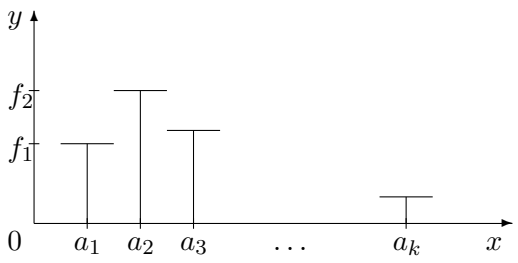
1. Die *absolute Häufigkeit* von Merkmalsausprägung a_i bzw. Klasse $(c_{i-1}, c_i]\,, \quad i = 1 \ldots k$ ist $n_i = \#\{x_j,\, j = 1, \ldots, n : x_j = a_i\}$ bzw. $n_i = \#\{x_j,\, j = 1, \ldots, n :\, x_j \in (c_{i-1}, c_i]\}$.

2. Die *relative Häufigkeit* von Merkmalsausprägung a_i bzw. Klasse $(c_{i-1}, c_i]$ ist $f_i = n_i/n\,,$ $i = 1 \ldots k$.

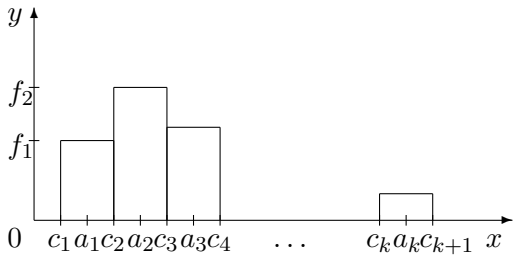
Es gilt offensichtlich $n = \sum_{i=1}^k n_i\,, \quad 0 \le f_i \le 1\,, \quad \sum_{i=1}^k f_i = 1$. Die absoluten und relativen Häufigkeiten werden oft in Häufigkeitstabellen zusammengefasst. Zu ihrer Visualisierung dienen so genannte *Diagramme*. Es wird grundsätzlich zwischen *Histogrammen* und *Kreisdiagrammen* unterschieden.

1. *Histogramme* werden gebildet, indem man die Paare (a_i, f_i) (bzw. $(1/2(c_1 + x_{(1)}), f_1)$, $(1/2(c_{i-1} + c_i), f_i)$, $i = 2, \ldots, k-1$, $(1/2(c_{k-1} + x_{(n)}), f_k)$ im absolut stetigen Fall, wobei hier die Bezeichnung $a_i = 1/2(c_{i-1} + c_i)$ verwendet wird) auf der Koordinatenebene (x, y) folgendermaßen aufträgt:

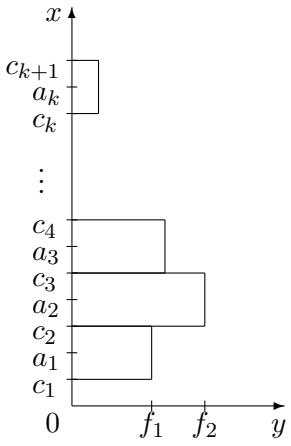
   - *Stabdiagramm:* f_i wird als Höhe des senkrechten Strichs über a_i dargestellt:



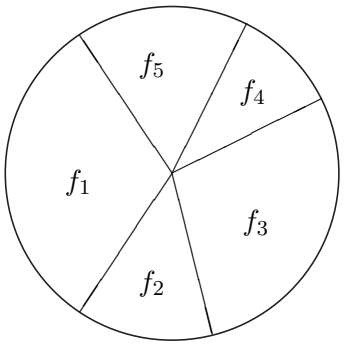
- *Säulendiagramm:* genauso wie ein Stabdiagramm, nur werden Striche durch Säulen der Form $(c_{i-1}, c_i] \times f_i$ ersetzt, wobei im diskreten Fall die Aufteilung der reellen Achse $-\infty = c_0 < c_1 < c_2 < \ldots < c_{k-1} < c_k = \infty$ in Intervalle beliebig vorgenommen werden kann.



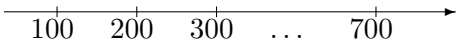
- *Balkendiagramm:* genauso wie Säulendiagramm, nur mit vertikalen statt horizontaler x-Achse.



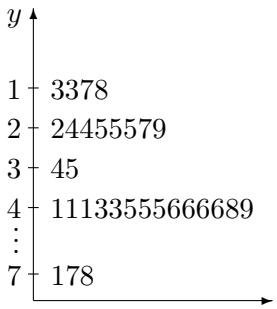
2. *Kreisdiagramme (Tortendiagramme):*
   Ein Kreis wird in Segmente mit Öffnungswinkel α_i eingeteilt, die proportional zu f_i sind: $\alpha_i = 2\pi f_i\,, \quad i = 1, \ldots, n.$



3. *Stamm-Blatt-Diagramme (stem-leaf display):*
   Diese werden heutzutage relativ selten und nur für kleine Datensätze verwendet. Dabei arbeitet man mit Stichprobenwerten, die auf ganze Zahlen gerundet sind. Sei $(x_1, \ldots, x_n)$ eine Stichprobe von solchen Werten, die Ausprägungen eines quantitativen Merkmals sind. Zunächst teilt man den Wertebereich $[x_{(1)}, x_{(n)}]$ in Klassen gleicher Breite $10^d, \quad d \in \mathbb{N}$, wobei jede Klasse mit den ersten Ziffern der dazugehörigen Beobachtungen markiert wird. Zum Beispiel, wenn die Klasseneinteilung so aussieht



   werden die Klassen $[100(i-1), 100i)$ mit den Zahlen i markiert und auf der y-Achse wie folgt aufgetragen:



   Auf diese Weise wird der Stamm des Baumes festgelegt. In jeder Klasse ordnet man Beobachtungen ihrer Größe nach und rundet sie auf die Stelle, die nach der gewählten Genauigkeit des Stammes folgt. Als Beispiel erhält man aus $127 \to 130$, aus $652 \to 650$ usw. und trägt diese Beobachtungen als Blätter des Baums horizontal ihrer Reihenfolge nach als 3 in Klasse 1 und 5 in Klasse 6 auf. Dabei darf man nicht vergessen, die Einheit zu notieren: $1/3 = 130$, um sich das Rückrechnen zu ermöglichen. Bei der Wahl der Klassenanzahl m hält man sich an die Faustregel $m \approx 10 \log_{10} n$, um einerseits den Dateverlust durch das unnötige Runden zu minimieren und andererseits dass Diagramm so übersichtlich wie möglich zu halten.

Bemerkung 2.1.1
Die in Abschnitt 2.1.1 betrachteten Methoden dienen der Visualisierung von (Zähl-) Dichten der

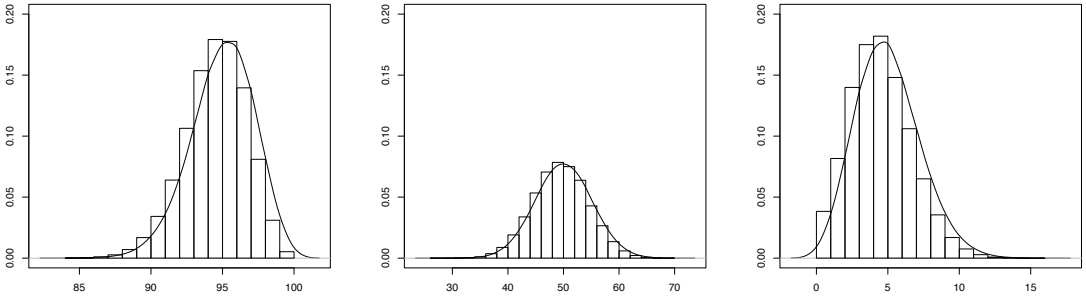


Abb. 2.1: Das Histogramm der Daten mit einer rechtssteilen (linksschiefen), symmetrischen und linkssteilen (rechtsschiefen) Verteilung und ihre Dichte.

Verteilung eines beobachteten Merkmals X. Aus dem Histogramm kann z.B. die Interpretation der Form der Dichte abgelesen werden:

Ist die zugrundeliegende Verteilung F_X symmetrisch bzw. linkssteil (rechtsschief) oder rechtssteil (linksschief) (vgl. Abb. 2.1) oder ist sie unimodal (d.h. eingipflig), bimodal (d.h. mit 2 Gipfeln) oder multimodal (also mit mehreren Gipfeln) (vgl. Abb. 2.2).

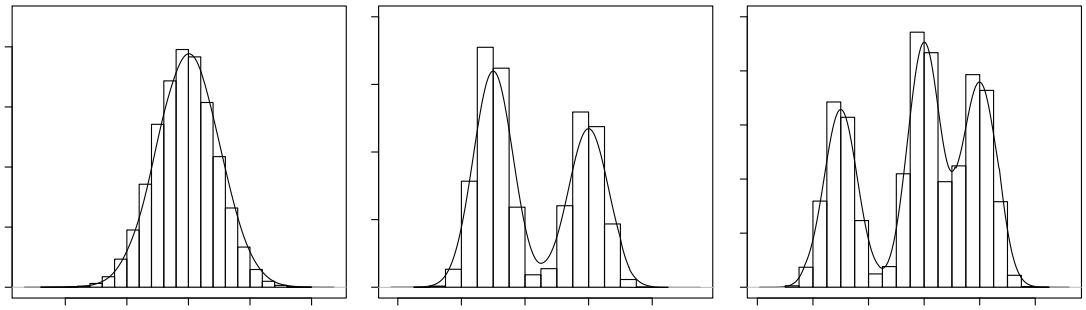


Abb. 2.2: Histogramm der Daten mit der Dichte einer unimodalen, bimodalen und multimodalen Verteilung

2.1.2 Empirische Verteilungsfunktion

Es sei eine konkrete Stichprobe $(x_1, \ldots, x_n)$ gegeben, die eine Realisierung des statistischen Modells $(X_1, \ldots, X_n)$ ist, wobei $X_1, \ldots, X_n$ unabhängig identisch verteilte Zufallsvariablen mit Verteilungsfunktion $F_X : X_i \stackrel{d}{=} X \sim F_X$ sind. Wie kann die unbekannte Verteilungsfunktion F_X aus den Daten $(x_1, \ldots, x_n)$ rekonstruiert (die Statistiker sagen „geschätzt“) werden? Dies ist mit Hilfe der sogenannten empirischen Verteilungsfunktion möglich:

Definition 2.1.2

1. Die Funktion $\hat{F}_n(x) = \#\{x_i : x_i \leq x\,,\, i = 1, \ldots, n\}/n\,, \quad \forall x \in \mathbb{R}$ heißt *empirische Verteilungsfunktion der konkreten Stichprobe* $(x_1, \ldots, x_n)$. Dabei gilt $\hat{F}_n : \mathbb{R}^{n+1} \to [0, 1]$, weil $\hat{F}_n(x) = \varphi(x_1, \ldots, x_n, x)$.

2. Die mit $x \in \mathbb{R}$ indizierte Zufallsvariable $\hat{F}_n : \Omega \times \mathbb{R} \to [0,1]$ heißt *empirische Verteilungsfunktion der Zufallsstichprobe* $(X_1, \ldots, X_n)$, wenn

$$\hat{F}_n(x,\omega) = \hat{F}_n(x) = \frac{1}{n}\#\{X_i,\, i = 1,\ldots,n :\, X_i(\omega) \le x\}\,, \quad x \in \mathbb{R}\,.$$

Äquivalent zur Definition 2.1.2 kann man

$$\hat{F}_n(x) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{I}(x_i \le x)\,, \quad x \in \mathbb{R}$$

schreiben, wobei

$$\mathbb{I}(x \in A) = \begin{cases} 1, & x \in A \\ 0, & \text{sonst.} \end{cases}$$

Es gilt

$$\hat{F}_n(x) = \begin{cases} 1, & x \ge x_{(n)}\,, \\ \frac{i}{n}, & x_{(i)} \le x < x_{(i+1)}\,, \quad i = 1,\ldots,n-1\,, \\ 0, & x < x_{(1)}\,. \end{cases}$$

Dabei ist die Höhe des Sprungs an Stelle $x_{(i)}$ gleich der relativen Häufigkeit f_i des Wertes $x_{(i)}$. Falls $x_{(i)} = x_{(i+1)}$ für ein $i \in \{1,\ldots,n\}$, so tritt der Wert i/n nicht auf. In Abbildung 2.3 sieht man, dass $\hat{F}_n(x)$ eine rechtsstetige monoton nichtfallende Treppenfunktion ist, für die

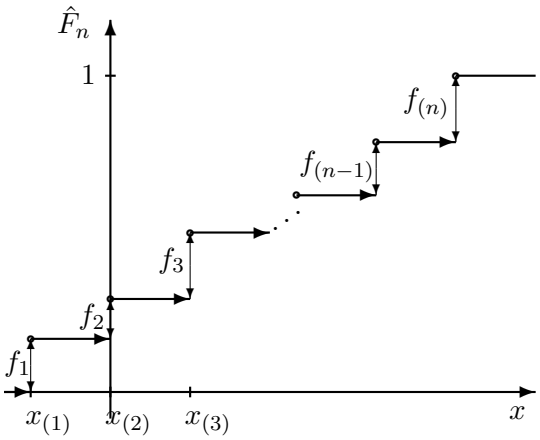
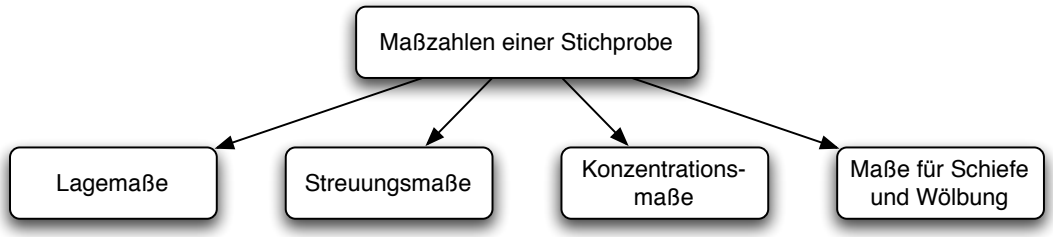


Abb. 2.3: Eine typische empirische Verteilungsfunktion

$\hat{F}_n(x) \underset{x\to-\infty}{\longrightarrow} 0\,,\ \hat{F}_n(x) \underset{x\to\infty}{\longrightarrow} 1$ gilt.

Übungsaufgabe 2.1.1
Zeigen Sie, dass $\hat{F}_n(x)$ eine Verteilungsfunktion ist.

2.2 Beschreibung von Verteilungen



Es sei eine konkrete Stichprobe $(x_1, \ldots, x_n)$ gegeben. Im Folgenden werden Kennzahlen (die sogenannten Maße) dieser Stichprobe betrachtet, welche die wesentlichen Aspekte der der Stichprobe zugrundeliegenden Verteilung wiedergeben:

1. Wo liegen die Werte x_i (Mittel, Ordnungsstatistiken, Quantile)? $\Longrightarrow$ Lagemaße
2. Wie stark streuen die Werte x_i (Varianz) $\Longrightarrow$ Streuungsmaße
3. Wie stark sind die Werte x_i in gewissen Bereichen von $\mathbb{R}$ konzentriert $\Longrightarrow$ Konzentrationsmaße
4. Wie schief bzw. gewölbt ist die Verteilung von $X \Longrightarrow$ Maße für Schiefe und Wölbung

2.2.1 Lagemaße

Man unterscheidet folgende wichtige Lagemaße:

- Mittelwerte: Stichprobenmittel (arithmetisch), geometrisches und harmonisches Mittel, gewichtetes Mittel, getrimmtes Mittel
- Ordnungsstatistiken und Quantile, insbesondere Median und Quartile
- Modus

Betrachten wir sie der Reihe nach:

1. *Mittelwertbildung:* Seit der Antike kennt man mindestens 3 Arten der *Mittelberechnung* von n Zahlen $(x_1, \ldots, x_n)$:
   - *arithmetisch:* $\bar{x}_n = 1/n \sum_{i=1}^n x_i\,, \quad \forall x_1, \ldots, x_n \in \mathbb{R}$,
   - *geometrisch:* $x_n^g = \sqrt[n]{x_1 \cdot \ldots \cdot x_n}\,, \quad x_1, \ldots, x_n > 0$,
   - *harmonisch:* $x_n^h = \left(1/n \sum_{i=1}^n x_i^{-1}\right)^{-1}\,, \quad x_1, \ldots, x_n \neq 0$.

   a) Das *arithmetische Mittel* wird in der Statistik am meisten benutzt, weil es keine Voraussetzungen über den Wertebereich von $x_1, \ldots, x_n$ braucht. Es wird auch *Stichprobenmittel* genannt. Offensichtlich ist $\bar{x}_n$ ein Spezialfall des sogenannten gewichteten Mittels $x_n^w = \sum_{i=1}^n w_i x_i$, wobei für die Gewichte $w_i \geq 0 \quad \forall i = 1, \ldots, n$ und $\sum_{i=1}^n w_i = 1$ gilt. Als eine natürliche Gewichtewahl kommt $w_i = 1/n\,, \quad \forall i = 1, \ldots, n$ bei einer konkreten Stichprobe $(x_1, \ldots, x_n)$ in Frage. Die Summe aller Abweichungen von $\bar{x}_n$ ist Null, denn $\sum_{i=1}^n (x_i - \bar{x}_n) = n\bar{x}_n - n\bar{x}_n = 0$, d.h. $\bar{x}_n$ stellt geometrisch den Schwerpunkt der Werte x_i dar, falls jedem Punkt eine Einheitsmasse zugeordnet

wird. Wenn es in der Stichprobe große Ausreißer gibt, so beeinflussen sie das Stichprobenmittel entscheident und erschweren so die objektive Datenanalyse. Deshalb verwendet man oft die robuste Version des arithmetischen Mittels, das sogenannte *getrimmte Mittel*:

$$\tilde{x}_n^{(k)} = \frac{1}{n-2k} \sum_{i=k+1}^{n-k} x_{(i)} \,,$$

bei dessen Berechnung die k kleinsten und k größten Ausreißer ausgelassen werden, wobei $k \ll n/2$.

b) Das *geometrische Mittel* wird hauptsächlich bei der Beobachtung von Wachstums- und Zinsfaktoren verwendet. Sei $x_i = B_i/B_{i-1}\,, \quad i = 1, \ldots, n$ der Wachstumsfaktor des Merkmals B_i das in den Jahren $i = 1, \ldots, n$ beobachtet wurde (z.B. Inflationsfaktor). Dann ist $B_n = B_0 \cdot x_1 \cdot \ldots \cdot x_n$ und somit wäre der Zins im Jahre n

$$B_n^g = B_0 \cdot x_1 \cdot \ldots \cdot x_n = B_0 \cdot (x_n^g)^n \,.$$

Für das geometrische Mittel gilt

$$\log x_n^g = \frac{1}{n} \sum_{i=1}^{n} \log x_i \leq \log \left(\frac{1}{n} \sum_{i=1}^{n} x_i \right)$$

wegen der Konkavität des Logarithmus, d.h. $\log x_n^g = \overline{\log x_n} \leq \log \bar{x}_n$ und somit $x_n^g \leq \bar{x}_n$, wobei $x_n^g = \bar{x}_n$ genau dann, wenn $x_1 = \ldots = x_n$.

c) Das *harmonische Mittel* wird bei der Ermittlung von z.B. durchschnittlicher Geschwindigkeiten gebraucht.

Beispiel 2.2.1
Seien x_i Geschwindigkeiten mit denen Bauteile eine Produktionslinie der Länge l durchlaufen. Die gesamte Bearbeitungszeit ist $l/x_1 + \ldots + l/x_n$ und die Durchschnittslaufgeschwindigkeit

$$\frac{l + \ldots + l}{l/x_1 + \ldots + l/x_n} = x_n^h \,.$$

2. *Ordnungsstatistiken und Quantile*

Definition 2.2.1
Die *Ordnungsstatistiken* $x_{(i)}\,, \quad i = 1, \ldots, n$ der Stichprobe $(x_1, \ldots, x_n)$ sind durch die messbare Permutation $\varphi(x_1, \ldots, x_n)$ gegeben, so dass

$$x_{(i)} = \min \left\{ x_j : \#\{k : x_k \leq x_j\} \geq i \right\}, \quad \forall i = 1, \ldots, n \,.$$

Somit gilt $x_{(1)} \leq x_{(2)} \leq \ldots \leq x_{(n)}$. Dieselbe Definition kann auch auf der Modellebene gegeben werden.

Definition 2.2.2

a) Sei nun X die Zufallsvariable, die das Merkmal modelliert. Sei F_X ihre Verteilungsfunktion. Die verallgemeinerte Inverse von F_X, definiert durch

$$F_X^{-1}(y) = \inf \left\{ x : F_X(x) \geq y \right\}, \quad y \in [0,1] \,,$$

heißt *Quantilfunktion* von F_X bzw. X. Es gilt $F_X^{-1} : [0,1] \to \mathbb{R} \cup \{\pm\infty\}$. Die Zahl $F_X^{-1}(\alpha)\,, \alpha \in [0,1]$ wird α-Quantil von F_X genannt.

b)
- $F_X^{-1}(0,25)$ heißt *unteres Quartil*,
- $F_X^{-1}(0,75)$ heißt *oberes Quartil*,
- $F_X^{-1}(0,5)$ heißt der *Median* der Verteilung von X.

Zwischen Ordnungsstatistiken und Quantilen besteht ein enger Zusammenhang. So bedeutet $F_X^{-1}(\alpha)\,, \quad \alpha \in (0,1)$, dass ca. $\alpha \cdot 100\%$ aller Merkmalsausprägungen in der Stichprobe $(x_1, \ldots, x_n)$ unter $F_X^{-1}(\alpha)$ und ca. $(1-\alpha) \cdot 100\%$ über $F_X^{-1}(\alpha)$ liegen (im absolut stetigen Fall). Insbesondere gilt $F_X^{-1}(\alpha) \approx x_{([n\alpha])}$, deshalb werden Ordnungsstatistiken auch *empirische Quantile* genannt. Dabei ist x_α definiert als

$$x_\alpha = \begin{cases} x_{([n\alpha]+1)}\,, & n\alpha \notin \mathbb{N} \\ 1/2(x_{([n\alpha])} + x_{([n\alpha]+1)})\,, & n\alpha \in \mathbb{N} \end{cases}.$$

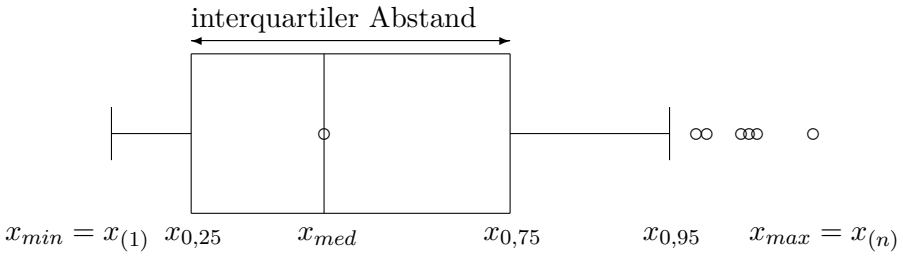
Dies ist die allgemeine Definition des *α-empirischen Quantils.*

Der *empirische Median* ist

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

Somit sind mindestens 50% aller Stichprobenwerte kleiner gleich und 50% größer gleich x_{med}. Der Median ist ein Lagemaß, das ein robuster Ersatz für den Mittelwert darstellt, denn er ist bzgl. Ausreißern in der Stichprobe nicht sensibel.

Die oben genannten Statistiken werden in einem Box-Plot zusammengefasst und grafisch dargestellt:



Manchmal werden $x_{(1)}$ und $x_{(n)}$ durch $x_{0,05}$ und $x_{0,95}$ ersetzt. Die restlichen Werte werden darüber hinaus als Einzelpunkte auf der x-Achse abgebildet. Dann liegt ein sogenannter *modifizierter Box-Plot* vor.

3. *Modus:* Sei $(x_1, \ldots, x_n)$ eine Stichprobe, die aus n unabhängigen Realisierungen des Merkmals X besteht. Sei $f(x)$ $(p(x))$ die (Zähl-) Dichte von X.

Definition 2.2.3

a) Der Wert $x_{mod} = \arg\max f(x)$ $(\arg\max p(x))$ wird der *Modus der Verteilung von X* genannt (vgl. Abb. 2.4).

b) Empirisch wird x_{mod} als $X_{(m)}$ für $m = \arg\max f_i$ definiert, also als ein Stichprobenwert mit der größten Häufigkeit des Vorkommens in der Stichprobe.

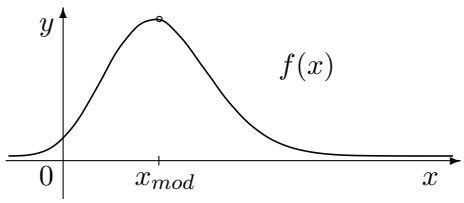


Abb. 2.4: Veranschaulichung des Modus

Den Mittelwert $\bar{x}_n$, Median x_{med} und Modus x_{mod} kann man auch wie folgt definieren:

$$\bar{x}_n = \arg\min_x \sum_{i=1}^n (x_i - x)^2$$
$$x_{med} = \arg\min_x \sum_{i=1}^n |x_i - x|$$
$$x_{mod} = \arg\min_x \sum_{i=1}^n \mathbb{I}(x_i \neq x)$$

Übungsaufgabe 2.2.1
Zeigen Sie die Äquivalenz der oben genannten Definitionen des Mittelwerts $\bar{x}_n$, Medians x_{med} und des Modus x_{mod} zu den bekannten Definitionen.

Die Größen $\bar{x}_n$, x_{med} und x_{mod} können auch zur Beschreibung der Symmetrie einer unimodalen Verteilung F_X von Daten $(x_1, \ldots, x_n)$ verwendet werden, da

- bei symmetrischen Verteilung F_X gilt $\bar{x}_n \approx x_{med} \approx x_{mod}$
- bei linkssteilen Verteilung F_X gilt $x_{mod} < x_{med} < \bar{x}_n$
- bei rechtssteilen Verteilung F_X gilt $\bar{x}_n < x_{med} < x_{mod}$.

2.2.2 Streuungsmaße

Bekannte Streuungsmaße einer konkreten Stichprobe $(x_1, \ldots, x_n)$ sind die folgenden Größen:

- *Spannweite* $x_{(n)} - x_{(1)}$,
- *empirische Varianz* $\bar{s}_n^2 = \frac{1}{n}\sum_{i=1}^n (x_i - \bar{x}_n)^2$,
- *Stichprobenvarianz* $s_n^2 = \frac{1}{n-1}\sum_{i=1}^n (x_i - \bar{x}_n)^2 = \frac{n}{n-1}\bar{s}_n^2$,
- *empirische Standardabweichungen* $\bar{s}_n = \sqrt{\bar{s}_n^2}$, $s_n = \sqrt{s_n^2}$,
- *empirischer Variationskoeffizient* $\gamma_n = s_n/\bar{x}_n$, falls $\bar{x}_n > 0$.

Die Spannweite zeigt die *maximale Streuung* in den Daten, wobei sich die empirische Varianz mit der *mittleren quadratischen Abweichung* vom Stichprobenmittel auseinandersetzt. Hier sind einige Eigenschaften von $\bar{s}_n^2$ (bzw. s_n^2, da sie sich nur durch einen Faktor unterscheiden):

Lemma 2.2.1

1. Für jedes $b \in \mathbb{R}$ gilt

$$\sum_{i=1}^{n}(x_i - b)^2 = \sum_{i=1}^{n}(x_i - \bar{x}_n)^2 + n(\bar{x}_n - b)^2$$

und somit für $b = 0$

$$\bar{s}_n^2 = \frac{1}{n}\sum_{i=1}^{n}\left(x_i^2 - \bar{x}_n^2\right) \qquad \text{bzw.} \qquad s_n^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(x_i^2 - \bar{x}_n^2\right).$$

2. *Transformationsregel:*
Falls die Daten $(x_1, \ldots, x_n)$ linear transformiert werden, d.h. $y_i = ax_i + b$, $a \neq 0$, $b \in \mathbb{R}$, dann gilt

$$\bar{s}_{n,y}^2 = a^2\bar{s}_{n,x}^2 \qquad \text{bzw.} \qquad \bar{s}_{n,y} = |a|\bar{s}_{n,x},$$

wobei

$$\bar{s}_{n,y}^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y}_n)^2, \qquad \bar{s}_{n,x}^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x}_n)^2$$

Beweis

1. $$\begin{aligned}\sum_{i=1}^{n}(x_i - b)^2 &= \sum_{i=1}^{n}(x_i - \bar{x}_n + \bar{x}_n - b)^2 \\ &= \sum_{i=1}^{n}(x_i - \bar{x}_n)^2 + 2\sum_{i=1}^{n}(x_i - \bar{x}_n)\cdot(\bar{x}_n - b) + \sum_{i=1}^{n}(\bar{x}_n - b)^2 \\ &= \sum_{i=1}^{n}(x_i - \bar{x}_n)^2 + 2(\bar{x}_n - b)\cdot\underbrace{\sum_{i=1}^{n}(x_i - \bar{x}_n)}_{=0} + n(\bar{x}_n - b)^2, \quad \forall b \in \mathbb{R}.\end{aligned}$$

2. $$\bar{s}_{n,y}^2 = \frac{1}{n}\sum_{i=1}^{n}(ax_i + b - a\bar{x}_n - b)^2 = \frac{a^2}{n}\sum_{i=1}^{n}(x_i - \bar{x}_n)^2 = a^2\bar{s}_{n,x}^2.$$

□

Der Skalierungsunterschied zwischen $\bar{s}_n^2$ und s_n^2 ist den Eigenschaften der *Erwartungstreue* von s_n^2 zu verdanken, die später im Laufe dieser Vorlesung behandelt wird, und besagt, dass für eine Zufallsstichprobe $(X_1, \ldots, X_n)$ mit X_i unabhängig identisch verteilt, $X_i \sim X$, $\operatorname{Var} X = \sigma^2 \in (0, \infty)$ gilt $\mathbb{E}s_n^2 = \sigma^2$, wobei $\mathbb{E}\bar{s}_n^2 = \frac{n}{n-1}\sigma^2 \underset{n\to\infty}{\longrightarrow} \sigma^2$. Das heißt, während bei der Verwendung von s_n^2 zur Schätzung von σ^2 kein Fehler „im Mittel“ gemacht wird, ist diese Aussage für $\bar{s}_n^2$ nur asymptotisch (für große Datenmengen n) richtig.

Aufgrund von $\sum_{i=1}^{n}(x_i - \bar{x}_n) = 0$ ist z.B. $x_n - \bar{x}_n$ durch $x_i - \bar{x}_n$, $i = 1, \ldots, n-1$ bestimmt. Somit verringert sich die *Anzahl der Freiheitsgrade* in der Summe $\sum_{i=1}^{n}(x_i - \bar{x}_n)^2$ um 1 und somit scheint die Normierung $\frac{1}{n-1}$ plausibel zu sein.

Die *Standardabweichungen* $\bar{s}_n$ und s_n werden verwendet, damit man die selben Einheiten (und nicht ihre Quadrate, also z.B. Euro und nicht Euro^2) erhält. Für normalverteilte Stichproben

($X \sim N(\mu, \sigma^2)$) liefert $\bar{s}_n$ auch die „k-Sigma-Regel“ (vgl. Vorlesung WR), die besagt, dass in den Intervallen

$$\begin{aligned} [\bar{x}_n - \bar{s}_n, \bar{x}_n + \bar{s}_n] \quad & \text{ca.} \quad 68\%\,, \\ [\bar{x}_n - 2\bar{s}_n, \bar{x}_n + 2\bar{s}_n] \quad & \text{ca.} \quad 95\%\,, \\ [\bar{x}_n - 3\bar{s}_n, \bar{x}_n + 3\bar{s}_n] \quad & \text{ca.} \quad 99\% \end{aligned}$$

aller Daten liegen.

Der Vorteil vom *empirischen Variationskoeffizienten* ist, dass er *maßstabsunabhängig* ist und somit den Vergleich von Streuungseigenschaften unterschiedlicher Stichproben zulässt.

2.2.3 Konzentrationsmaße

Insbesondere in den Wirtschaftswissenschaften interessiert man sich oft für die Konzentration von Merkmalsausprägungen in der Stichprobe, z.B. wie sich das Familieneinkommen einer demographischen Einheit auf unterschiedliche Einkommensbereiche (Vielverdiener, Mittelstand, Wenigverdiener) aufteilt, oder wie sich der Markt auf Marktanbieter aufteilt (Marktkonzentration). Dabei ist es wünschenswert, diese Relation mit Hilfe weniger Zahlen oder einer Grafik zum Ausdruck zu bringen. Dies ist mit Hilfe folgender Stichprobenfunktionen möglich:

- *Lorenzkurve* L,
- *Gini-Koeffizient* G,
- *Konzentrationsrate* CR_g,
- *Herfindahl-Index* H.

1. Die Lorenzkurve wurde von M. Lorenz am Anfang des XX. Jahrhunderts für die Charakterisierung der Vermögenskonzentration benutzt. Sei $(x_1, \ldots, x_n)$ eine Stichprobe, die in aufsteigender Reihenfolge geordnet werden muss: $(x_{(1)}, \ldots, x_{(n)})$. Die *Lorenzkurve* verbindet Punkte
$$(0,0),\ (u_1, v_1),\ \ldots,\ (u_n, v_n),\ (1,1)$$
durch Liniensegmente, wobei $u_j = j/n$ der Anteil des j kleinsten Merkmalsträger und $v_j = \sum_{i=1}^{j} x_{(i)} / \sum_{i=1}^{n} x_i$ die kumulierte relative Merkmalssumme ist. Der Grundgedanke ist darzustellen, welcher Anteil des Merkmalsträgers auf welchen Anteil der Gesamtmerkmalssumme entfällt. Zum Beispiel lassen sich dadurch Aussagen wie etwa „Auf 20% aller Haushalte im Land entfällt 78% des Gesamteinkommens“ machen. Eine Interpretation der Lorenzkruve L ist nur an den Knoten (u_j, v_j) möglich: „Auf $u_j \cdot 100\%$ der kleinsten Merkmalsträger konzentrieren sich $v_j \cdot 100\%$ der Merkmalssumme“. Dabei liegt L auf $[0,1]^2$ immer zwischen der „line of perfect equality“ (l.p.e.) $v_i = u_i \quad \forall\, i$ (Einkommen ist absolut gleichmäßig—also „gerecht“—verteilt) und „line of perfect inequality“ (l.p.i.) $v = 0$, $u \in [0,1)$ und $(1,1)$ (das Gesamteinkommen besitzt nur die reichste Familie) und ist immer monoton und konvex. Auf Modellebene gibt es ein Analogon der Lorenzkurve. Dieses ist
$$L = \left\{ (u,v) \in [0,1]^2 : \quad v = \frac{\int_0^u F_X^{-1}(t)dt}{\int_0^1 F_X^{-1}(t)dt}\,, \qquad u \in [0,1] \right\}\,,$$

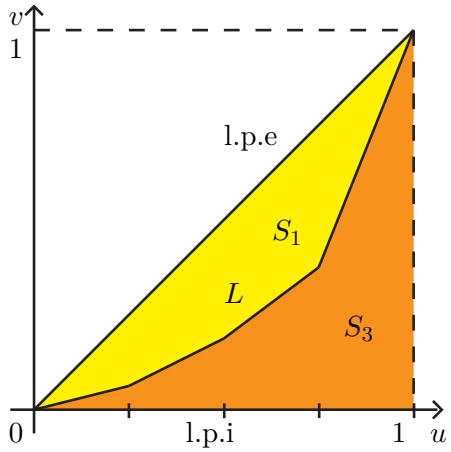


Abb. 2.5: Abbildung einer typischen Lorenzkurve

wobei

$$\mathbb{E}X = \int_0^1 F_X^{-1}(t)dt$$

(vgl. WR Satz 4.3.2). Dementsprechend können die Knoten (u_i, v_j) der oben eingeführten empirischen Lorenzkurve als

$$v_j = \frac{\sum_{i=1}^{j} \frac{x_{(i)}}{n}}{\bar{x}_n}$$

interpretiert werden.

2. Der *Gini-Koeffizient* G ist gegeben durch $G = S_1/S_2$, wobei S_1 die Fläche zwischen der Lorenzkurve L und der Diagonalen $v = u$, S_2 die Fläche zwischen der Diagonalen und der u-Achse $(= 1/2|[0,1]^2| = 1/2)$ ist.

   Satz 2.2.1 (Darstellung des Gini-Koeffizienten):
   Es gilt

$$G = 2S_1 = \frac{2\sum_{i=1}^{n} i x_{(i)}}{n\sum_{i=1}^{n} x_i} - \frac{n+1}{n}\,.$$

   Beweis Beginnen wir mit die Darstellung $G = n + 1/n - 2\bar{v}_n$ zu zeigen. Nach Definition ist

$$G = \frac{S_1}{S_2} = \frac{S_2 - S_3}{S_2} = 1 - \frac{S_3}{S_2} = 1 - 2S_3\,,$$

   wobei S_3 die Fläche zwischen der Lorenzkurve und der x-Achse ist (vgl. Abb. 2.5). Berechnen wir S_3:

   $S_3 = \sum_{j=1}^{n} F_j$, wobei $F_j = 1/n \cdot v_{j-1} + \frac{1}{2}\frac{1}{n} \cdot (v_j + v_{j-1}) = \frac{1}{2n}(v_j - v_{j-1})$ die Fläche unter einem Liniensegment der Lorenzkurve ist (vgl. Abb. 2.6). Es gilt

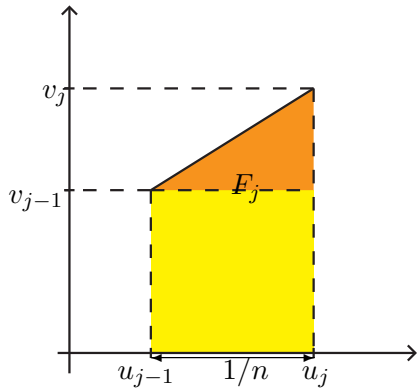


Abb. 2.6: Liniensegment der Lorenzkurve

$$S_3 = \frac{1}{2n}\sum_{j=1}^{n}(v_j + v_{j-1}) = \frac{1}{2n}\left(2\sum_{j=1}^{n} v_j - 1\right) = \bar{v}_n - \frac{1}{2n}\,,$$

somit

$$G = 1 - 2\bar{v}_n + \frac{1}{n} = \frac{n+1}{n} - 2\bar{v}_n\,.$$

Beweisen wir jetzt, dass

$$G = \frac{2\sum_{i=1}^{n} i x_{(i)}}{n\sum_{i=1}^{n} x_i} - \frac{n+1}{n}$$

ist. Sei $w = \sum_{i=1}^{n} i x_{(i)}$. Aufgrund der Definition von v_j gilt $s_j = \sum_{i=1}^{j} x_{(i)} = s_n \cdot v_j$, $\forall j = 1, \ldots, n$ und $x_{(i)} = s_i - s_{i-1}\,, \quad s_0 = 0$. Daher erhalten wir

$$\begin{aligned} w &= \sum_{i=1}^{n} i(s_i - s_{i-1}) = \sum_{i=1}^{n} i s_i - \sum_{i=0}^{n-1}(i+1)s_i = n s_n - \sum_{i=0}^{n-1} s_i \\ &= (n+1)s_n - \sum_{i=1}^{n} s_i = (n+1)s_n - s_n \cdot \sum_{i=1}^{n} v_i = (n+1)s_n - s_n \cdot n\bar{v}_n \end{aligned}$$

und somit

$$\frac{2\omega}{n s_n} - \frac{n+1}{n} = \frac{2w - (n+1)s_n}{n s_n} = \frac{2(n+1)s_n - 2 s_n n\bar{v}_n - (n+1)s_n}{n s_n} = \frac{n+1}{n} - 2\bar{v}_n = G\,.$$

□

Es gilt $G \in [0, (n-1)x/n]$, wobei

$$\begin{aligned} G_{min} &= 0 \quad \text{bei} \quad x_1 = x_2 = \ldots = x_n && \text{„perfect equality“,} \\ G_{max} &= \frac{n-1}{n} \quad \text{bei} \quad x_1 = \ldots = x_{n-1} = 0,\ x_n \neq 0 && \text{„perfect inequality“}\,. \end{aligned}$$

Somit hängt G_{max} vom Datenumfang ab. Um dies zu vermeiden, betrachtet man oft den normierten Gini-Koeffizienten

$$G^* = \frac{G}{G_{max}} = \frac{n}{n-1}G \quad \in [0,1]$$

(Lorenz-Münzner-Koeffizient).

3. *Konzentrationsrate CR_g:*
   In den Punkten 1) und 2) betrachteten wir die *relative Konzentration*, wie etwa bei der Fragestellung „Wieviel % der Familien teilen sich wieviel % des Gesamteinkommens?“. Dabei beantwortet die Konzentrationsrate die Frage „Wieviele Familien haben wieviel Prozent des Gesamteinkommens?“ für die g reichsten Familien, somit wird auch die absolute Anzahl aller Familien berücksichtigt.

   Sei $g \in \{1, \ldots, n\}$ und seien $x_{(1)} \leq \ldots \leq x_{(n)}$ die Ordnungsstatistiken der Stichprobe $(x_1, \ldots, x_n)$. Für $i \in \{1, \ldots, n\}$ sei

$$p_i = \frac{x_{(i)}}{\sum_{j=1}^n x_j} = \frac{x_{(i)}}{n\bar{x}_n} \tag{2.2.1}$$

4. der Merkmalsanteil der i-ten Eiheit.

   Dann gibt die *Konzentrationsrate* $CR_g = \sum_{i=n-g+1}^n p_i$ wieder, welcher Anteil des Gesamteinkommens von g reichsten Familien gehalten wird.

5. Der *Herfindahl-Index* ist definiert durch $M = \sum_{i=1}^n p_i^2$, wobei der Merkmalsanteil p_i nach (2.2.1) definiert ist. Bei der gleichen Verteilung des Einkommens ($x_1 = x_2 = \ldots = x_n$) gilt $H_{min} = 1/n$, bei völlig ungerechter Verteilung ($x_1 = \ldots = x_{n-1} = 0$, $x_n \neq 0$) $H_{max} = 1$. Sonst gilt $H \in [H_{min}, H_{max}]$, also $1/n \leq H \leq 1$. H ist umso kleiner, je gerechter das Gesamteinkommen verteilt ist.

2.2.4 Maße für Schiefe und Wölbung

Im Vorlesungsskript WR, Abschnitt 4.5 S. 99 wurden folgende Maße für Schiefe bzw. Wölbung der Verteilung einer Zufallsvariable X eingeführt:
Schiefe oder Symmetriekoeffizient:

$$\gamma_1 = \frac{\mu_3}{\sigma^3} = E(\tilde{X}^3)\,,$$

wobei

$$\mu_k = \mathbb{E}(X - \mathbb{E}X)^k\,, \quad \sigma^2 = \mu_2 = \operatorname{Var} X\,, \quad \tilde{X} = \frac{X - \mathbb{E}X}{\sigma}\,.$$

Wölbung (Exzess)

$$\gamma_2 = \frac{\mu_4}{\sigma^4} - 3 = \mathbb{E}(\tilde{X}^4) - 3\,,$$

vorausgesetzt, dass $\mathbb{E}(X^4) < \infty$. Für ihre Bedeutung und Interpretation siehe die oben genannten Seiten des WR-Vorlesungsskriptes. Falls nun das Merkmal X statistisch in einer Stichprobe $(x_1, \ldots, x_n)$ beobachtet wird, wie können γ_1 und γ_2 aus diesen Daten geschätzt und interpretiert werden?

Als Schätzer für das k-te zentrierte Moment $\mu_k = \mathbb{E}(X - \mathbb{E}X)^k$, $k \in \mathbb{N}$ schlagen wir

$$\hat{\mu}_k = \frac{1}{n}\sum_{i=1}^n (x_i - \bar{x}_n)^k$$

vor, die Varianz σ^2 wird durch

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

geschätzt. Somit bekommt man den Momentenkoeffizient der Schiefe (engl. „skewness“)

$$\hat{\gamma}_1 = \frac{\hat{\mu}_3}{\bar{s}_n^3} = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x}_n)^3}{\left(\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x}_n)^2\right)^{3/2}}\,.$$

Falls die Verteilung von X linksteil ist, überwiegen positive Abweichungen im Zähler und somit gilt $\hat{\gamma}_1 > 0$ für linkssteile Verteilungen. Analog gilt $\hat{\gamma}_1 \approx 0$ für symmetrische und $\hat{\gamma}_1 < 0$ für rechtssteile Verteilungen.

Das *Wölbungsmaß von Fisher* (engl. „kurtosis“) ist gegeben durch

$$\hat{\gamma}_2 = \frac{\hat{\mu}_4}{\bar{s}_n^4} - 3 = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x}_n)^4}{\left(\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x}_n^2)\right)^2} - 3\,.$$

Falls $\hat{\gamma}_2 > 0$ so ist die Verteilung von X steilgipflig, für $\hat{\gamma}_2 < 0$ ist sie flachgipflig. Falls $X \sim N(\mu, \sigma^2)$, so gilt $\hat{\gamma}_2 \approx 0$. Die Ursache dafür ist, dass die steilgifpligen Verteilungen schwerere Tails haben als die flachgipfligen. Als Maß dient dabei die Normalverteilung, für die $\gamma_1 = \gamma_2 = 0$ und somit $\hat{\gamma}_1 \approx 0$, $\hat{\gamma}_2 \approx 0$. So definiert, sind $\hat{\gamma}_1$ und $\hat{\gamma}_2$ nicht resistent gegenüber Ausreissern. Eine robuste Variante von $\hat{\gamma}_1$ ist beispielsweise durch den sogennanten *Quantilskoeffizienten der Schiefe*

$$\hat{\gamma}_q(\alpha) = \frac{(x_{1-\alpha} - x_{med}) - (x_{med} - x_\alpha)}{x_{1-\alpha} - x_\alpha}\,, \quad \alpha \in (0, 1/2)$$

gegeben.

Für $\alpha = 0,25$ erhält man den Quartilskoeffizienten. $\hat{\gamma}_q(\alpha)$ misst den Unterschied zwischen der Entfernung des α- und $(1-\alpha)$-Quantils zum Median. Bei linkssteilen (bzw. rechtssteilen) Verteilungen liegt das (untere) x_α-Quantil näher an (bzw. weiter entfernt von) dem Median. Somit gilt

- $\hat{\gamma}_q(\alpha) > 0$ für linkssteile Verteilungen,
- $\hat{\gamma}_q(\alpha) < 0$ für rechtssteile Verteilungen,
- $\hat{\gamma}_q(\alpha) = 0$ für symmetrische Verteilungen.

Durch das zusätzliche Normieren (Nenner) gilt $-1 \leq \hat{\gamma}_q(\alpha) \leq 1$.

2.3 Quantilplots (Quantil-Grafiken)

Nach der ersten beschreibenden Analyse eines Datensatzes $(x_1, \ldots, x_n)$ soll überlegt werden, mit welcher Verteilung diese Stichprobe modelliert werden kann. Hier sind die sogenannten *Quantilplots* behilflich, da sie grafisch zeigen, wie gut die Daten $(x_1, \ldots, x_n)$ mit dem Verteilungsgesetz G übereinstimmen, wobei G die Verteilungsfunktion einer hypothetischen Verteilung ist.

Sei X eine Zufallsvariable mit (unbekannter) Verteilungsfunktion F_X. Auf Basis der Daten $(X_1, \ldots, X_n)$, X_i unabhängig identisch verteilt und $X_i \stackrel{d}{=} X$ möchte man prüfen, ob $F_X = G$ für eine bekannte Verteilungsfunktion G gilt. Die Methode der *Quantil-Grafiken* besteht darin, dass man die entsprechenden Quantil-Funktionen $\hat{F}_n^{-1}$ und G^{-1} von $\hat{F}_n$ und G grafisch vergleicht. Hierzu

- plotte man $G^{-1}(k/n)$ gegen $\hat{F}_n^{-1}(k/n) = X_{(k)}\,, \quad k = 1, \ldots, n\,.$
- Falls die Punktwolke
$$\left\{\left(G^{-1}(k/n),\, X_{(k)}\right), \quad k = 1, \ldots, n\right\}$$
näherungsweise auf einer Geraden $y = ax + b$ liegt, so sagt man, dass $F_X(x) \approx G\left(\frac{x-a}{b}\right)$, $x \in \mathbb{R}$.

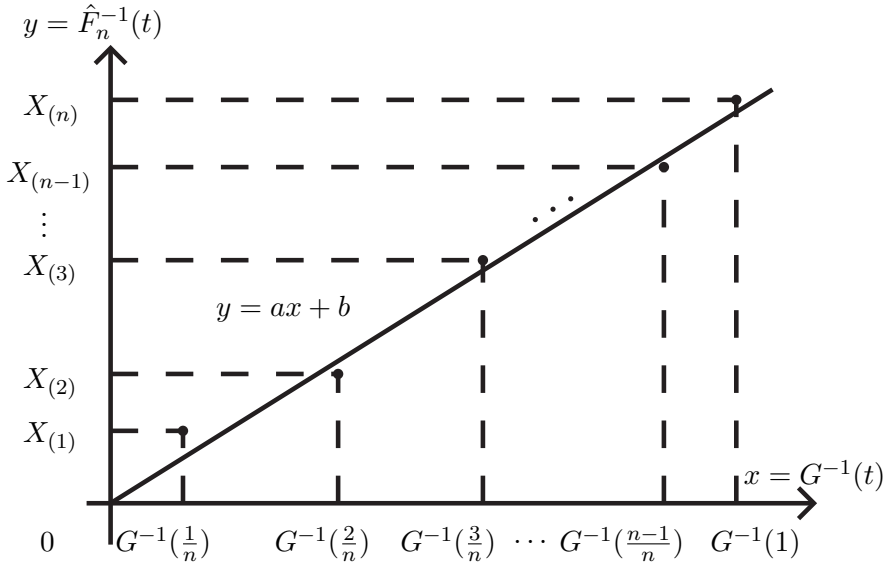


Abb. 2.7: Quantil-Grafik

Diese empirische Vergleichsmethode beruht auf folgenden Überlegungen:

- Man ersetzt die unbekannte Funktion F_X durch die aus den Daten berechenbare Funktion $\hat{F}_n$. Dabei macht man einen Fehler, der allerdings asymptotisch (für $n \to \infty$) klein ist. Dies folgt aus dem Satz 3.3.9 von Gliwenko-Cantelli, der besagt, dass
$$\sup_{x\in\mathbb{R}} \left|\hat{F}_n(x) - F_X(x)\right| \underset{n\to\infty}{\longrightarrow} 0\,.$$
Der Vergleich der entsprechenden Quantil-Funktionen wird durch folgendes Ergebnis bestärkt: Falls $\mathbb{E}X < \infty$, dann gilt
$$\sup_{t\in[0,1]} \left|\int_0^t \left(\hat{F}_n^{-1}(y) - F_X^{-1}(y)\right) dy\right| \xrightarrow[n\to\infty]{\text{f.s.}} 0\,.$$
Somit setzt man bei der Verwendung der Quantil-Grafiken voraus, dass der Stichprobenumfang n ausreichend groß ist, um $\hat{F}_n^{-1} \approx F_X^{-1}$ zu gewährleisten.
- Man setzt zusätzlich voraus, dass die Gleichungen
$$\begin{aligned} y &= ax + b\,, \\ y &= F_X^{-1}(t)\,, \\ x &= G^{-1}(t) \end{aligned}$$

für alle t (und nicht nur näherungsweise für $t = k/n$, $k = 1, \ldots, n$) gelten. Daraus folgt, dass $G(x) = t = F_X(y) = F_X(ax + b)$ für alle x, oder $F_X(y) = G\left(\frac{y-b}{a}\right)$ für alle y, weil $x = \frac{y-a}{b}$ ist.

Aus praktischer Sicht ist es besser, Paare $\left(G^{-1}\left(\frac{k}{n+1}\right), X_{(k)}\right)$, $k = 1, \ldots, n$ zu plotten. Dadurch wird vermieden, dass $G^{-1}(n/n) = G^{-1}(1) = \infty$ vorkommt, wie es zum Beispiel im Falle einer Verteilung G der Fall ist, bei der $F(x) < 1$ gilt für alle $x \in \mathbb{R}$. Tatsächlich gilt für $k = n$, dass $\frac{n}{n+1} < 1$ und somit $G^{-1}\left(\frac{k}{n+1}\right) < \infty$.

Beispiel 2.3.1 (Exponential-Verteilung, $G(x) = (1 - e^{-\lambda x}) \cdot \mathbb{I}(x \geq 0)$):
Es gilt $G^{-1} = -1/\lambda \log(1 - y)$, $y \in (0, 1)$. So wird man beim Quantil-Plot Paare

$$\left(-\frac{1}{\lambda} \log\left(1 - \frac{k}{n+1}\right), X_{(k)}\right), \quad k = 1, \ldots, n$$

zeichnen, wobei der Faktor $1/\lambda$ für die Linearität unwesentlich ist und weggelassen werden kann.

Beispiel 2.3.2 (Normalverteilung, $G(x) = \Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-t^2/2}\, dt$, $x \in \mathbb{R}$):
Leider ist die analytische Berechnung von Φ^{-1} mit einer geschlossenen Formel nicht möglich. Aus diesem Grund wird $\Phi^{-1}\left(\frac{k}{n+1}\right)$ numerisch berechnet und in Tabellen oder statistischen Software-Paketen (wie z.B. R) abgelegt. Um die empirische Verteilung der Daten mit der Normalverteilung zu vergleichen, trägt man Punkte mit Koordinaten

$$\left(\Phi^{-1}\left(\frac{k}{n+1}\right), X_{(k)}\right), \quad k = 1, \ldots, n$$

auf der Ebene auf und prüft, ob sie eine Gerade bilden (vgl. Abb. 2.8).

Übungsaufgabe 2.3.1
Entwerfen Sie die Quantil-Grafiken für den Vergleich der empirischen Verteilung mit der Lognormal und der Weibull-Verteilung.

Bemerkung 2.3.1
Falls $\bar{x}_n = 0$ und die Verteilung F_X linkssteil ist, so sind die Quantile von F_X kleiner als die von Φ. Somit ist der Normal-Quantilplot konvex. Falls $\bar{x}_n = 0$ und F_X rechtssteil ist, so wird der Normal-Quantilplot konkav sein.

Beispiel 2.3.3 (Haftpflichtversicherung (Belgien, 1992)):
In Abbildung 2.9 sind Ordnungsstatistiken der Stichprobe von $n = 227$ Schadenhöhen der Industrie-Unfälle in Belgien im Jahr 1992 (Haftpflichtversicherung) gegen Quantile von Exponential-, Pareto-, Standardnormal- und Weibull-Verteilungen geplottet. Im Bereich von Kleinschäden zeigen die Exponential- und Pareto-Verteilungen eine gute Übereinstimmung mit den Daten. Die Verteilung von mittelgroßen Schäden kann am besten durch die Normal- und Weibul-Verteilungen modelliert werden. Für Großschäden erweist sich die Weibull-Verteilung als geeignet.

Beispiel 2.3.4 (Rendite der BMW-Aktie):
In Abbildung 2.10 ist der Quantilplot für Renditen der BMW-Aktie beispielhaft zu sehen.

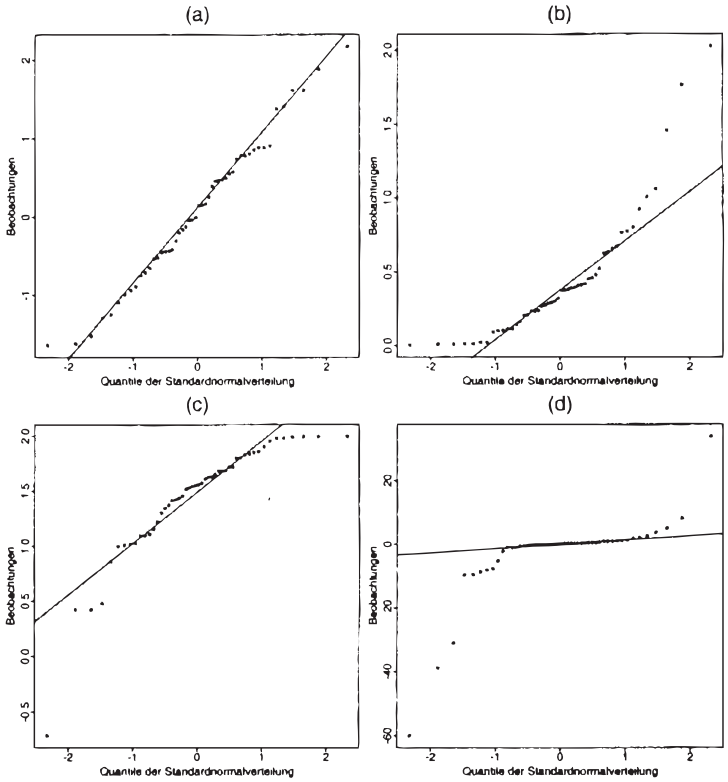


Abb. 2.8: QQ-Plot einer Normalverteilung (a), einer linkssteilen Verteilung (b), einer rechtssteilen Verteilung (c) und einer symmetrischen, aber stark gekrümmten Verteilung (d)

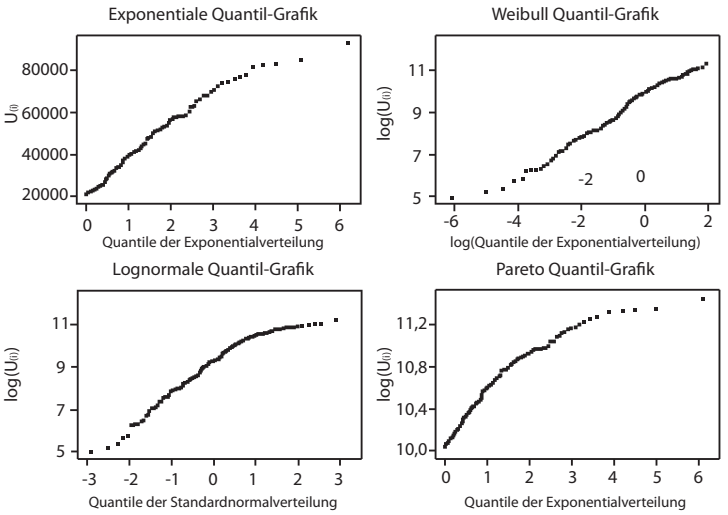


Abb. 2.9: Ordnungsstatistiken einer Stichprobe von Schadenhöhen der Industrie-Unfälle in Belgien im Jahr 1992

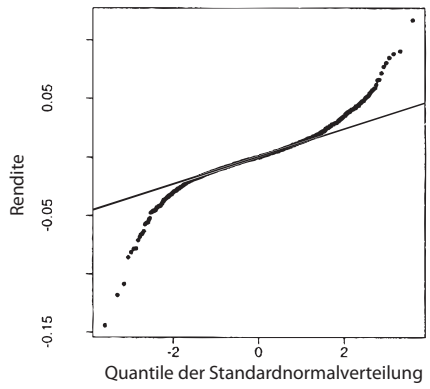


Abb. 2.10: Quantilplot der Rendite der BMW-Aktie

2.4 Dichteschätzung

Sei eine Stichprobe $(x_1, \ldots, x_n)$ von unabhängigen Realisierungen eines absolut stetig verteilten Merkmals X mit Dichte f_X gegeben. Mit Hilfe der in Abschnitt 2.1.1 eingeführten Histogramme lässt sich f_X grafisch durch eine Treppenfunktion $\hat{f}_X$ darstellen. Dabei gibt es zwei entscheidende Nachteile der Histogrammdarstellung:

1. Willkür in der Wahl der Klasseneinteilung $[c_{i-1}, c_i]$,

2. Eine (möglicherweise) stetige Funktion f_X wird durch eine Treppenfunktion $\hat{f}_X$ ersetzt.

In diesem Abschnitt werden wir versuchen, diese Nachteile zu beseitigen, indem wir eine Klasse von Kerndichtenschätzern einführen, die (je nach Wahl des Kerns) auch zu stetigen Schätzern $\hat{f}_X$ führen.

Definition 2.4.1
Der Kern $K(x)$ wird definiert als eine nicht-negative messbare Funktion auf $\mathbb{R}$ mit der Eigenschaft $\int_{\mathbb{R}} K(x)\,dx = 1$.

Definition 2.4.2
Der *Kerndichteschätzer* der Dichte f_X aus den Daten $(x_1, \ldots, x_n)$ mit Kernfunktion $K(x)$ ist gegeben durch

$$\hat{f}_X(x) = \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{x - x_i}{h}\right), \quad x \in \mathbb{R},$$

wobei $h > 0$ die sogenannte *Bandbreite* ist.

Beispiele für Kerne:

1. *Rechteckskern:*

$$K(x) = 1/2 \cdot \mathbb{I}(x \in [-1, 1)).$$

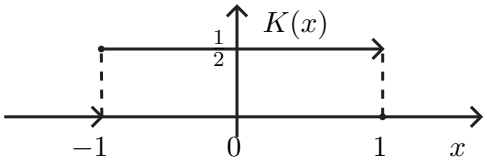
Dabei ist

$$\frac{1}{h} K\left(\frac{x - x_i}{h}\right) = \begin{cases} 1/(2h), & x_i - h \leq x < x_i + h, \\ 0, & \text{sonst}, \end{cases}$$

und somit

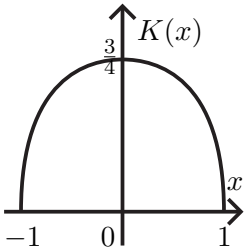
$$\hat{f}_X(x) = \frac{1}{nh} \sum_{i=1}^{k} K\left(\frac{x - x_i}{h}\right) = \frac{\#\{x_i \in [x-h, x+h)\}}{2nh}\,,$$

das auch *gleitendes Histogramm* genannt wird. Dieser Dichteschätzer ist (noch) nicht stetig, was durch die (besonders einfache rechteckige unstetige) Form des Kerns erklärt wird.



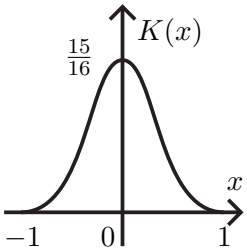
2. *Epanechnikov-Kern:*

$$K(x) = \begin{cases} 3/4(1-x^2)\,, & x \in [-1,1) \\ 0\,, & \text{sonst.} \end{cases}$$



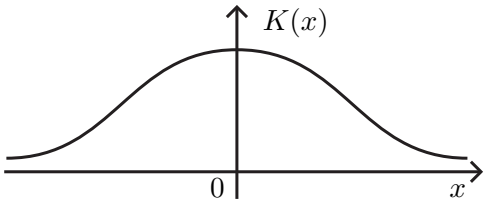
3. *Bisquare-Kern:*

$$K(x) = \frac{15}{16}\Big((1-x^2)^2 \cdot \mathbb{I}(x \in [-1,1))\Big)\,.$$



4. *Gauss-Kern:*

$$K(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}\,, \quad x \in \mathbb{R}\,.$$



Dabei ist die Wahl der Bandbreite h entscheidend für die Qualität der Schätzung. Je größer $h > 0$, desto glatter wird $\hat{f}_X$ sein und desto mehr „Details“ werden „herausgemittelt“. Für kleinere h wird $\hat{f}_X$ rauer. Dabei können aber auch Details auftreten, die rein stochastischer Natur sind und keine Gesetzmäßigkeiten zeigen. Mit der adäquaten Wahl von h beschäftigen sich viele wissenschaftliche Arbeiten, die empirische Faustregeln, aber auch kompliziertere Optimierungsmethoden dafür vorschlagen. Insgesamt ist das Problem der optimalen Dichteschätzung in der Statistik immer noch offen.

2.5 Beschreibung und Exploration von bivariaten Datensätzen

Im Gegensatz zu der Datenlage in den Abschnitten 2.1 bis 2.4 betrachten wir im Folgenden Datensätze bestehend aus 2 Stichproben $(x_1, \ldots, x_n)$ und $(y_1, \ldots, y_n)$, die als Realisierungen von stochastischen Stichproben $(X_1, \ldots, X_n)$ und $(Y_1, \ldots, Y_n)$ aufgefasst werden, wobei $X_1, \ldots, X_n$ unabhängige identisch verteilte Zufallsvariablen mit $X_i \stackrel{d}{=} X \sim F_X$, $Y_1, \ldots, Y_n$ unabhängige identisch verteilte Zufallsvariablen mit $Y_i \stackrel{d}{=} Y \sim F_Y$ sind. Wir betrachten hier ausschließlich quantitative Merkmale X und Y. Es wird ein Zusammenhang zwischen X und Y vermutet, der an Hand von (konkreten) Stichproben $(x_1, \ldots, x_n)$ und $(y_1, \ldots, y_n)$ näher untersucht werden soll. Mit anderen Worten, wir interessieren uns für die Eigenschaften der bivariaten Verteilung $F_{X,Y}(x, y) = P(X \leq x, Y \leq y)$ des Zufallsvektors (X, Y).

2.5.1 Grafische Darstellung von bivariaten Datensätzen

Um die Verteilung von $(x_1, \ldots, x_n)$ und $(y_1, \ldots, y_n)$ zu visualisieren, betrachten wir drei Möglichkeiten:

1. *Streudiagramme*
2. *Zweidimensionale Histogramme*
3. *Kerndichteschätzer* (im Falle eines absolut stetig verteilten Zufallsvektors (X, Y))

1. *Streudiagramme* sind die erste sehr einfache und intuitive Visualisierungsmöglichkeit von bivariaten Daten. Um ein Streudiagramm zu erstellen, plottet man die „Punktwolke“ $(x_i, y_i)_{i=1,\ldots,n}$ auf einer Koordinatenebene im $\mathbb{R}^2$. Dabei zeigt die Form der Punktwolke, ob ein linearer ($y = ax + b$) bzw. polynomialer ($y = P_d(x)$) Zusammenhang in den Daten zu erwarten ist. Später werden solche Zusammenhänge im Rahmen der Regressionstheorie untersucht (vgl. Abschnitt 2.5.3 für die einfache lineare Regression).

2. *Zweidimensionale Histogramme* dienen der Darstellung der bivariaten Zähldichte $p(x, y)$ des Zufallsvektors (X, Y), falls er diskret verteilt ist, bzw. seiner Dichte $f(x, y)$ im Falle einer absolut stetigen Verteilung von (X, Y) aus den Daten $(x_1, \ldots, x_n)$ und $(y_1, \ldots, y_n)$. Dabei teilt man den Wertebereich von X in Intervalle

$$[c_{i-1}, c_i)\,, \quad i = 1 \ldots k\,, \quad -\infty = c_0 < c_1 < \ldots < c_k = +\infty$$

und den Wertebereich von Y in Intervalle

$$[e_{i-1}, e_i)\,, \quad i = 1 \ldots m\,, \quad -\infty = e_0 < e_1 < \ldots < e_m = +\infty\,.$$

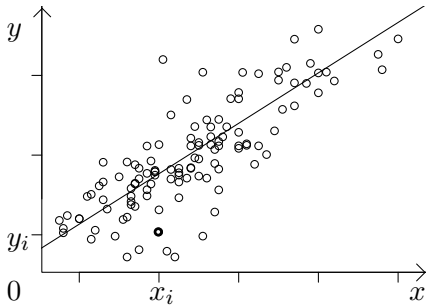


Abb. 2.11: Punktwolke

Bezeichnen wir

$$h_{ij} = \#\{(x_k, y_l),\, k, l = 1, \ldots, n :\, x_k \in [c_{i-1}, c_i),\, y_l \in [e_{j-1}, e_j)\}$$

als die absolute Häufigkeit von (X, Y) in $[c_{i-1}, c_i) \times [e_{j-1}, e_j)$, $f_{ij} = h_{ij}/n$ als die relative Häufigkeit. Das zweidimensionale Histogramm setzt sich aus den Säulen mit Grundriss $[c_{i-1}, c_i) \times [e_{j-1}, e_j)$ und Höhe

$$\frac{h_{ij}}{(c_i - c_{i-1})(e_j - e_{j-1})}$$

für das Histogramm absoluter Häufigkeiten bzw.

$$\frac{f_{ij}}{(c_i - c_{i-1})(e_j - e_{j-1})}$$

für das Histogramm relativer Häufigkeiten zusammen, damit das Volumen dieser Säulen h_{ij} bzw. f_{ij} ist. Dabei hat solch ein Histogramm dieselben Vor- bzw. Nachteile wie ein ein-

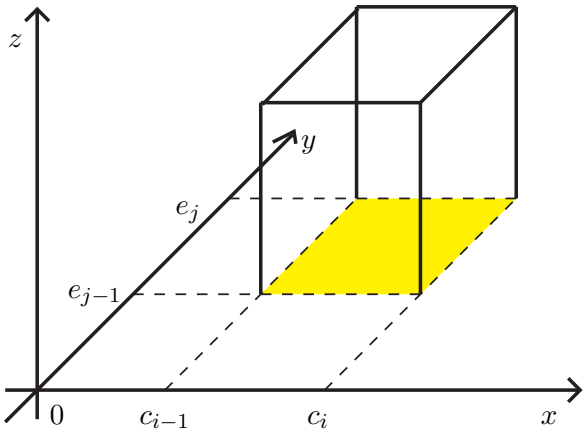


Abb. 2.12: Zweidimensionales Histogramm

dimensionales, wenn es um die grafische Darstellung einer bivariaten Dichte $f(x, y)$ geht. Deshalb benutzt man oft Kerndichteschätzer, um eine glatte Darstellung zu bekommen.

3. *Zweidimensionale Kerndichteschätzer* haben die Form

$$\hat{f}(x,y) = \frac{1}{nh_1h_2}\sum_{i=1}^{n} K\left(\frac{x-x_i}{h_1}\right) K\left(\frac{y-y_i}{h_2}\right)$$

für die Bandbreiten h_1, $h_2 > 0$, die Glättungsparameter sind. Dabei ist $K(\cdot)$ eine Kernfunktion (vgl. Abschnitt 2.4). Seine Eigenschaften übertragen sich aus dem eindimensionalen Fall.

2.5.2 Zusammenhangsmaße

Jetzt wird uns die Frage beschäftigen, in welchem Maße die Merkmale X und Y voneinander abhängig sind. Um die $\mathrm{Cov}(X,Y) = \mathbb{E}(X-\mathbb{E}X)(Y-\mathbb{E}Y)$ aus den Daten zu schätzen, setzt man die sogenannte *empirische Kovarianz*

$$S_{xy}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x}_n)(y_i-\bar{y}_n)$$

ein. Dabei ist S_{xy}^2 jedoch von den Skalen von X und Y abhängig.

1. Um eine skaleninvariantes Zusammenhangsmaß zu bekommen, betrachtet man die empirische Variante des Korrelationskoeffizienten

$$\varrho(X,Y) = \frac{\mathrm{Cov}(X,Y)}{\sqrt{\mathrm{Var}\,X}\cdot\sqrt{\mathrm{Var}\,Y}}\,,$$

den sogenannten *Bravais-Pearson-Korrelationskoeffizienten*

$$\varrho_{xy} = \frac{S_{xy}^2}{\sqrt{S_{xx}^2\cdot S_{yy}^2}}\,,$$

wobei

$$S_{xx}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x}_n)^2\,, \qquad S_{yy}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i-\bar{y}_n)^2$$

die Stichprobenvarianzen der Stichproben $(x_1,\ldots,x_n)$ und $(y_1,\ldots,y_n)$ sind. Dabei erbt ϱ_{xy} alle Eigenschaften des Korrelationskoeffizienten $\varrho(X,Y)$:

a) $|\varrho_{xy}| \le 1$

b) $\varrho_{xy} = \pm 1$, falls ein linearer Zusammenhang in den Daten $(x_i,y_i)_{i=1,\ldots,n}$ vorliegt, d.h. alle Punkte (x_i,y_i), $i=1,\ldots,n$ liegen auf einer Gerade mit positivem (bei $\varrho_{xy}=1$) bzw. negativem (bei $\varrho_{xy}=-1$) Anstieg.

c) Wenn $|\varrho_{xy}|$ klein ist ($\varrho_{xy}\approx 0$), so sind die Datensätze unkorreliert. Dabei wird oft folgende grobe Einteilung vorgenommen:

Merkmale X und Y sind

- *„schwach korreliert“*, falls $|\varrho_{xy}| < 0.5$,
- *„stark korreliert“*, falls $|\varrho_{xy}| \ge 0.8$.

Ansonsten liegt ein mittlerer Zusammenhang zwischen X und Y vor.

Lemma 2.5.1
Für ϱ_{xy} gilt die alternative rechengünstige Darstellung

$$\varrho_{xy} = \frac{\sum_{i=1}^n x_i y_i - n\bar{x}_n\bar{y}_n}{\sqrt{\left(\sum_{i=1}^n x_i^2 - n\bar{x}_n^2\right)\left(\sum_{i=1}^n y_i^2 - n\bar{y}_n^2\right)}}\,. \tag{2.5.1}$$

Beweis Man muss lediglich zeigen, dass

$$\sum_{i=1}^n (x_i - \bar{x}_n)(y_i - \bar{y}_n) = \sum_{i=1}^n x_i y_i - n\bar{x}_n\bar{y}_n\,.$$

Alles andere folgt daraus für $x_i = y_i\,, \quad i = 1, \ldots, n$. Es gilt

$$\begin{aligned}\sum_{i=1}^n (x_i - \bar{x}_n)(y_i - \bar{y}_n) &= \sum_{i=1}^n x_i y_i - \bar{x}_n \sum_{i=1}^n y_i - \bar{y}_n \sum_{i=1}^n x_i + n\bar{x}_n\bar{y}_n \\ &= \sum_{i=1}^n x_i y_i - n\bar{x}_n\bar{y}_n - n\bar{y}_n\bar{x}_n + n\bar{x}_n\bar{y}_n = \sum_{i=1}^n x_i y_i - n\bar{x}_n\bar{y}_n\end{aligned}$$

□

Falls die vorliegenden Daten $(x_1, \ldots, x_n)$ und $(y_1, \ldots, y_n)$ nur 2 Ausprägungen zeigen und somit *binär* kodiert werden können, d.h. x_i, $y_i \in \{0, 1\}$, dann gilt

$$\varrho_{xy} = \frac{h_{11}h_{22} - h_{12}h_{21}}{\sqrt{h_{1\cdot} \cdot h_{2\cdot} \cdot h_{\cdot 2} \cdot h_{\cdot 2}}} = \varphi$$

(der sogenannte Phi-Koeffizient), wobei

$$\begin{aligned}
h_{11} &= \#\{(x_i, y_i): \quad x_i = y_i = 0\} \\
h_{22} &= \#\{(x_i, y_i): \quad x_i = y_i = 1\} \\
h_{12} &= \#\{(x_i, y_i): \quad x_i = 0,\ y_i = 1\} \\
h_{21} &= \#\{(x_i, y_i): \quad x_i = 1,\ y_i = 0\} \\
h_{1\cdot} &= h_{11} + h_{12} \\
h_{\cdot 1} &= h_{11} + h_{21} \\
h_{2\cdot} &= h_{22} + h_{21} \\
h_{\cdot 2} &= h_{22} + h_{12}
\end{aligned}$$

Übungsaufgabe 2.5.1
Zeigen Sie diese Darstellungsform!

2. *Spearmans Korrelationskoeffizient*
Einen alternativen Korrelationskoeffizienten erhält man, wenn man die Stichprobenwerte x_i bzw. y_i in ϱ_{xy} durch ihre *Ränge* $\mathrm{rg}(x_i)$ bzw. $\mathrm{rg}(y_i)$ ersetzt, die als Position dieser Werte in den ansteigend geordneten Stichproben zu verstehen sind:
$\mathrm{rg}(x_i) = j$, falls $x_i = x_{(j)}$ für ein $j \in \{1, \ldots, n\}$, $\forall i = 1, \ldots, n$. Es bedeutet, dass $\mathrm{rg}(x_{(i)}) = i\ \forall i = 1, \ldots, n$, falls $x_i \neq x_j$ für $i \neq j$.

Falls die Stichprobe $(x_1, \ldots, x_n)$ k identische Werte x_i (die sogenannten *Bindungen*) enthält, so wird diesen Werten der sogenannte Durchschnittsrang $\mathrm{rg}(x_i)$ zugewiesen, der als arithmetisches Mittel der k in Frage kommenden Ränge errechnet wird. Zum Beispiel findet folgende Zuordnung statt:

x_i	$(3, 1, 7, 5, 3, 3)$
$\mathrm{rg}(x_i)$	$(a, 1, 6, 5, a, a)$

wobei der Durchschnittsrang a von Stichprobeneintrag 3 gleich $a = 1/3(2 + 3 + 4) = 3$ ist.
Somit wird der sogenannte *Spearmans Korrelationskoeffizient* (Rangkorrelationskoeffizient) der Stichproben

$$(x_1, \ldots, x_n) \quad \text{und} \quad (y_1, \ldots, y_n)$$

als der *Bravais-Pearson-Koeffizient* der Stichproben ihrer Ränge

$$\big(\mathrm{rg}(x_1), \ldots, \mathrm{rg}(x_n)\big) \quad \text{und} \quad \big(\mathrm{rg}(y_1), \ldots, \mathrm{rg}(y_n)\big)$$

definiert:

$$\varrho_{sp} = \frac{\sum_{i=1}^n \big(\mathrm{rg}(x_i) - \overline{\mathrm{rg}}_x\big)\big(\mathrm{rg}(y_i) - \overline{\mathrm{rg}}_y\big)}{\sqrt{\sum_{i=1}^n \big(\mathrm{rg}(x_i) - \overline{\mathrm{rg}}_x\big)^2 \sum_{i=1}^n \big(\mathrm{rg}(y_i) - \overline{\mathrm{rg}}_y\big)^2}}\,,$$

wobei

$$\begin{aligned}
\overline{\mathrm{rg}}_x &= \frac{1}{n}\sum_{i=1}^n \mathrm{rg}(x_i) = \frac{1}{n}\sum_{i=1}^n \mathrm{rg}(x_{(i)}) = \frac{1}{n}\sum_{i=1}^n i = \frac{n(n+1)}{2n} = \frac{n+1}{2}\,,\\
\overline{\mathrm{rg}}_y &= \frac{1}{n}\sum_{i=1}^n \mathrm{rg}(y_i) = \frac{n+1}{2}\,.
\end{aligned}$$

Dieser Koeffizient misst monotone Zusammenhänge in den Daten. Aus den Eigenschaften der Bravais-Pearson-Koeffizienten folgt $|\varrho_{sp}| \leq 1$. Betrachten wir die Fälle $\varrho_{sp} = \pm 1$ gesondert:

- $\varrho_{sp} = 1$ bedeutet, dass die Punkte $(\mathrm{rg}(x_i), \mathrm{rg}(y_i))$, $i = 1, \ldots, n$ auf einer Geraden mit positiver Steigung liegen. Da aber $\mathrm{rg}(x_i), \mathrm{rg}(y_i) \in \mathbb{N}$, kann diese Steigung nur 1 sein. Es bedeutet, dass dem kleinsten Wert in der Stichprobe $(x_1, \ldots, x_n)$ der kleinste Wert in $(y_1, \ldots, y_n)$ entspricht, usw., d.h., für wachsende x_i wachsen auch die y_i streng monoton: $x_i < x_j \Longrightarrow y_i < y_j \quad \forall i \neq j$.
- Analog gilt dann für $\varrho_{sp} = -1$, dass $x_i < x_j \Longrightarrow y_i > y_j \quad \forall i \neq j$.

Dies kann folgendermaßen zusammengefaßt werden:

- $\varrho_{sp} > 0$: gleichsinniger monotoner Zusammenhang (x_i groß $\Longleftrightarrow$ y_i groß)
- $\varrho_{sp} < 0$: gegensinniger monotoner Zusammenhang (x_i groß $\Longleftrightarrow$ y_i klein)
- $\varrho_{sp} \approx 0$: kein monotoner Zusammenhang.

Da der Spearmans Korrelationskoeffizient nur Ränge von x_i und y_i betrachtet, eignet er sich auch für ordinale (und nicht nur quantitative) Daten.

Lemma 2.5.2
Falls die Stichproben $(x_1, \ldots, x_n)$ und $(y_1, \ldots, y_n)$ keine Bindung enthalten ($x_i \neq x_j$, $y_i \neq y_j$ $\forall i \neq j$), dann gilt

$$\varrho_{sp} = 1 - \frac{6}{(n^2-1)n} \sum_{i=1}^{n} d_i^2 \,,$$

wobei $d_i = \mathrm{rg}(x_i) - \mathrm{rg}(y_i) \quad \forall i = 1, \ldots, n$.

Beweis Als Übungsaufgabe. □

Satz 2.5.1 (Invarianzeigenschaften):

1. Wenn die Merkmale X und Y linear transformiert werden:

$$\begin{aligned} f(X) &= a_x X + b_x \,, \quad \forall a_x \neq 0,\, b_x \in \mathbb{R} \\ g(Y) &= a_y Y + b_y \,, \quad \forall a_y \neq 0,\, b_y \in \mathbb{R} \end{aligned}$$

   dann gilt $\varrho_{f(x)g(y)} = \mathrm{sgn}(a_x a_y) \cdot \varrho_{xy}$.

2. Falls Funktionen $f : \mathbb{R} \to \mathbb{R}$ und $g : \mathbb{R} \to \mathbb{R}$ beide monoton wachsend oder beide monoton fallend sind, dann gilt

$$\varrho_{sp}(f(x), g(y)) = \varrho_{sp}(x, y) \,.$$

   Falls f monoton wachsend und g monoton fallend (oder umgekehrt) sind, dann gilt $\varrho_{sp}(f(x), g(y)) = -\varrho_{sp}(x, y)$.

Beweis Beweisen wir nur 1), weil 2) offensichtlich ist.

1.

$$\begin{aligned} \varrho_{f(x)g(y)} &= \frac{\sum_{i=1}^n \big((a_x x_i + b_x) - (a_x \bar{x}_n + b_x)\big)\big((a_y y_i + b_y) - (a_y \bar{y}_n + b_y)\big)}{\sqrt{a_x^2 \sum_{i=1}^n (x_i - \bar{x}_n)^2 a_y^2 \sum_{i=1}^n (y_i - \bar{y}_n)^2}} \\ &= \frac{a_x a_y}{|a_x||a_y|} \cdot \frac{\sum_{i=1}^n (x_i - \bar{x}_n)(y_i - \bar{y}_n)}{\sqrt{\sum_{i=1}^n (x_i - \bar{x}_n)^2 \sum_{i=1}^n (y_i - \bar{y}_n)^2}} = \mathrm{sgn}(a_x a_y) \cdot \varrho_{xy} \,. \end{aligned}$$

□

Bemerkung 2.5.1

1. Da lineare Transformationen monoton sind, gilt Aussage 1) auch für Spearmans Korrelationskoeffizienten ϱ_{sp}.

2. Der Koeffizient ϱ_{xy} erfasst lineare Zusammenhänge, während ϱ_{sp} monotone Zusammenhänge aufspürt.

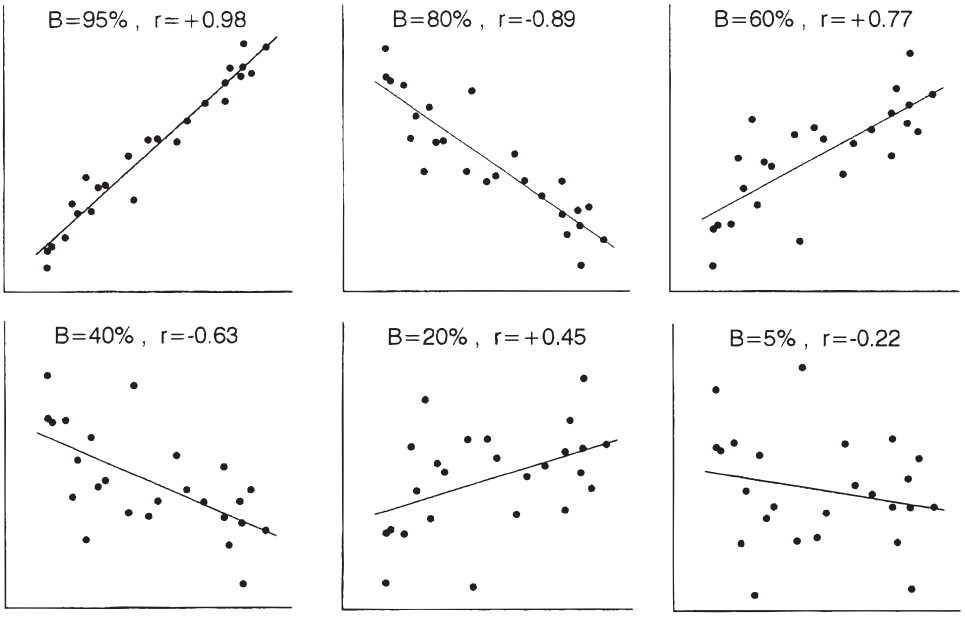


Abb. 2.13: Vergleich verschiedenwertiger Bestimmtheitsmaße. Es sind Regressionsgerade, Bestimmtheitsmaß B und Korrelationskoeffizient r verschiedener (fiktiver) Punktwolken vom Umfang $n = 25$ dargestellt. Die Beschriftung der Achsen ist weggelassen, weil sie hier ohne Bedeutung ist.

2.5.3 Einfache lineare Regression

Wenn man den Zusammenhang von Merkmalen X und Y mit Hilfe von Streudiagrammen visualisiert, wird oft ein linearer Trend erkennbar, obwohl der Bravais-Pearson-Korrelationskoeffizient einen Wert kleiner als 1 liefert, z.B. $\varrho_{xy} \approx 0,6$ (vgl. Abb. 2.13). Dies ist der Fall, weil die Datenpunkte (x_i, y_i), $i = 1, \ldots, n$ oft um eine Gerade streuen und nicht exakt auf einer Geraden liegen. Um solche Situationen stochastisch modellieren zu können, nimmt man den Zusammenhang der Form

$$Y = f(X) + \varepsilon$$

an, wobei ε die sogenannte Störgröße ist, die auf mehrere Ursachen wie z.B. Beobachtungsfehler (Messfehler, Berechnungsfehler, usw.) zurückzuführen sein kann. Dabei nennt man die Zufallsvariable Y *Zielgröße* oder *Regressand*, die Zufallsvariable X *Einflussfaktor*, *Regressor* oder *Ausgangsvariable*. Der Zusammenhang $Y = f(X) + \varepsilon$ wird *Regression* genannt, wobei man oft über ε voraussetzt, dass $\mathbb{E}\varepsilon = 0$ (kein systematischer Beobachtungsfehler). Wenn $f(x) = \alpha + \beta x$ eine lineare Funktion ist, so spricht man von der *einfachen linearen Regression*. Es sind aber durchaus andere Arten der Zusammenhänge denkbar, wie z.B.

$$f(x) = \sum_{i=0}^{n} \alpha_i x^i$$

(*polynomiale Regression*), usw. Beispiele für mögliche Ausgangs- bzw. Zielgrößen sind in Tabelle 2.1 zusammengefasst, einige Beispiele in Abbildung 2.14. Auf Modellebene ist damit

X	Y
Geschwindigkeit	Länge des Bremswegs
Körpergröße des Vaters	Körpergröße des Sohnes
Produktionsfaktor	Qualität des Produktes
Spraydosen-Verbrauch	Ozongehalt der Atmosphäre
Noten im Vordiplom	Noten im Hauptdiplom

Tab. 2.1: Beispiele möglicher Ausgangs- und Zielgrößen

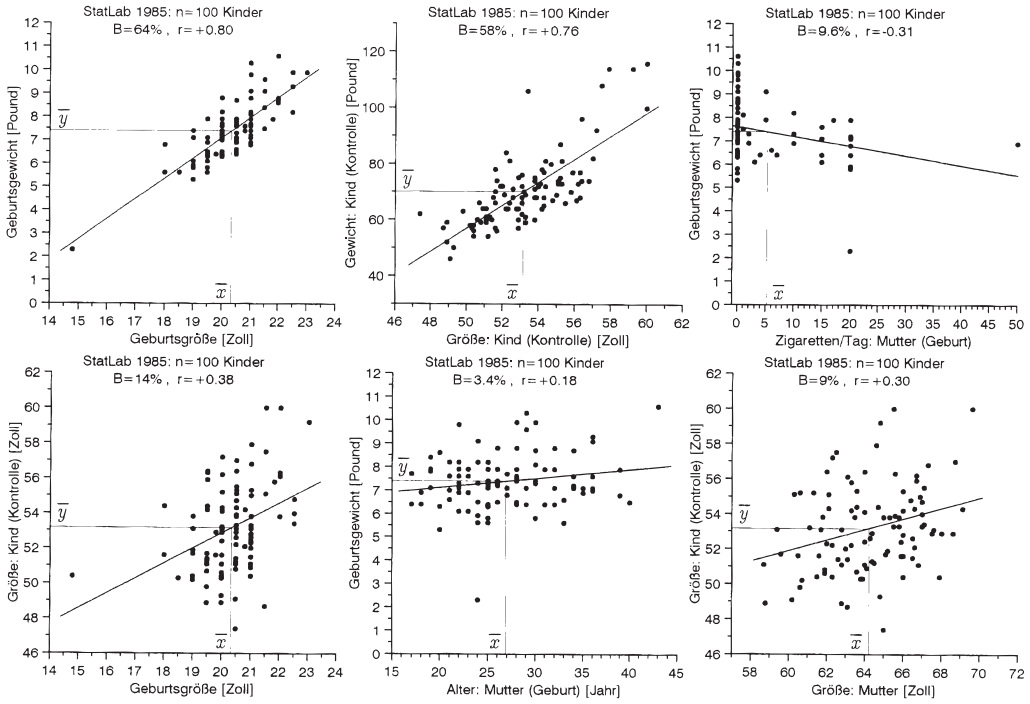


Abb. 2.14: Punktwolken verschiedener Merkmale der StatLab-Auswahl 1985 mit Regressionsgerade, Bestimmtheitsmaß B und Korrelationskoeffizient r.

folgende Fragestellung gegeben: Es gebe Zufallsstichproben von Ziel- bzw. Ausgangsvariablen $(Y_1, \ldots, Y_n)$ und $(X_1, \ldots, X_n)$, zwischen denen ein verrauschter linearer Zusammenhang $Y_i = \alpha + \beta X_i + \varepsilon_i$ besteht, wobei ε_i Störgrößen sind, die nicht direkt beobachtbar und uns somit unbekannt sind. Meistens nimmt man an, dass $\mathbb{E}\,\varepsilon_i = 0 \quad \forall\, i = 1, \ldots, n$ und $\mathrm{Cov}(\varepsilon_i, \varepsilon_j) = \sigma^2 \delta_{ij}$, d.h. $\varepsilon_1 \ldots \varepsilon_n$ sind unkorreliert mit $\mathrm{Var}\,\varepsilon_i = \sigma^2$. Wenn wir über die Eigenschaften der Schätzer für α, β und σ^2 reden, gehen wir davon aus, dass die X-Werte nicht zufällig sind, also $X_i = x_i \quad \forall i = 1, \ldots, n$. Wenn man von einer konkreten Stichprobe $(y_1, \ldots, y_n)$ für $(Y_1, \ldots, Y_n)$ ausgeht, so sollen anhand von den Stichproben $(x_1, \ldots, x_n)$ und $(y_1, \ldots, y_n)$ Regressionsparameter α (*Regressionskonstante*) und β (*Regressionskoeffizient*) sowie *Regressionsvarianz* σ^2 geschätzt werden. Dabei verwendet man die sogenannte *Methode der kleinsten Quadrate*, die den mittleren quadratischen Fehler von den Datenpunkten $(x_i, y_i)_{i=1,\ldots,n}$ des Streudiagramms

zur *Regressionsgeraden* $y = \alpha + \beta x$ minimiert:

$$(\alpha, \beta) = \arg\min_{\alpha,\beta\in\mathbb{R}} e(\alpha, \beta) \quad \text{mit} \quad e(\alpha, \beta) = \frac{1}{n}\sum_{i=1}^{n}(y_i - \alpha - \beta x_i)^2 \,.$$

Da die Darstellung $y_i = \alpha + \beta x_i + \varepsilon_i$ gilt, kann man $e(\alpha, \beta) = 1/n \sum_{i=1}^{n} \varepsilon_i^2$ schreiben. Es

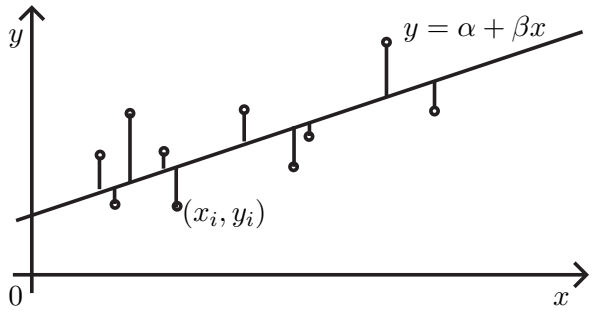


Abb. 2.15: Methode kleinster Quadrate

ist der vertikale mittlere quadratische Abstand von den Datenpunkten (x_i, y_i) zur Geraden $y = \alpha + \beta x$ (vgl. Abb. 2.15). Das Minimierungsproblem $e(\alpha, \beta) \mapsto \min$ löst man durch das zweifache Differenzieren von $e(\alpha, \beta)$. Somit erhält man $\hat{\alpha} = \bar{y}_n - \hat{\beta}\bar{x}_n$, wobei

$$\hat{\beta} = \frac{S_{xy}^2}{S_{xx}^2}\,, \quad \bar{x}_n = \frac{1}{n}\sum_{i=1}^{n} x_i\,, \quad \bar{y}_n = \frac{1}{n}\sum_{i=1}^{n} y_i\,,$$
$$S_{xy}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x}_n)(y_i - \bar{y}_n)\,, \quad S_{xx}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x}_n)^2\,.$$

Übungsaufgabe 2.5.2
Leiten Sie die Schätzer $\hat{\alpha}$ und $\hat{\beta}$ selbstständig her.

Die Varianz σ^2 schätzt man durch $\hat{\sigma}^2 = \frac{1}{n-2}\sum_{i=1}^{n}\hat{\varepsilon}_i^2$, wobei $\hat{\varepsilon}_i = y_i - \hat{\alpha} - \hat{\beta}x_i\,, \quad i = 1, \ldots, n$ die sogenannten *Residuen* sind. Die Gründe, warum $\hat{\sigma}^2$ diese Gestalt hat, können an dieser Stelle noch nicht angegeben werden, weil wir noch nicht die Maximum-Likelihood-Methode kennen. Zu gegebener Zeit (in der Vorlesung Statistik II) wird jedoch klar, dass diese Art der Schätzung sehr natürlich ist.

Bemerkung 2.5.2
Die angegebenen Schätzer für α und β sind nicht symmetrisch bzgl. Variablen x_i und y_i. Wenn man also die *horizontalen* Abstände (statt vertikaler) zur Bildung des mittleren quadratischen Fehlers nimmt (was dem Rollentausch $x \leftrightarrow y$ entspricht), so bekommt man andere Schätzer für α und β, die mit $\hat{\alpha}$ und $\hat{\beta}$ nicht übereinstimmen müssen:

$$d_i = y_i - \alpha - \beta x_i \mapsto d_i' = x_i - \frac{(y_i - \alpha)}{\beta}\,.$$

Ein Ausweg aus dieser asymmetrischen Situation wäre die orthogonalen Abstände o_i von

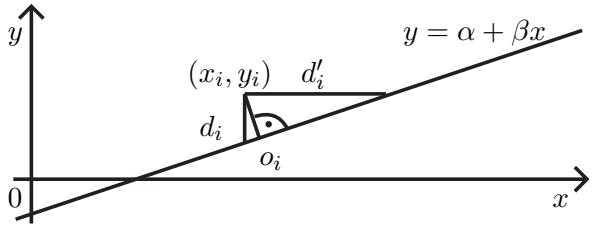


Abb. 2.16: Orthogonale Abstände

Kind i	1	2	3	4	5	6	7	8	9
Fernsehzeit x_i	0,3	2,2	0,5	0,7	1,0	1,8	3,0	0,2	2,3
Tiefschlafdauer y_i	5,8	4,4	6,5	5,8	5,6	5,0	4,8	6,0	6,1

Tab. 2.2: Daten von Fernsehzeit und korrespondierender Tiefschlafdauer

(x_i, y_i) zur Geraden $y = \alpha + \beta x$ zu betrachten (vgl. Abb. 2.16). Diese Art der Regression, die „errors-in-variables regression“ genannt wird, hat aber eine Reihe von Eigenschaften, die sie zur Prognose von Zielvariablen y_i durch die Ausgangsvariablen x_i unbrauchbar machen. Sie sollte zum Beispiel nur dann verwendet werden, wenn die Standardabweichungen für X und Y etwa gleich groß sind.

Beispiel 2.5.1
Ein Kinderpsychologe vermutet, dass sich häufiges Fernsehen negativ auf das Schlafverhalten von Kindern auswirkt. Um diese Hypothese zu überprüfen, wurden 9 Kinder im gleichen Alter befragt, wie lange sie pro Tag fernsehen dürfen, und zusätzlich die Dauer ihrer Tiefschlafphase gemessen. So ergibt sich der Datensatz in Tabelle 2.2 und die Regressionsgerade aus Abbildung 2.17.

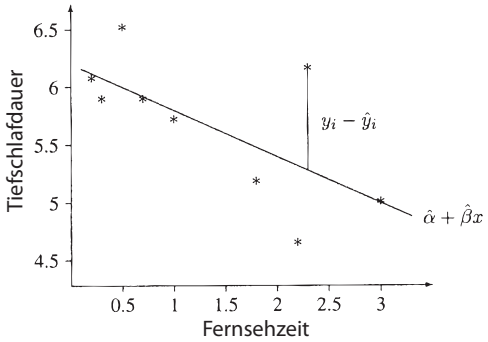


Abb. 2.17: Streudiagramm und Ausgleichsgerade zur Regression der Dauer des Tiefschlafs auf die Fernsehzeit

Es ergibt sich für die oben genannten Stichproben $(x_1, \ldots, x_9)$ und $(y_1, \ldots, y_9)$

$$\bar{x}_9 = 1,33\,, \quad \bar{y}_9 = 5,56\,, \quad \hat{\beta} = -0,45\,, \quad \hat{\alpha} = 6,16\,.$$

Somit ist

$$y = 6,16 - 0,45x$$

die Regressionsgerade, die eine negative Steigung hat, was die Vermutung des Kinderpsychologen bestätigt. Außerdem ist es mit Hilfe dieser Geraden möglich, Prognosen für die Dauer des Tiefschlafs für vorgegebene Fernsehzeiten anzugeben. So wäre z.B. für die Fernsehzeit von 1 Stunde der Tiefschlaf von $6,16 - 0,45 \cdot 1 = 5,71$ Stunden plausibel.

Bemerkung 2.5.3 (Eigenschaften der Regressionsgerade):

1. Es gilt $\operatorname{sgn}(\hat{\beta}) = \operatorname{sgn}(r_{xy})$, was aus $\hat{\beta} = s_{xy}^2/s_{xx}^2$ folgt. Dies bedeutet (falls $s_{yy}^2 > 0$):
   a) Die Regressionsgerade $y = \hat{\alpha} + \hat{\beta}x$ steigt an, falls die Stichproben $(x_1, \ldots, x_n)$ und $(y_1, \ldots, y_n)$ positiv korreliert sind.
   b) Die Regressionsgerade fällt ab, falls sie negativ korreliert sind.
   c) Die Regressionsgerade ist konstant, falls die Stichproben unkorreliert sind.

   Falls $s_{yy}^2 = 0$, dann ist die Regressionsgerade konstant ($y = \bar{y}_n$).
2. Die Regressionsgerade $y = \hat{\alpha}+\hat{\beta}x$ verläuft immer durch den Punkt $(\bar{x}_n, \bar{y}_n)$: $\hat{\alpha}+\hat{\beta}\bar{x}_n = \bar{y}_n$.
3. Seien $\hat{y}_i = \hat{\alpha} + \hat{\beta}x_i$, $i = 1, \ldots, n$. Dann gilt

   $$\overline{\hat{y}_n} = \frac{1}{n}\sum_{i=1}^{n} \hat{y}_i = \bar{y}_n \quad \text{und somit} \quad \sum_{i=1}^{n}(\underbrace{y_i - \hat{y}_i}_{\hat{\varepsilon}_i}) = 0\,.$$

   Dabei sind $\hat{\varepsilon}_i$ die schon vorher eingeführten Residuen. Mit ihrer Hilfe ist es möglich, die Güte der Regressionsprognose zu beurteilen.

Residualanalyse und Bestimmtheitsmaß

Definition 2.5.1
Der relative Anteil der Streuungsreduktion an der Gesamtstreuung S_{yy}^2 heißt das *Bestimmtheitsmaß* der Regressionsgeraden:

$$R^2 = \frac{S_{yy}^2 - \frac{1}{n-1}\sum_{i=1}^{n}\hat{\varepsilon}_i^2}{S_{yy}^2} = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y}_n)^2}\,.$$

Es ist nur im Fall $S_{xx}^2 > 0$, $S_{yy}^2 > 0$ definiert, d.h., wenn nicht alle Werte x_i bzw. y_i übereinstimmen.

Warum R^2 in dieser Form eingeführt wird, zeigt folgende Überlegung, die *Streuungszerlegung* genannt wird:

Lemma 2.5.3
Die Gesamtstreuung („sum of squares total“) $\mathrm{SQT} = (n-1)S_{yy}^2 = \sum_{i=1}^{n}(y_i-\bar{y}_n)^2$ lässt sich in die Summe der sogenannten erklärten Streuung „sum of squares explained“ $\mathrm{SQE} = \sum_{i=1}^{n}(\hat{y}_i - \bar{y}_n)^2$ und der Residualstreuung „sum of squared residuals“ $\mathrm{SQR} = \sum_{i=1}^{n}\hat{\varepsilon}_i^2 = \sum_{i=1}^{n}(y_i-\hat{y}_i)^2$ zerlegen:

$$\mathrm{SQT} = \mathrm{SQE} + \mathrm{SQR}$$

bzw.

$$\sum_{i=1}^{n}(y_i - \bar{y}_n)^2 = \sum_{i=1}^{n}(\hat{y}_i - \bar{y}_n)^2 + \sum_{i=1}^{n}(y_i - \hat{y}_i)^2\,.$$

Beweis

$$\begin{aligned}
\mathrm{SQT} &= \sum_{i=1}^{n}(y_i-\bar{y}_n)^2 = \sum_{i=1}^{n}(y_i-\hat{y}_i+\hat{y}_i-\bar{y}_n)^2 \\
&= \underbrace{\sum_{i=1}^{n}(y_i-\hat{y}_i)^2}_{=\mathrm{SQR}} + 2\sum_{i=1}^{n}(y_i-\hat{y}_i)(\hat{y}_i-\bar{y}_n) + \underbrace{\sum_{i=1}^{n}(\hat{y}_i-\bar{y}_n)^2}_{=\mathrm{SQE}} \\
&= \mathrm{SQE}+\mathrm{SQR}+2\sum_{i=1}^{n}\hat{y}_i(y_i-\hat{y}_i) - 2\bar{y}_n \underbrace{\sum_{i=1}^{n}(y_i-\hat{y}_i)}_{=0,\text{ vgl. Eig. 3 S. 37}} = \mathrm{SQE}+\mathrm{SQR}+E\,,
\end{aligned}$$

wobei noch zu zeigen ist, dass $E = 2\sum_{i=1}^{n}\hat{y}_i(y_i-\hat{y}_i) = 0$, also

$$\begin{aligned}
E &= 2\sum_{i=1}^{n}(\hat{\alpha}+\hat{\beta}x_i)(y_i-\hat{\alpha}-\hat{\beta}x_i) = 2\hat{\alpha}\underbrace{\sum_{i=1}^{n}\hat{\varepsilon}_i}_{=0} + 2\hat{\beta}\sum_{i=1}^{n}x_i(y_i-\hat{\alpha}-\hat{\beta}x_i) \\
&= 2\hat{\beta}\left(\sum_{i=1}^{n}x_iy_i - \hat{\alpha}\sum_{i=1}^{n}x_i - \hat{\beta}\sum_{i=1}^{n}x_i^2\right) \underset{\hat{\alpha}=\bar{y}_n-\bar{x}_n\hat{\beta}}{=} 2\hat{\beta}\left(\underbrace{\sum_{i=1}^{n}x_iy_i - n\bar{x}_n\bar{y}_n}_{=(n-1)S_{xy}^2} + \hat{\beta}n\bar{x}_n^2 - \hat{\beta}\sum_{i=1}^{n}x_i^2\right) \\
&= 2\hat{\beta}\left((n-1)S_{xy}^2 - \hat{\beta}(n-1)S_{xx}^2\right) \underset{\hat{\beta}=\frac{S_{xy}^2}{S_{xx}^2}}{=} 2\hat{\beta}(n-1)\left(S_{xy}^2 - \frac{S_{xy}^2}{S_{xx}^2}\cdot S_{xx}^2\right) = 0\,.
\end{aligned}$$

□

Die erklärte Streuung gibt die Streuung der Regressionsgeradenwerte um $\bar{y}_n$ an. Sie stellt damit die auf den linearen Zusammenhang zwischen X und Y zurückgeführende Variation der y-Werte dar. Das oben eingeführte Bestimmtheitsmaß ist somit der Anteil dieser Streuung an der Gesamtstreuung:

$$R^2 = \frac{\mathrm{SQE}}{\mathrm{SQT}} = \frac{\sum_{i=1}^{n}(\hat{y}_i-\bar{y}_n)^2}{\sum_{i=1}^{n}(y_i-\bar{y}_n)^2} = \frac{\mathrm{SQT}-\mathrm{SQR}}{\mathrm{SQT}} = 1-\frac{\mathrm{SQR}}{\mathrm{SQT}}\,.$$

Es folgt aus dieser Darstellung, dass $R^2 \in [0,1]$ ist.

1. $R^2 = 0$ bedeutet $\mathrm{SQE} = \sum_{i=1}^{n}(\hat{y}_i-\bar{y}_n)^2 = 0$ und somit $\hat{y}_i = \bar{y}_n\ \forall i$. Dies weist darauf hin, dass das lineare Modell in diesem Fall schlecht ist, denn aus $\hat{y}_i = \hat{\alpha}+\hat{\beta}x_i = \bar{y}_n$ folgt $\hat{\beta} = \frac{S_{xy}^2}{S_{xx}^2} = 0$ und somit $S_{xy}^2 = 0$. Also sind die Merkmale X und Y unkorreliert.

2. $R^2 = 1$ bedingt $\mathrm{SQR} = \sum_{i=1}^{n}\hat{\varepsilon}_i^2 = 0$. Somit liegen alle (x_i, y_i) perfekt auf der Regressionsgeraden. Dies bedeutet, dass die Daten x_i und y_i, $i = 1,\ldots,n$ perfekt linear abhängig sind.

   Faustregel zur Beurteilung der Güte der Anpassung eines linearen Modells an Hand von Bestimmtheitsmaß R^2:
   R^2 ist deutlich von Null verschieden (d.h. es besteht noch ein linearer Zusammenhang), falls $R^2 > \frac{4}{n+2}$, wobei n der Stichprobenumfang ist.

Allgemein gilt folgender Zusammenhang zwischen dem Bestimmtheitsmaß R^2 und dem Bravais-Pearson-Korrelationskoeffizienten ϱ_{xy}:

Lemma 2.5.4

$$R^2 = \varrho_{xy}^2$$

Beweis Aus der Eigenschaft 3 S. 37 folgt $\bar{y}_n = \overline{\hat{y}_n}$. Somit gilt

$$\text{SQE} = \sum_{i=1}^{n}(\hat{y}_i - \bar{y}_n)^2 = \sum_{i=1}^{n}(\hat{y}_i - \overline{\hat{y}_n})^2 = \sum_{i=1}^{n}(\hat{\alpha} + \hat{\beta}x_i - \hat{\alpha} - \hat{\beta}\bar{x}_n)^2 = \hat{\beta}^2 \sum_{i=1}^{n}(x_i - \bar{x}_n)^2$$

und damit

$$R^2 = \frac{\text{SQE}}{\text{SQT}} = \frac{\hat{\beta}^2 \sum_{i=1}^{n}(x_i - \bar{x}_n)^2}{\sum_{i=1}^{n}(y_i - \bar{y}_n)^2} = \frac{(S_{xy}^2)^2}{(S_{xx}^2)^2} \cdot \frac{(n-1)S_{xx}^2}{(n-1)S_{yy}^2} = \left(\frac{S_{xy}^2}{S_{yy}S_{xx}}\right)^2 = \varrho_{xy}^2$$

□

Folgerung 2.5.1

1. Der Wert von R^2 ändert sich bei einer Lineartransformation der Daten $(x_1, \ldots, x_n)$ und $(y_1, \ldots, y_n)$ nicht. Grafisch kann man die Güte der Modellanpassung bei der linearen Regression folgendermaßen überprüfen:

   Man zeichnet Punktepaare $(\hat{y}_i, \hat{\varepsilon}_i)_{i=1,\ldots,n}$ als Streudiagramm (der sogenannte *Residualplot*). Falls diese Punktewolke gleichmäßig um Null streut, so ist das lineare Modell gut gewählt worden. Falls das Streudiagramm einen erkennbaren Trend aufweist, bedeutet das, dass die Annahme des linearen Modells für diese Daten ungeeigenet sei (vgl. Abb. 2.18)

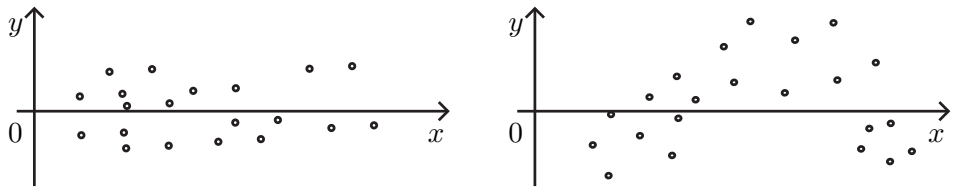


Abb. 2.18: Links: Gute, Rechts: Schlechte Übereinstimmung mit dem linearen Modell

2. Da $R^2 = \varrho_{xy}^2$, ist der Wert von R^2 symmetrisch bzgl. der Stichproben $(x_1, \ldots, x_n)$ und $(y_1, \ldots, y_n)$:

   $$\varrho_{xy}^2 = R^2 = \varrho_{yx}^2 \qquad \text{bzw.} \qquad R_{xy}^2 = R_{yx}^2\,,$$

   wobei R_{xy}^2 das Bestimmtheitsmaß bezeichnet, das sich aus der normalen Regression ergibt und R_{yx}^2 das mit vertauschten Achsen.

3 Punktschätzer

3.1 Parametrisches Modell

Sei $(x_1, \ldots, x_n)$ eine konkrete Stichprobe. Es wird angenommen, dass $(x_1, \ldots, x_n)$ eine Realisierung einer Zufallsstichprobe $(X_1, \ldots, X_n)$ ist, wobei $X_1, \ldots, X_n$ unabhängige identisch verteilte Zufallsvariablen mit der unbekannten Verteilungsfunktion F sind und F zu einer bekanten parametrischen Familie $\{F_\theta : \theta \in \Theta\}$ gehört. Hier ist $\theta = (\theta_1, \ldots, \theta_m) \in \Theta$ der *m-dimensionale Parametervektor* der Verteilung F_θ und $\Theta \subset \mathbb{R}^m$ der sogenannte *Parameterraum* (eine Borel-Teilmenge von $\mathbb{R}^m$, die die Menge aller zugelassenen Parameterwerte darstellt). Es wird vorausgesetzt, dass die Parametrisierung $\theta \to F_\theta$ *identifizierbar* ist, indem $F_{\theta_1} \neq F_{\theta_2}$ für $\theta_1 \neq \theta_2$ gilt.

Eine wichtige Aufgabe der Statistik, die wir in diesem Kapitel betrachten werden, besteht in der Schätzung des Parametervektors θ (oder eines Teils von θ) an Hand von der konkreten Stichprobe $(x_1, \ldots, x_n)$. In diesem Fall spricht man von einem *Punktschätzer* $\hat{\theta} : \mathbb{R}^n \to \mathbb{R}^m$, der eine gültige Stichprobenfunktion ist. Meistens wird angenommen, dass

$$\mathbb{P}\left(\hat{\theta}(X_1, \ldots, X_n) \in \Theta\right) = 1\,,$$

wobei es zu dieser Regel auch Ausnahmen gibt. Bisher haben wir den Wahrscheinlichkeitsraum $(\Omega, \mathcal{F}, \mathbb{P})$, auf dem unsere Zufallsstichprobe definiert ist, nicht näher spezifiziert. Dies kan man aber leicht tun, indem man den sogenannten *kanonischen Wahrscheinlichkeitsraum* angibt, wobei

$$\Omega = \mathbb{R}^\infty\,, \qquad \mathcal{F} = \mathcal{B}_{\mathbb{R}}^\infty = \mathcal{B}_{\mathbb{R}} \times \mathcal{B}_{\mathbb{R}} \times \cdots$$

und das Wahrscheinlichkeitsmaß $\mathbb{P}$ durch

$$\mathbb{P}\left(\{\omega = (\omega_1, \ldots, \omega_n, \ldots) \in \mathbb{R}^\infty : \omega_{i_1} \leq x_{i_1}, \ldots, \omega_{i_k} \leq x_{i_k}\}\right) = F_\theta(x_{i_1}) \ldots F_\theta(x_{i_k})$$

$\forall\, k \in \mathbb{N}, \quad 1 \leq i_1 < \ldots < i_k$ gegeben sei. Um zu betonen, dass $\mathbb{P}$ vom Parameter θ abhängt, werden wir Bezeichnungen $\mathbb{P}_\theta$, $\mathbb{E}_\theta$ und Var_θ für das Maß $\mathbb{P}$, den Erwartungswert und die Varianz bzgl. $\mathbb{P}$ verwenden.

Auf dem kanonischen Wahrscheinlichkeitsraum $(\Omega, \mathcal{F}, \mathbb{P}_\theta)$ gilt $X_i(\omega) = \omega_i$ (Projektion auf die Koordinate i), $i = 1, \ldots, n$,

$$\mathbb{P}_\theta(X_i \leq x_i) = \mathbb{P}_\theta\big(\{\omega \in \Omega : \omega_i \leq x_i\}\big) = F_\theta(x_i)\,, \quad i = 1, \ldots, n, \quad x_i \in \mathbb{R}\,.$$

Beispiel 3.1.1

1. Sei X die Dauer des fehlerfreien Arbeitszyklus eines technischen Systems. Oft wird $X \sim Exp(\lambda)$ angenommen. Dann stellt $\{F_\theta : \theta \in \Theta\}$ mit $m = 1$, $\theta = \lambda$, $\Theta = \mathbb{R}_+$ und

   $$F_\theta(x) = (1 - e^{-\theta x}) \cdot \mathbb{I}(x \geq 0)$$

   ein parametrisches Modell dar, wobei der Parameterraum eindimensional ist. Später wird für λ der (Punkt-) Schätzer $\hat{x}(x_1, \ldots, x_n) = 1/\bar{x}_n$ vorgeschlagen.

2. In den Fragestellungen der statistischen Qualitätskontrolle werden n Erzeugnisse auf Mängel untersucht. Falls $p \in (0,1)$ die unbekannte Wahrscheinlichkeit des Mangels ist, so wird mit $X \sim Bin(n,p)$ die Gesamtanzahl der mangelhaften Produkte beschrieben. Dabei wird folgendes parametrische Modell unterstellt:

$$\Theta = \{(n,p) : n \in \mathbb{N},\, p \in (0,1)\}\,, \qquad \theta = (n,p)\,, \quad m = 2\,,$$

$$F_\theta(x) = \mathbb{P}_\theta(X \le x) = \sum_{k=0}^{[x]} \binom{n}{k} p^k (1-p)^{n-k} \cdot \mathbb{I}(x \ge 0)\,.$$

Falls n bekannt ist, kann die Wahrscheinlichkeit p des Ausschusses durch den Punktschätzer $\hat{p}(x_1,\ldots,x_n) = \bar{x}_n\,, \quad x_i \in \{0,1\}$ näherungsweise berechnet werden.

3.2 Parametrische Familien von statistischen Prüfverteilungen

In der Vorlesung Wahrscheinlichkeitsrechung wurden bereits einige parametrische Familien von Verteilungen eingeführt. Hier geben wir weitere Verteilungsfamilien an, die in der Statistik eine besondere Stellung einnehmen, weil sie als Referenzverteilungen in der Schätztheorie, statistischen Tests und Vertrauensintervallen ihre Anwendung finden.

3.2.1 Gamma-Verteilung

Als erstes führen wir zwei spezielle Funktionen aus der Analysis ein:

1. Die *Gamma-Funktion:*

$$\Gamma(p) = \int_0^\infty x^{p-1} e^{-x} dx \qquad \text{für } p > 0\,.$$

Es gelten folgende Eigenschaften:

$$\Gamma(1) = 1\,, \qquad \Gamma(1/2) = \sqrt{\pi}$$
$$\Gamma(p+1) = p\Gamma(p) \qquad \forall\, p > 0\,, \qquad \Gamma(n+1) = n!\,, \qquad \forall\, n \in \mathbb{N}\,.$$

2. Die *Beta-Funktion:*

$$B(p,q) = \int_0^1 t^{p-1}(1-t)^{q-1}\, dt\,, \quad p,q > 0\,.$$

Es gelten folgende Eigenschaften:

$$B(p,q) = B(q,p)\,, \qquad B(p,q) = \frac{\Gamma(p)\Gamma(q)}{\Gamma(p+q)}\,, \qquad p,q > 0\,.$$

Definition 3.2.1
Die *Gamma-Verteilung* mit Parametern $\lambda > 0$ und $p > 0$ ist eine absolut stetige Verteilung mit der Dichte

$$f_X(x) = \begin{cases} \frac{\lambda^p x^{p-1}}{\Gamma(p)} e^{-\lambda x}\,, & x \ge 0\,, \\ 0\,, & x < 0\,. \end{cases} \tag{3.2.1}$$

Dabei verwenden wir die Bezeichnung $X \sim \Gamma(\lambda,p)$ für eine Zufallsvariable X, die Gamma-verteilt mit Parametern λ und p ist. Es gilt offensichtlich $X \ge 0$ fast sicher für $X \sim \Gamma(\lambda,p)$.

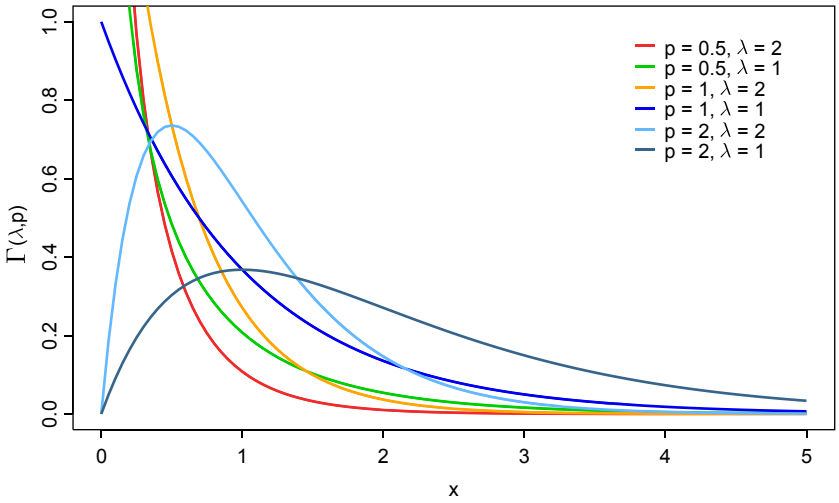


Abb. 3.1: Dichte der Gammaverteilung

Übungsaufgabe 3.2.1
Zeigen Sie, dass (3.2.1) eine Dichte ist.

Beispiel 3.2.1

1. In der Kraftfahrzeugversicherung wird die Gamma-Verteilung oft zur Modellierung des Gesamtschadens verwendet.

2. Falls $p = 1$, dann ist $\Gamma(\lambda, 1) = Exp(\lambda)$.

Satz 3.2.1 (Momenterzeugende und charakteristische Funktion der Gammaverteilung):
Falls $X \sim \Gamma(\lambda, p)$, dann gilt Folgendes:

1. Die momenterzeugende Funktion der Gammaverteilung $\Psi_X(s)$ ist gegeben durch

$$\Psi_X(s) = \mathbb{E}e^{sX} = \frac{1}{(1 - s/\lambda)^p}\,, \quad s < \lambda\,.$$

   Die charakteristische Funktion der Gammaverteilung $\varphi_X(s)$ ist gegeben durch

$$\varphi_X(s) = \mathbb{E}e^{isX} = \frac{1}{(1 - is/\lambda)^p}\,, \quad s \in \mathbb{R}\,.$$

2. k-te Momente:

$$\mathbb{E}X^k = \frac{p(p+1)\cdot\ldots\cdot(p+k-1)}{\lambda^k}\,, \quad k \in \mathbb{N}\,.$$

Beweis 1. Betrachte

$$\begin{aligned}\Psi_X(s) &= \int_0^\infty e^{sx} f_X(x)\,dx = \frac{\lambda^p}{\Gamma(p)}\int_0^\infty x^{p-1} e^{\overbrace{(s-\lambda)}^{<0}x}\,dx \\ &\underset{-(s-\lambda)x=y}{=} \frac{\lambda^p}{\Gamma(p)}\int_0^\infty \frac{y^{p-1}}{-(s-\lambda)^p} e^{-y}\,dy = \frac{\lambda^p\Gamma(p)}{\Gamma(p)(\lambda-s)^p} \\ &= \left(\frac{\lambda}{\lambda-s}\right)^p = \frac{1}{(1-s/\lambda)^p}\,, \quad \lambda > s\,.\end{aligned}$$

Falls $s \in \mathbb{C}$, $\mathrm{Re}\,(s) < \lambda$, dann ist $\Psi_X(s)$ holomorph auf $D = \{z = x+iy \in \mathbb{C} : x < \lambda\}$. Es gilt

$$\Psi_X(s) = \varphi_X(-is)\,, \qquad s = it,\ t < \lambda$$

Daraus folgt

$$\Psi_X(s) = \varphi_X(-is)\,, \qquad s \in D \Longrightarrow \varphi_X(s) = \frac{1}{(1-is/\lambda)^p}\,, \quad s \in \mathbb{R}\,.$$

2.

$$\mathbb{E}X^k = \Psi^{(k)}(0) \Longrightarrow \mathbb{E}X^k = \frac{p\cdot(p+1)\cdot\ldots\cdot(p+k-1)}{\lambda^k}\,, \quad k \in \mathbb{N}\,.$$

□

Folgerung 3.2.1 (Faltungsstabilität der Γ-Verteilung):
Falls $X \sim \Gamma(\lambda, p_1)$ und $Y \sim \Gamma(\lambda, p_2)$, X, Y unabhängig, dann ist $X + Y \sim \Gamma(\lambda, p_1 + p_2)$.

Beweis Es gilt

$$\varphi_{X+Y}(s) = \varphi_X(s)\cdot\varphi_Y(s) = \frac{1}{(1-is/\lambda)^{p_1}}\cdot\frac{1}{(1-is/\lambda)^{p_2}} = \left(\frac{1}{1-is/\lambda}\right)^{p_1+p_2} = \varphi_{\Gamma(\lambda,p_1+p_2)}(s)\,.$$

Da die charakteristischen Funktionen die Verteilungen eindeutig bestimmen, folgt damit $X + Y \sim \Gamma(\lambda, p_1 + p_2)$. □

Beispiel 3.2.2
Seien $X_1, \ldots, X_n \sim Exp(\lambda)$ unabhängig. Nach der Folgerung 3.2.1 gilt $X = X_1 + \ldots + X_n \sim \Gamma(\lambda, \underbrace{1 + \ldots + 1}_{n}) = \Gamma(\lambda, n)$, denn $Exp(\lambda) = \Gamma(\lambda, 1)$. Dabei heißt X *Erlang-verteilt* mit Parametern λ und n. Man schreibt $X \sim Erl(\lambda, n)$.

Zusammengefasst: $\qquad Erl(\lambda, n) = \Gamma(\lambda, n)$

Interpretation: In der Risikotheorie z.B. sind X_i Zwischenankunftszeiten der Einzelschäden. Dann ist $X = \sum_{i=1}^n X_i$ die Ankunftszeit des n-ten Schadens, $X \sim Erl(\lambda, n)$.

Definition 3.2.2 (χ^2-Verteilung):
X ist eine χ^2-verteilte Zufallsvariable mit k Freiheitsgraden ($X \sim \chi^2_k$), falls $X \overset{d}{=} X_1^2 + \ldots + X_k^2$, wobei $X_1, \ldots, X_k \sim N(0,1)$ unabhängige identisch verteilte Zufallsvariablen sind.

Satz 3.2.2 (χ^2-Verteilung: Spezialfall der Γ-Verteilung mit $\lambda = 1/2$, $p = k/2$):
Falls $X \sim \chi^2_k$, dann gilt:

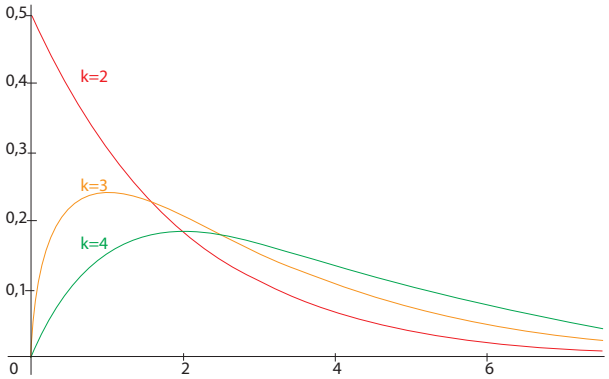


Abb. 3.2: Dichte der χ^2-Verteilung für $k = 2, 3, 4$

1. $X \sim \Gamma(1/2, k/2)$, d.h.
$$f_X(x) = \begin{cases} \dfrac{x^{k/2-1}e^{-x/2}}{2^{k/2}\Gamma(k/2)}, & x \geq 0 \\ 0, & x < 0 \end{cases}. \tag{3.2.2}$$

2. Insbesondere ist $\mathbb{E}X = k$, $\operatorname{Var} X = 2k$.

Beweis 1. Sei $X = X_1^2 + \ldots + X_k^2$ mit $X_i \sim N(0,1)$ unabhängingen identisch verteilten Zufallsvariablen. Errechnen wir zunächst die *Verteilung der* X_i^2*:*

$$\begin{aligned} P(X_1^2 \leq x) = P(X_1 \in [-\sqrt{x}, \sqrt{x}]) &= \int_{-\sqrt{x}}^{\sqrt{x}} \frac{1}{\sqrt{2\pi}} e^{-\frac{y^2}{2}}\, dy \\ &= \int_0^{\sqrt{x}} \frac{1}{\sqrt{2\pi}} e^{-\frac{y^2}{2}}\, dy + \int_{-\sqrt{x}}^{0} \frac{1}{\sqrt{2\pi}} e^{-\frac{y^2}{2}}\, dy \\ &\underset{y^2=t}{=} \int_0^x \frac{1}{\sqrt{2\pi}} e^{\frac{-t}{2}} \frac{1}{2\sqrt{t}}\, dt + \int_x^0 \frac{1}{\sqrt{2\pi}} e^{-t/2} \frac{-1}{2\sqrt{t}}\, dt \\ &= \int_0^x \frac{(1/2)^{-1/2} t^{1/2-1}}{\Gamma(1/2)} e^{-t/2}\, dt\,, \qquad x \geq 0\,. \end{aligned}$$

Somit folgt $X_1^2 \sim \Gamma(1/2, 1/2) \Longrightarrow X \sim \Gamma(1/2, \underbrace{1/2 + \ldots + 1/2}_{k}) = \Gamma(1/2, k/2)$ und daher gilt der Ausdruck (3.2.2) für die Dichte.

2. Wegen der Additivität des Erwartungswertes und der Unabhängigkeit von X_i gilt
$$\mathbb{E}X = k \cdot \mathbb{E}X_1^2\,, \quad \operatorname{Var} X = k \operatorname{Var} X_1^2\,, \quad \mathbb{E}(X_1^2) = \mathbb{E}\big(\Gamma(1/2, 1/2)\big)\,.$$
Bitte zeigen Sie selbstständig, dass $\mathbb{E}X_1^2 = 1\,, \quad \operatorname{Var} X_1^2 = 2$.

□

3.2.2 Student-Verteilung (t-Verteilung)

Definition 3.2.3
Seien X, Y unabhängige Zufallsvariablen, wobei $X \sim N(0,1)$ und $Y \sim \chi_r^2$. Dann heißt die

Zufallsvariable

$$U \stackrel{d}{=} \frac{X}{\sqrt{Y/r}}$$

Student- oder t-verteilt mit r Freiheitsgraden. Wir schreiben $U \sim t_r$.

Satz 3.2.3 (Dichte der t-Verteilung):
Falls $X \sim t_r$, dann gilt:

1. $$f_X(x) = \frac{1}{\sqrt{r}B\left(\frac{r}{2}, \frac{1}{2}\right)} \cdot \frac{1}{\left(1 + \frac{x^2}{r}\right)^{\frac{r+1}{2}}}, \qquad x \in \mathbb{R}.$$

2. $\mathbb{E}X = 0, \quad \operatorname{Var} X = \frac{r}{r-2}, \quad r \geq 3.$

Bemerkung 3.2.1

1. **Grafik von f:** Die t-Verteilung ist symmetrisch. Insbesondere gilt:

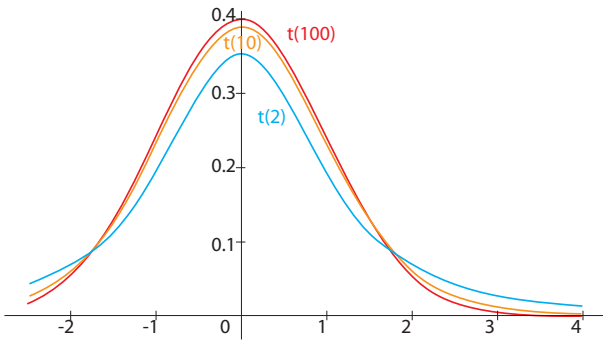


Abb. 3.3: Dichte der t-Verteilung

$$t_{r,\alpha} = -t_{r,1-\alpha}, \quad \alpha \in (0,1),$$

wobei $t_{r,\alpha}$ das α-Quantil der Student-Verteilung mit r Freiheitsgraden ist.

2. Falls $r \to \infty$, dann $f_r(x) \to \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}, \quad x \in \mathbb{R}$. (Übungsaufgabe)

3. Für $r = 1$ gilt: $t_1 = Cauchy(0,1)$ mit Dichte $f(x) = \frac{1}{\pi(1+x^2)}$. Der Erwartungswert von t_1 existiert nicht.

Beweis des Satzes 3.2.3:

1. Es gilt $X := \varphi(Y,Z)$, wobei $\varphi(x,y) = \frac{x}{\sqrt{y/r}}$ und $V = (Y,Z)$ ein zweidimensionaler Zufallsvektor ist, $Y \sim N(0,1)$, $Z \sim \chi^2_r$, Y und Z unabhängig.

   Wir wollen den sogenannten *Dichtetransformationssatz für Zufallsvektoren* verwenden, der besagt, dass unter bestimmten Voraussetzungen

$$f_{\varphi(V)}(x) = f_V(\varphi^{-1}(x))|J|$$

gilt, wobei $|J| = |\det J|$, $J = \left(\frac{\partial \varphi_i^{-1}(x)}{\partial x_j}\right)_{i,j=1}^n$, $\varphi = (\varphi_1, \ldots, \varphi_n) : \mathbb{R}^n \to \mathbb{R}^n$. Berechnen wir hier φ^{-1} von $\varphi : (x, y) \mapsto (v, w)$, wobei $v = \frac{x}{\sqrt{y/r}}$, $w = y$:

$$\varphi^{-1} : v = \frac{x}{\sqrt{\frac{y}{r}}} \Longrightarrow x = v\sqrt{\frac{y}{r}} = v\sqrt{\frac{w}{r}}\,. \quad \text{Somit } \varphi^{-1} : (v, w) \mapsto \left(v\sqrt{\frac{w}{r}}, w\right)$$

und die Jacobi-Matrix ist gleich

$$J = \begin{pmatrix} \frac{\partial \varphi_1^{-1}}{\partial v} & \frac{\partial \varphi_2^{-1}}{\partial w} \\ \frac{\partial \varphi_2^{-1}}{\partial v} & \frac{\partial \varphi_2^{-1}}{\partial w} \end{pmatrix} = \begin{pmatrix} \sqrt{\frac{w}{r}} & \frac{v}{2\sqrt{wr}} \\ 0 & 1 \end{pmatrix}.$$

Falls $V = (Y, Z)$, Y und Z unabhängig, dann

$$f_V(x, y) = f_Y(x) \cdot f_Z(y) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} \cdot \frac{y^{r/2-1} e^{-y/2}}{\Gamma(r/2) 2^{r/2}} = \frac{y^{r/2-1} e^{-\frac{y+x^2}{2}}}{2^{\frac{r+1}{2}} \Gamma(1/2)\Gamma(r/2)}\,, \quad x \in \mathbb{R}\,,\, y > 0\,,$$

und nach dem Dichtetransformationssatz gilt

$$\begin{aligned} f_X(v) &= \int_0^\infty f_{\varphi(V)}(u, w) dw = \int_0^\infty f_V(\varphi^{-1}(v, w)) |J|\, dw \\ &= \int_0^\infty \frac{e^{-(v^2 \frac{w}{r} + w)/2} w^{r/2-1}}{2^{\frac{r+1}{2}} \Gamma(1/2)\Gamma(r/2)} \sqrt{w/r}\, dw \\ &= \int_0^\infty w^{\frac{r-1}{2}} e^{-\overbrace{\frac{\frac{v^2}{r}+1}{2} \cdot w}^{=t}} dw \\ &\underset{w=\frac{2t}{v^2/r+1}}{=} \frac{1}{\sqrt{r} 2^{\frac{r+1}{2}} \Gamma(1/2)\Gamma(r/2)} \cdot \int_0^\infty \frac{2^{\frac{r-1}{2}+1} t^{\frac{r-1}{2}}}{(v^2/r+1)^{\frac{r-1}{2}+1}} e^{-t} dt \\ &= \frac{2^{\frac{r+1}{2}} \Gamma(\frac{r+1}{2})}{(\frac{v^2}{r}+1)^{\frac{r+1}{2}} \sqrt{r} 2^{\frac{r+1}{2}} \Gamma(1/2)\Gamma(r/2)} = \frac{1}{\sqrt{r} B(r/2, 1/2)(1 + v^2/r)^{\frac{r+1}{2}}} \end{aligned}$$

□

Da im Wahrscheinlichkeits-Skript der Dichtetransformationssatz nur für Zufallsvariablen formuliert wurde, geben wir hier die notwendigen Begriffe und verallgemeinerten Sätze für Zufallsvektoren (ohne Beweis). Hierbei verwenden wir die folgende Notation:

Für Vektoren $x = (x_1, \ldots, x_n)$ und $y = (y_1, \ldots, y_n)$ schreiben wir $x \leq y$, falls $x_i \leq y_i$ für $i = 1, \ldots, n$. Ferner sei für einen Zufallsvektor $X = (X_1, \ldots, X_n)$ die Verteilungsfunktion definiert als $F(x) = \mathbb{P}(X \leq x)$ für $x = (x_1, \ldots, x_n)$.

Definition 3.2.4
Die Zufallsvektoren $X_i : \Omega \to \mathbb{R}^{m_i}$, $\quad i = 1, \ldots, n$ sind *unabhängig*, falls

$$F_{(X_1,\ldots,X_n)}(x_1, \ldots, x_n) = P(X_1 \leq x_1, \ldots, X_n \leq x_n) = \prod_{i=1}^n P(X_i \leq x_i) = \prod_{i=1}^n F_{X_i}(x_i)\,,$$

$x_i \in \mathbb{R}^{m_i}$, $\quad i = 1, \ldots, n$.

Satz 3.2.4
Falls X_i absolut stetig verteilte und unabhängige Zufallsvektoren mit Dichten f_{X_i}, $i = 1, \ldots, n$, sind, dann ist auch $(X_1, \ldots, X_n)$ absolut stetig verteilt mit Dichte

$$f_{X_1,\ldots,X_n}(x_1, \ldots, x_n) = \prod_{i=1}^{n} f_{X_i}(x_i)\,, \quad x_i \in \mathbb{R}^{m_i}\,, \qquad i = 1, \ldots, n\,.$$

Satz 3.2.5
Falls $X_i : \Omega \to \mathbb{R}^{m_i}$, $i = 1, \ldots, n$ unabhängige Zufallsvektoren sind, und $\varphi : \mathbb{R}^{m_i} \to \mathbb{R}^{n_i}$, $\forall i = 1, \ldots, n$ Borel-messbare Funktionen, dann sind Zufallsvektoren $\varphi(X_1), \ldots, \varphi_n(X_n)$ unabhängig.

Satz 3.2.6 (Dichtetransformationssatz für Zufallsvektoren):
Sei $X = (X_1, \ldots, X_m) : \Omega \to \mathbb{R}^m$ ein absolut stetig verteilter Zufallsvektor mit Dichte f_X. Sei $\varphi = (\varphi_1, \ldots, \varphi_m) : \mathbb{R}^m \to \mathbb{R}^m$ eine Borel-messbare Abbildung, die innerhalb von einem Quader $B \subset \mathbb{R}^m$ stetig differenzierbar ist. Falls $\mathrm{supp} f_X \subset B$ und $det\left(\frac{\partial \varphi_i}{\partial x_j}\right)_{i,j=1,\ldots,m} \neq 0$ auf B, dann $\exists \varphi^{-1} : \varphi(B) \to B$ stetig differenzierbar und

$$f_{\varphi(X)}(x) = \begin{cases} f_X\left(\varphi^{-1}(x)\right) \cdot |J|, & x \in \varphi(B)\,, \\ 0, & x \notin \varphi(B)\,, \end{cases}$$

wobei $J = det\left(\frac{\partial \varphi_i^{-1}}{\partial x_j}\right)_{i,j=1,\ldots,m}$

3.2.3 Fisher-Snedecor-Verteilung (F-Verteilung)

Definition 3.2.5
Falls $X \overset{d}{=} \frac{U_r/r}{U_s/s}$, wobei $U_r \sim \chi_r^2$, $U_s \sim \chi_s^2$, $r, s \in \mathbb{N}$, U_r, U_s unabhängig, dann hat X eine F-Verteilung mit Freiheitsgraden r, s. Bezeichnung: $X \sim F_{r,s}$.

Lemma 3.2.1
Falls $X \sim F_{r,s}$, dann ist X absolut stetig verteilt mit Dichte

$$f_X(x) = \frac{x^{r/2-1}}{B(r/2, s/2)(r/s)^{-r/2}(1 + (r/s) \cdot x)^{\frac{r+s}{2}}} \cdot \mathbb{I}(x > 0)\,.$$

Beweis Da $U_r \sim \chi_r^2$, gilt für ihre Dichte

$$f_{U_r}(x) = \frac{x^{r/2-1} e^{x/2}}{\Gamma(r/2) 2^{r/2}}\,, \quad x > 0\,, \quad \forall r \in \mathbb{N}.$$

Somit

$$P\left(U_r/r \leq x\right) = P(U_r \leq rx) = F_{U_r}(rx)$$

und deshalb

$$\begin{aligned} f_{U_r/r}(x) &= (F_{U_r}(rx))' = r \cdot f_{U_r}(rx) = \frac{r(rx)^{r/2-1} e^{\frac{-rx}{2}}}{\Gamma(r/2) 2^{r/2}} \cdot \mathbb{I}(x > 0) \\ &= \frac{r^{r/2} x^{r/2-1} e^{-r/2 \cdot x}}{\Gamma(r/2) 2^{r/2}} \cdot \mathbb{I}(x > 0)\,. \end{aligned}$$

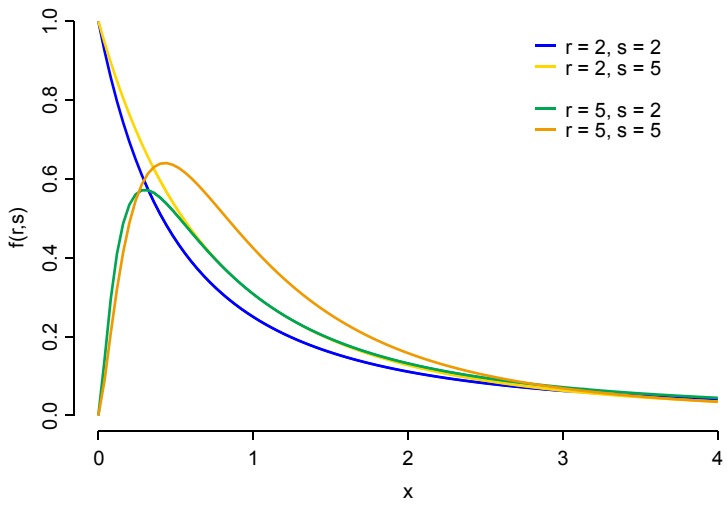


Abb. 3.4: Dichte der F-Verteilung für verschiedene Parameter r und s.

Nach dem Dichtetransformationssatz für das Verhältnis von zwei Zufallsvariablen (vgl. Wahrscheinlichkeitsskript Satz 3.15) gilt

$$f_{\frac{U_r/r}{U_s/s}}(x) = \int_0^\infty t f_{U_r/r}(xt) \cdot f_{U_s/s}(t)\, dt \cdot \mathbb{I}(x > 0)\,.$$

Somit

$$\begin{aligned}
f_X(x) &= \int_0^\infty t \frac{r^{r/2}(tx)^{r/2-1}e^{-\frac{rtx}{2}}}{\Gamma(r/2)2^{r/2}} \cdot \frac{s^{s/2}t^{s/2-1}e^{-st/2}}{\Gamma(s/2)2^{s/2}}\, dt \\
&= \frac{r^{r/2}s^{s/2}x^{r/2-1}}{\Gamma(r/2)\Gamma(s/2)2^{\frac{r+s}{2}}} \cdot \int_0^\infty t^{r/2+s/2-1} e^{-\overbrace{\frac{rx+s}{2}}^{=y} t}\, dt \\
&= \frac{r^{r/2}s^{s/2}x^{r/2-1}}{\Gamma(r/2)\Gamma(s/2)} \cdot \int_0^\infty \frac{y^{\frac{r+s}{2}-1}}{(rx+s)^{\frac{r+s}{2}}} \cdot e^{-y}\, dy \\
&\underset{t=\frac{y}{\frac{rx+s}{2}}}{=} \frac{r^{r/2}s^{s/2}x^{r/2-1}}{\Gamma(r/2)\Gamma(s/2)s^{\frac{r+s}{2}}(1+\frac{r}{s}\cdot x)^{\frac{r+s}{2}}} \cdot \Gamma\left(\frac{r+s}{2}\right) \\
&= \frac{(r/s)^{r/2}x^{r/2-1}}{B(r/2,s/2)(1+\frac{r}{s}x)^{\frac{r+s}{2}}} \cdot \mathbb{I}(x > 0)\,.
\end{aligned}$$

□

Bemerkung 3.2.2
Sei $X \sim F_{r,s},\ r, s \in \mathbb{N}$ mit Dichte f_X.

1. Einige Graphen der F-Verteilung sind in Abbildung 3.4 dargestellt.

2. Einige Eigenschaften der F-Verteilung:

Lemma 3.2.2
Es gilt:

a)
$$\mathbb{E}X = \frac{s}{s-2}, \qquad s \geq 3\,.$$

b)
$$\operatorname{Var} X = \frac{2s^2(r+s-2)}{r(s-4)(s-2)^2}\,, \qquad s \geq 5\,.$$

c) Falls $F_{r,s,\alpha}$ das α-Quantil der $F_{r,s}$-Verteilung ist, dann gilt

$$F_{r,s,\alpha} = \frac{1}{F_{s,r,1-\alpha}}\,, \quad \forall \alpha \in (0,1)\,.$$

Übungsaufgabe 3.2.2
Beweisen Sie Lemma 3.2.2!

3. Für Quantile $F_{r,s,\alpha}$ gilt folgende Näherungsformel (Abramowitz, Stegun (1972)):
$F_{r,s,\alpha} \approx e^{\omega}$, wobei

$$\begin{aligned}
\omega &= 2\left(\frac{\alpha(h+a)^{1/2}}{h} - \left(\frac{1}{r-1} - \frac{1}{s-1}\right)\cdot\left(a + \frac{5}{6} - \frac{2}{3h}\right)\right)\,,\\
h &= 2\left(\frac{1}{r-1} + \frac{1}{s-1}\right)^{-1}\,,\\
a &= \frac{z_\alpha^2 - 3}{6}
\end{aligned}$$

und z_α das α-Quantil der $N(0,1)$-Verteilung ist.

3.3 Punktschätzer und ihre Grundeigenschaften

Sei $(X_1, \ldots, X_n)$ eine Zufallsstichprobe, definiert auf dem kanonischen Wahrscheinlichkeitsraum $(\Omega, \mathcal{F}, \mathbb{P}_\theta)$. Seien X_i, $i = 1, \ldots, n$ unabhängige identisch verteilte Zufallsvariablen mit Verteilungsfunktion $F \in \{F_\theta : \theta \in \Theta\}$, $\Theta \subset \mathbb{R}^m$. Finde einen Schätzer $\hat{\theta}(X_1, \ldots, X_n)$ für den Parameter θ mit vorgegebenen Eigenschaften.

Unser Ziel im nächsten Abschnitt ist es, zunächst grundlegende Eigenschaften der Schätzer kennenzulernen.

3.3.1 Eigenschaften von Punktschätzern

Definition 3.3.1 (Erwartungstreue):
Ein Schätzer $\hat{\theta}(X_1, \ldots, X_n)$ für θ heißt *erwartungstreu* oder *unverzerrt*, falls

$$\mathbb{E}_\theta \hat{\theta}(X_1, \ldots, X_n) = \theta\,, \qquad \theta \in \Theta\,.$$

Dabei wird vorausgesetzt, dass

$$\mathbb{E}_\theta |\hat{\theta}(X_1, \ldots, X_n)| < \infty\,, \qquad \theta \in \Theta\,.$$

Der *Bias* (*Verzerrung*) eines Schätzers $\hat{\theta}(X_1, \ldots, X_n)$ ist gegeben durch

$$\mathrm{Bias}(\hat{\theta}) = \mathbb{E}_\theta\, \hat{\theta}(X_1, \ldots, X_n) - \theta\,.$$

Falls $\hat{\theta}(X_1, \ldots, X_n)$ erwartungstreu ist, dann gilt $\mathrm{Bias}(\hat{\theta}) = 0$ (kein systematischer Schätzfehler).

Definition 3.3.2 (Asymptotische Erwartungstreue):
Der Schätzer $\hat{\theta}(X_1, \ldots, X_n)$ für θ heißt *asymptotisch erwartungstreu* (oder *asymptotisch unverzerrt*), falls (für große Datenmengen)

$$\mathbb{E}_\theta\, \hat{\theta}(X_1, \ldots, X_n) \underset{n\to\infty}{\longrightarrow} \theta\,.$$

Definition 3.3.3 (Konsistenz):
Falls

$$\hat{\theta}(X_1, \ldots, X_n) \underset{n\to\infty}{\longrightarrow} \theta$$

in L^2, stochastisch bzw. fast sicher, dann heißt der Schätzer $\hat{\theta}(X_1, \ldots, X_n)$ ein *konsistenter Schätzer* für θ im *mittleren quadratischen, schwachen bzw. starken Sinne.*

- $\hat{\theta}$ *L^2-konsistent*: für $\mathbb{E}_\theta\, \hat{\theta}^2(X_1, \ldots, X_n) < \infty$ gilt

$$\hat{\theta} \underset{n\to\infty}{\overset{L^2}{\longrightarrow}} \theta \Longleftrightarrow \mathbb{E}_\theta |\hat{\theta}(X_1, \ldots, X_n) - \theta|^2 \underset{n\to\infty}{\longrightarrow} 0\,, \qquad \theta \in \Theta\,.$$

- $\hat{\theta}$ *schwach konsistent*:

$$\hat{\theta} \underset{n\to\infty}{\overset{\mathbb{P}}{\longrightarrow}} \theta \Longleftrightarrow P_\theta(|\hat{\theta}(X_1, \ldots, X_n) - \theta| > \varepsilon) \underset{n\to\infty}{\longrightarrow} 0\,, \qquad \varepsilon > 0\,, \quad \theta \in \Theta$$

- $\hat{\theta}$ *stark konsistent*:

$$\hat{\theta} \underset{n\to\infty}{\overset{\text{f.s.}}{\longrightarrow}} \theta \Longleftrightarrow P_\theta\left(\lim_{n\to\infty} \hat{\theta}(X_1, \ldots, X_n) = \theta\right) = 1\,, \qquad \theta \in \Theta$$

Daraus ergibt sich folgendes Diagramm (vgl. Wahrscheinlichkeitsrechungsskript, Kapitel 6).

$$\boxed{L^2 - \text{Konsistenz}} \Longrightarrow \boxed{\text{schwache Konsistenz}} \Longleftarrow \boxed{\text{starke Konsistenz}}$$

Definition 3.3.4 (Mittlerer quadratischer Fehler (mean squared error)):
Der mittlere quadratische Fehler eines Schätzers $\hat{\theta}(X_1, \ldots, X_n)$ für θ ist definiert als

$$MSE(\hat{\theta}) = \mathbb{E}_\theta \big|\hat{\theta}(X_1, \ldots, X_n) - \theta\big|^2\,.$$

Lemma 3.3.1
Falls $m = 1$ und $\mathbb{E}_\theta\, \hat{\theta}^2(X_1, \ldots, X_n) < \infty\,, \quad \theta \in \Theta$, dann gilt

$$MSE(\hat{\theta}) = \mathrm{Var}_\theta\, \hat{\theta}(X_1, \ldots, X_n) + \big(\mathrm{Bias}(\hat{\theta})\big)^2\,.$$

Beweis

$$\begin{aligned} MSE(\hat{\theta}) &= \mathbb{E}_\theta(\hat{\theta} - \theta)^2 = \mathbb{E}_\theta(\hat{\theta} - \mathbb{E}_\theta\hat{\theta} + \mathbb{E}_\theta\hat{\theta} - \theta)^2 \\ &= \underbrace{\mathbb{E}_\theta(\hat{\theta} - \mathbb{E}_\theta\, \hat{\theta})^2}_{\mathrm{Var}_\theta\, \hat{\theta}} + 2\, \underbrace{\mathbb{E}_\theta(\hat{\theta} - \mathbb{E}_\theta\, \hat{\theta})}_{=0}\, \underbrace{(\mathbb{E}_\theta\, \hat{\theta} - \theta)}_{=const} + \underbrace{(\mathbb{E}_\theta\, \hat{\theta} - \theta)^2}_{=\mathrm{Bias}(\hat{\theta})^2} \\ &= \mathrm{Var}_\theta\, \hat{\theta} + \Big(\mathrm{Bias}(\hat{\theta})\Big)^2\,. \end{aligned}$$

□

Bemerkung 3.3.1
Falls $\hat{\theta}$ *erwartungstreu* für θ ist, dann gilt $MSE(\hat{\theta}) = \mathrm{Var}_\theta\, \hat{\theta}$.

Definition 3.3.5 (Vergleich von Schätzern):
Seien $\hat{\theta}_1(X_1, \ldots, X_n)$ und $\hat{\theta}_2(X_1, \ldots, X_n)$ zwei Schätzer für θ. Man sagt, dass $\hat{\theta}_1$ *besser* ist als $\hat{\theta}_2$, falls

$$MSE(\hat{\theta}_1) < MSE(\hat{\theta}_2)\,, \qquad \theta \in \Theta\,.$$

Falls $m = 1$ und die Schätzer $\hat{\theta}_1$, $\hat{\theta}_2$ erwartungstreu sind, so ist $\hat{\theta}_1$ *besser als* $\hat{\theta}_2$, falls $\hat{\theta}_1$ die kleinere Varianz besitzt. Dabei wird stets vorausgesetzt, dass $\mathbb{E}_\theta\, \hat{\theta}_i^2 < \infty\,, \quad \theta \in \Theta$.

Definition 3.3.6 (Asymptotische Normalverteiltheit):
Sei $\hat{\theta}(X_1, \ldots, X_n)$ ein Schätzer für θ $(m = 1)$. Falls $0 < \mathrm{Var}_\theta\, \hat{\theta}(X_1, \ldots, X_n) < \infty\,, \quad \theta \in \Theta$ und

$$\frac{\hat{\theta}(X_1, \ldots, X_n) - \mathbb{E}_\theta\, \hat{\theta}(X_1, \ldots, X_n)}{\sqrt{\mathrm{Var}_\theta\, \hat{\theta}(X_1, \ldots, X_n)}} \xrightarrow[n\to\infty]{d} Y \sim N(0,1)\,,$$

dann ist $\hat{\theta}(X_1, \ldots, X_n)$ *asymptotisch normalverteilt.*

Definition 3.3.7 (Bester erwartungstreuer Schätzer):
Der Schätzer $\hat{\theta}(X_1, \ldots, X_n)$ für θ ist der *beste erwartungstreue Schätzer*, falls

$$\mathbb{E}_\theta\, \hat{\theta}^2(X_1, \ldots, X_n) < \infty\,, \quad \theta \in \Theta\,, \qquad \mathbb{E}_\theta\, \hat{\theta}(X_1, \ldots, X_n) = \theta\,, \quad \theta \in \Theta\,,$$

und $\hat{\theta}$ die minimale Varianz in der Klasse aller erwartungstreuen Schätzer für θ besitzt. Das heißt, dass für einen beliebigen erwartungstreuen Schätzer $\tilde{\theta}(X_1, \ldots, X_n)$ mit

$$\mathbb{E}_\theta\, \tilde{\theta}^2(X_1, \ldots, X_n) < \infty \quad \text{gilt} \quad \mathrm{Var}_\theta\, \hat{\theta} \le \mathrm{Var}_\theta\, \tilde{\theta}\,, \qquad \theta \in \Theta\,.$$

3.3.2 Schätzer des Erwartungswertes und empirische Momente

Sei $X \overset{d}{=} X_i\,, \quad i = 1, \ldots, n$ ein statistisches Merkmal. Sei weiter $\mathbb{E}|X_i|^k < \infty$ für ein $k \in \mathbb{N}$, $m = 1$ und der zu schätzende Parameter $\theta = \mu_k = \mathbb{E}X_i^k$. Insbesondere gilt im Fall $k = 1$, dass $\theta = \mu_1 = \mu$ der Erwartungswert ist.

Definition 3.3.8
Das *k-te empirische Moment* von X wird als

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

definiert. Unter dieser Definition gilt, dass $\hat{\mu}_1 = \bar{X}_n$, also das erste empirische Moment gleich dem Stichprobenmittel ist.

Satz 3.3.1 (Eigenschaften der empirischen Momente):
Unter obigen Voraussetzungen gelten folgende Eigenschaften:

1. $\hat{\mu}_k$ ist erwartungstreu für μ_k (insbesondere $\bar{X}_n$).
2. $\hat{\mu}_k$ ist stark konsistent.
3. Falls $\mathbb{E}_\theta|X|^{2k} < \infty\,, \quad \forall \theta \in \Theta$, dann ist $\hat{\mu}_k$ asymptotisch normalverteilt.

4. Es gilt $\operatorname{Var} \bar{X}_n = \frac{\sigma^2}{n}$, wobei $\sigma^2 = \operatorname{Var}_\theta X$. Falls $X_i \sim N(\mu, \sigma^2)\,, \quad i = 1, \ldots, n$ (eine normalverteilte Stichprobe), dann gilt:

$$\bar{X}_n \sim N\left(\mu, \frac{\sigma^2}{n}\right).$$

Beweis

1. $$\mathbb{E}_\theta\, \hat{\mu}_k = \frac{1}{n}\sum_{i=1}^n \mathbb{E}_\theta X_i^k = \frac{1}{n}\sum_{i=1}^n \mu_k = \frac{n\mu_k}{n} = \mu_k\,.$$

2. Aus dem starken Gesetz der großen Zahlen folgt

$$\frac{1}{n}\sum_{i=1}^n X_i^k \xrightarrow[n\to\infty]{\text{f.s.}} \mathbb{E}_\theta\, X_i^k = \mu_k\,.$$

3. Mit dem zentralen Grenzwertsatz gilt

$$\frac{\sum_{i=1}^n X_i^k - n\cdot \mathbb{E}X^k}{\sqrt{n\cdot \operatorname{Var} X^k}} = \frac{\frac{1}{n}\sum_{i=1}^n X_i^k - \mu_k}{\frac{1}{\sqrt{n}}\sqrt{\operatorname{Var} X^k}} = \sqrt{n}\frac{\hat{\mu}_k - \mu_k}{\sqrt{\operatorname{Var} X^k}} \xrightarrow[n\to\infty]{d} Y \sim N(0,1)\,.$$

Insbesondere gilt für den Spezialfall $k = 1$

$$\sqrt{n}\frac{\bar{X}_n - \mu}{\sigma} \xrightarrow[n\to\infty]{d} Y \sim N(0,1)\,.$$

4. $$\operatorname{Var} \bar{X}_n = \operatorname{Var}\left(\frac{1}{n}\sum_{i=1}^n X_i\right) \underset{X_i \text{ u.i.v.}}{=} \frac{1}{n^2}\sum_{i=1}^n \operatorname{Var} X_i = \frac{n\cdot\sigma^2}{n^2} = \frac{\sigma^2}{n}\,.$$

Falls $X_i \sim N(\mu, \sigma^2)\,, \quad i = 1, \ldots, n$, dann gilt wegen der Faltungsstabilität der Normalverteilung $\bar{X}_n \sim N(\cdot, \cdot)$, weil

$$\frac{1}{n}X_i \sim N\left(\frac{\mu}{n}, \frac{\sigma^2}{n^2}\right), \quad X_i \text{ u.i.v.}$$

Somit folgt aus 1) und 4) $\bar{X}_n \sim N\left(\mu, \frac{\sigma^2}{n}\right)$.

Damit ist der Satz bewiesen. □

Bemerkung 3.3.2
Aus Satz 3.3.1, 3) folgt

$$\begin{aligned}
\mathbb{P}(|\bar{X}_n - \mu| > \varepsilon) &= 1 - \mathbb{P}(-\varepsilon \le \bar{X}_n - \mu \le \varepsilon) \\
&= 1 - \mathbb{P}\left(-\frac{\varepsilon\sqrt{n}}{\sigma} \le \sqrt{n}\frac{\bar{X}_n - \mu}{\sigma} \le \frac{\varepsilon\sqrt{n}}{\sigma}\right) \\
&\underset{n\to\infty}{\approx} 1 - \left(\Phi\left(\frac{\varepsilon\sqrt{n}}{\sigma}\right) - \Phi\left(-\frac{\varepsilon\sqrt{n}}{\sigma}\right)\right) \\
&\underset{\Phi(-x)=1-\Phi(x)}{=} 1 - \left(\Phi\left(\frac{\varepsilon\sqrt{n}}{\sigma}\right) - 1 + \Phi\left(\frac{\varepsilon\sqrt{n}}{\sigma}\right)\right) \\
&= 1 - \left(2\Phi\left(\frac{\varepsilon\sqrt{n}}{\sigma}\right) - 1\right),
\end{aligned}$$

wobei $\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{t^2}{2}} dt$ die Verteilungsfunktion der $N(0,1)$-Verteilung ist.
Insgesamt gilt also für großes n

$$\mathbb{P}(|\bar{X}_n - \mu| > \varepsilon) \approx 2\left(1 - \Phi\left(\frac{\varepsilon\sqrt{n}}{\sigma}\right)\right).$$

3.3.3 Schätzer der Varianz

Seien $X_i\,, \quad i = 1, \ldots, n$ unabhängig identisch verteilt, $X_i \stackrel{d}{=} X\,, \quad \mathbb{E}_\theta X^2 < \infty \quad \forall \theta \in \Theta\,, \quad \theta = (\theta_1, \ldots, \theta_m)\,,\ \theta_i = \sigma^2 = \mathrm{Var}_\theta X$ für ein $i \in \{1, \ldots, m\}$. Die *Stichprobenvarianz*

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

ist dann ein Schätzer für σ^2. Falls der Erwartungswert $\mu = \mathbb{E}_\theta X$ der Stichprobenvariablen explizit benannt ist, so kann ein Schätzer für σ^2 auch als

$$\tilde{S}_n^2 = \frac{1}{n} \sum_{i=1}^{n} (X_i - \mu)^2$$

definiert werden.
Wir werden nun die Eigenschaften von S_n^2 und $\tilde{S}_n^2$ untersuchen und sie miteinander vergleichen.

Satz 3.3.2

1. Die Stichprobenvarianz S_n^2 ist erwartungstreu für σ^2:

$$\mathbb{E}_\theta S_n^2 = \sigma^2\,, \qquad \theta \in \Theta\,.$$

2. Wenn $\mathbb{E}_\theta X^4 < \infty$, dann gilt

$$\mathrm{Var}_\theta S_n^2 = \frac{1}{n}\left(\mu_4' - \frac{n-3}{n-1}\sigma^4\right),$$

wobei $\mu_4' = \mathbb{E}_\theta (X - \mu)^4$.

Beweis 1. Aus Lemma 2.2.1 1), 2) folgt, dass

$$S_n^2 = \frac{1}{n-1}\left(\sum_{i=1}^{n} X_i - n\bar{X}_n^2\right),$$

und dass man o.B.d.A. $\mu = \mathbb{E}_\theta X_i = 0$ annehmen kan, woraus insbesondere $\mathbb{E}_\theta \bar{X}_n = 0\,, \quad \theta \in \Theta$ folgt. Dann gilt

$$\begin{aligned}\mathbb{E}_\theta S_n^2 &= \frac{1}{n-1}\left(\sum_{i=1}^{n} \mathbb{E}_\theta X_i^2 - n\mathbb{E}\bar{X}_n^2\right) = \frac{1}{n-1}\left(\sum_{i=1}^{n} \mathrm{Var}_\theta X_i - n\mathrm{Var}\,\bar{X}_n\right) \\ &\underset{\text{S. 3.3.1, 4)}}{=} \frac{1}{n-1}\left(n\sigma^2 - n \cdot \frac{\sigma^2}{n}\right) = \sigma^2\,, \qquad \theta \in \Theta\,.\end{aligned}$$

2. Berechnen wir $\operatorname{Var}_\theta S_n^2 = \mathbb{E}_\theta (S_n^2)^2 - (\mathbb{E}_\theta S_n^2)^2 = \mathbb{E}_\theta (S_n^2)^2 - \sigma^4$. Es gilt

$$\begin{aligned}
\mathbb{E}_\theta S_n^4 &= \frac{1}{(n-1)^2} \mathbb{E}_\theta \left(\sum_{i=1}^n X_i^2 - n\bar{X}_n^2 \right)^2 \\
&= \frac{1}{(n-1)^2} \left(\underbrace{\mathbb{E}_\theta \left(\sum_{i=1}^n X_i^2 \right)^2}_{=I_1} - 2n \underbrace{\mathbb{E}_\theta \left(\bar{X}_n^2 \sum_{i=1}^n X_i^2 \right)}_{=I_2} + n^2 \underbrace{\mathbb{E}_\theta \bar{X}_n^4}_{=I_3} \right) .
\end{aligned}$$

Dabei gilt

$$\begin{aligned}
I_1 &= \mathbb{E}_\theta \left(\sum_{i=1}^n X_i^2 \sum_{j=1}^n X_j^2 \right) = \mathbb{E}_\theta \left(\sum_{i=1}^n X_i^4 + \sum_{i \neq j} X_i^2 X_j^2 \right) = \sum_{i=1}^n \mathbb{E}_\theta X_i^4 + \sum_{i \neq j} \mathbb{E}_\theta (X_i^2 X_j^2) \\
&\underset{X_i \text{ u.i.v, } \mu=0}{=} \sum_{i=1}^n \mu_4' + \sum_{i \neq j} \operatorname{Var}_\theta X_i \cdot \operatorname{Var}_\theta X_j = n\mu_4' + n(n-1)\sigma^4 ,
\end{aligned}$$

$$\begin{aligned}
I_2 &= \mathbb{E}_\theta \left(\left(\frac{1}{n} \sum_{i=1}^n X_i \right)^2 \sum_{j=1}^n X_j^2 \right) = \frac{1}{n^2} \mathbb{E}_\theta \left(\left(\sum_{i=1}^n X_i^2 + \sum_{i \neq j} X_i X_j \right) \sum_{j=1}^n X_j^2 \right) \\
&= \frac{1}{n^2} \mathbb{E}_\theta \left(\sum_{i=1}^n X_i^2 \sum_{j=1}^n X_j^2 \right) + \frac{1}{n^2} \mathbb{E} \left(\sum_{i \neq j} X_i X_j \sum_{k=1}^n X_k^2 \right) \\
&= \frac{1}{n^2} I_1 + \frac{1}{n^2} \sum_{i \neq j} \sum_k \underbrace{\mathbb{E} \left(X_i X_j X_k^2 \right)}_{=0, \text{ da } X_i \text{ u.i.v. und } \mu=0} = \frac{I_1}{n^2} = \frac{\mu_4' + (n-1)\sigma^4}{n} ,
\end{aligned}$$

$$\begin{aligned}
I_3 &= \mathbb{E}_\theta \left(\left(\frac{1}{n} \sum_{i=1}^n X_i \right)^2 \left(\frac{1}{n} \sum_{j=1}^n X_j \right)^2 \right) \\
&= \frac{1}{n^4} \mathbb{E}_\theta \left(\left(\sum_{k=1}^n X_k^2 + \sum_{i \neq j} X_i X_j \right) \cdot \left(\sum_{r=1}^n X_r^2 + \sum_{s \neq t} X_s X_t \right) \right) \\
&= \frac{1}{n^4} \mathbb{E}_\theta \left(\sum_{k,r=1}^n X_k^2 X_r^2 + 2 \sum_{k=1}^n X_k^2 \sum_{i \neq j} X_i X_j + \sum_{i \neq j} X_i X_j \sum_{s \neq t} X_s X_t \right) \\
&= \frac{1}{n^4} \Bigg(\mathbb{E} \Big(\sum_{k=1}^n X_k^4 \Big) + \mathbb{E}_\theta \Big(\sum_{k \neq r} X_k^2 X_r^2 \Big) + 2 \underbrace{\mathbb{E}_\theta \Big(\sum_{k=1}^n X_k^2 \sum_{i \neq j} X_i X_j \Big)}_{=0, \text{ da } X_i \text{ u.i.v. und } \mu=0} + \\
&\quad + 2 \underbrace{\mathbb{E}_\theta \Big(\sum_{i \neq j} X_i^2 X_j^2 \Big)}_{\text{weil } (i,j) \text{ und } (j,i) \text{ zählen}} + \underbrace{\mathbb{E}_\theta \left(\sum_{i \neq j \neq t} X_i^2 X_j X_t \right)}_{=0, \text{ da } X_j \text{ u.i.v. und } \mu=0} + \underbrace{\mathbb{E}_\theta \Big(\sum_{i \neq j \neq s \neq t} X_i X_j X_s X_t \Big)}_{=0, \text{ da } X_i \text{ u.i.v. und } \mu=0} \Bigg) \\
&= \frac{1}{n^4} \left(n\mu_4' + 3\mathbb{E}_\theta \sum_{i \neq j} X_i^2 X_j^2 \right) = \frac{n\mu_4' + 3n(n-1)\sigma^4}{n^4} = \frac{\mu_4' + 3(n-1)\sigma^4}{n^3} .
\end{aligned}$$

Somit gilt insgesamt

$$\begin{aligned}\mathbb{E}_\theta S_n^4 &= \frac{1}{(n-1)^2}\left(n\mu_4' + n(n-1)\sigma^4 - 2(\mu_4' + (n-1)\sigma^4) + \frac{\mu_4' + 3(n-1)\sigma^4}{n}\right)\\ &= \frac{(n^2-2n+1)\mu_4' + (n^2-2n+3)(n-1)\sigma^4}{n(n-1)^2}\\ &= \frac{(n-1)^2}{(n-1)^2}\frac{\mu_4^2}{n} + \frac{n^2-2n+3}{n(n-1)}\sigma^4 = \frac{\mu_4'}{n} + \frac{n^2-2n+3}{n(n-1)}\sigma^4\end{aligned}$$

und deshalb

$$\operatorname{Var}_\theta S_n^2 = \frac{\mu_4'}{n} + \frac{n^2-2n+3-n^2+n}{n(n-1)}\sigma^4 = \frac{\mu_4'}{n} - \frac{n-3}{n(n-1)}\sigma^4 = \frac{1}{n}\left(\mu_4' - \frac{n-3}{n-1}\sigma^4\right).$$

□

Satz 3.3.3 1. Der Schätzer $\tilde{S}_n^2$ für σ^2 ist erwartungstreu.

2. Es gilt $\operatorname{Var}_\theta \tilde{S}_n^2 = 1/n(\mu_4' - \sigma^4)$.

Beweis

1. $$\mathbb{E}_\theta \tilde{S}_n^2 = \frac{1}{n}\sum_{i=1}^n \underbrace{\mathbb{E}_\theta(X_i-\mu)^2}_{=\operatorname{Var}_\theta X_i} = \frac{1}{n}\sum_{i=1}^n \sigma^2 = \sigma^2.$$

2. Setzen wir wie in Satz 3.3.2 o.B.d.A. $\mu = 0$ voraus. Dann gilt

$$\begin{aligned}\operatorname{Var}\tilde{S}_n^2 &= \mathbb{E}_\theta\left(\frac{1}{n}\sum_{i=1}^n X_i^2\right)^2 - \left(\mathbb{E}_\theta \tilde{S}_n^2\right)^2 = \frac{1}{n^2}\mathbb{E}\left(\sum_{i=1}^n X_i^2\right) - \sigma^4\\ &\underset{I_1 \text{ Beweis S. 3.3.2}}{=} \frac{n\mu_4' + n(n-1)\sigma^4}{n^2} - \sigma^4 = \frac{\mu_4' + (n-1)\sigma^4}{n} - \sigma^4 = \frac{\mu_4' - \sigma^4}{n}.\end{aligned}$$

□

Folgerung 3.3.1
Der Schätzer $\tilde{S}_n^2$ für σ^2 ist besser als S_n^2, weil beide erwartungstreu sind und

$$\operatorname{Var}_\theta \tilde{S}_n^2 = \frac{\mu_4' - \sigma^4}{n} < \frac{\mu_4' - \frac{n-3}{n-1}\sigma^4}{n} = \operatorname{Var}_\theta S_n^2.$$

Diese Eigenschaft von $\tilde{S}_n^2$ im Vergleich zu S_n^2 ist intuitiv klar, da man in $\tilde{S}_n^2$ mehr Informationen über die Verteilung der Stichprobenvariablen X_i (nämlich den bekannten Erwartungswert μ) reingesteckt hat.

Satz 3.3.4
Die Schätzer S_n^2 bzw. $\tilde{S}_n^2$ sind stark konsistent und asymptotisch normalverteilt:

$$S_n^2 \xrightarrow[n\to\infty]{\text{f.s.}} \sigma^2, \qquad \sqrt{n}\frac{S_n^2 - \sigma^2}{\sqrt{\mu_4' - \sigma^4}} \xrightarrow[n\to\infty]{d} Y \sim N(0,1),$$

$$\tilde{S}_n^2 \xrightarrow[n\to\infty]{\text{f.s.}} \sigma^2, \qquad \sqrt{n}\frac{\tilde{S}_n^2 - \sigma^2}{\sqrt{\mu_4' - \sigma^4}} \xrightarrow[n\to\infty]{d} Y \sim N(0,1),$$

falls $\mu_4' < \infty$.

Beweis Zeigen wir nur, dass S_n^2 die obigen Eigenschaften besitzt. Der Beweis für $\tilde{S}_n^2$ verläuft analog. Die starke Konsistenz von S_n^2 folgt aus dem starken Gesetz der großen Zahlen, nach dem

$$\frac{1}{n}\sum_{i=1}^{n} X_i^2 \xrightarrow[n\to\infty]{\text{f.s.}} \mathbb{E}X^2 \quad \text{und} \quad \bar{X}_n \xrightarrow[n\to\infty]{\text{f.s.}} \mu$$

gilt und somit auch

$$\bar{X}_n^2 \xrightarrow[n\to\infty]{\text{f.s.}} \mu^2 .$$

Dann

$$\begin{aligned} S_n^2 &= \frac{1}{n-1}\left(\sum_{i=1}^{n} X_i - n\bar{X}_n^2\right) = \frac{n}{n-1}\left(\frac{1}{n}\sum_{i=1}^{n} X_i^2 - \bar{X}_n^2\right) \\ &\xrightarrow[n\to\infty]{\text{f.s.}} \mathbb{E}_\theta X^2 - \mu^2 = \operatorname{Var}_\theta X = \sigma^2 , \end{aligned}$$

und die starke Konsistenz ist bewiesen. Um die asymptotische Normalverteiltheit zu beweisen, nehmen wir o.B.d.A. an, dass $\mu = \mathbb{E}_\theta X = 0$. Dann folgt mit Hilfe des Satzes von Slutsky (vgl. Sätze 6.8 - 6.9 aus dem WR-Skript)

$$\begin{aligned} \sqrt{n}\frac{S_n^2 - \sigma^2}{\sqrt{\mu_4' - \sigma^4}} &= \sqrt{n}\frac{\frac{1}{n-1}\sum_{i=1}^{n} X_i^2 - \frac{n}{n-1}\bar{X}_n^2 - \sigma^2}{\sqrt{\mu_4' - \sigma^4}} \\ &= \sqrt{n}\frac{1}{n-1}\frac{\sum_{i=1}^{n} X_i^2 - n\sigma^2}{\sqrt{\mu_4' - \sigma^4}} - \underbrace{\frac{\sqrt{n}\frac{n}{n-1}\bar{X}_n^2}{\sqrt{\mu_4' - \sigma^4}}}_{=R_n^1} - \underbrace{\left(1 - \frac{n}{n-1}\right)\frac{\sigma^2\sqrt{n}}{\sqrt{\mu_4' - \sigma^4}}}_{=R_n^2} \\ &\underset{n\to\infty}{\overset{d}{\sim}} \frac{\sum_{i=1}^{n} X_i^2 - n\sigma^2}{\sqrt{n(\mu_4' - \sigma^4)}} , \end{aligned}$$

weil

$$R_n^2 = \left(1 - \frac{n}{n-1}\right)\sigma^2\frac{\sqrt{n}}{\sqrt{\mu_4' - \sigma^4}} = -\frac{\sigma^2}{\sqrt{\mu_4' - \sigma^4}}\frac{\sqrt{n}}{n-1} \xrightarrow[n\to\infty]{\text{f.s.}} 0 ,$$

also auch stochastisch und in Verteilung. Es gilt

$$R_n^1 \sim \sqrt{n}\frac{\bar{X}_n^2}{\sqrt{\mu_4' - \sigma^4}} \xrightarrow[n\to\infty]{d} 0 ,$$

weil

$$\mathbb{E}_\theta\left(\sqrt{n}\bar{X}_n^2\right) \underset{\mu=0}{=} \sqrt{n}\operatorname{Var}_\theta\bar{X}_n \underset{\text{S. 3.3.1, 4)}}{=} \sqrt{n}\frac{\sigma^2}{n} = \frac{\sigma^2}{\sqrt{n}} \xrightarrow[n\to\infty]{} 0$$

und somit

$$\sqrt{n}(\bar{X}_n)^2 \xrightarrow[n\to\infty]{L^1} 0 \Longrightarrow \sqrt{n}(\bar{X}_n)^2 \xrightarrow[n\to\infty]{\mathbb{P}} 0 \Longrightarrow \sqrt{n}(\bar{X}_n)^2 \xrightarrow[n\to\infty]{d} 0 .$$

Dann gilt

$$\lim_{n\to\infty}\sqrt{n}\frac{S_n^2 - \sigma^2}{\sqrt{\mu_4' - \sigma^4}} \overset{d}{=} \lim_{n\to\infty}\frac{\sum_{i=1}^{n} X_i^2 - n\sigma^2}{\sqrt{n(\mu_4' - \sigma^4)}} \overset{d}{=} Y \sim N(0,1)$$

nach dem zentralen Grenzwertsatz für die Folge von unabhängigen identisch verteilten Zufallsvariablen $\{X_i^2\}_{i\in\mathbb{N}}$, weil $\mathbb{E}_\theta X_i^2 \underset{\mu=0}{=} \mathrm{Var}_\theta X = \sigma^2$ und

$$\mathrm{Var}_\theta X_i^2 = \mathbb{E}_\theta X^4 - \left(\mathbb{E}_\theta X^2\right)^2 = \mu_4' - \sigma^4 .$$

□

Folgerung 3.3.2
Es gilt

1. $$\sqrt{n}\frac{\bar{X}_n - \mu}{S_n} \xrightarrow[n\to\infty]{d} Y \sim N(0,1)$$

   und somit

2. $$\mathbb{P}\left(\mu \in \left[\bar{X}_n - \frac{z_{1-\alpha/2}S_n}{\sqrt{n}}\,,\, \bar{X}_n + \frac{z_{1-\alpha/2}S_n}{\sqrt{n}}\right]\right) \xrightarrow[n\to\infty]{} 1-\alpha \tag{3.3.1}$$

   für ein $\alpha \in (0,1)$, wobei z_α das α-Quantil der $N(0,1)$-Verteilung ist.

Bemerkung 3.3.3
Das Intervall in (3.3.1) nennt man *asymptotisches Konfidenz- oder Vertrauensintervall* für den Parameter μ. Falls α klein ist (z.B. $\alpha = 0,05$), so liegt μ mit einer asymptotisch großen Wahrscheinlichkeit $1-\alpha$ im vorgegebenen Intervall. Diese Art der Schätzung von μ stellt eine Alternative zu den Punktschätzern dar und wird ausführlich in der Vorlesung Satistik II behandelt.

Beweis der Folgerung 3.3.2

1. Aus Satz 3.3.4 folgt

   $$S_n^2 \xrightarrow[n\to\infty]{\text{f.s.}} \sigma^2 \Longrightarrow \frac{\sigma}{S_n} \xrightarrow[n\to\infty]{\text{f.s.}} 1 \Longrightarrow \frac{\sigma}{S_n} \xrightarrow[n\to\infty]{d} 1$$

   und somit nach der Verwendung des Satzes von Slutsky

   $$\sqrt{n}\frac{\bar{X}_n - \mu}{S_n} = \sqrt{n}\frac{\bar{X}_n - \mu}{\sigma}\cdot\frac{\sigma}{S_n} \xrightarrow[n\to\infty]{} Y\cdot 1 = Y \sim N(0,1)\,,$$

   wobei wir die asymptotische Normalverteiltheit von $\bar{X}_n$ benutzt haben.

2. Aus 1) folgt

   $$\mathbb{P}_\theta\left(\sqrt{n}\frac{\bar{X}_n - \mu}{S_n} \in \left[z_{\alpha/2}\,,\, z_{1-\alpha/2}\right]\right) \xrightarrow[n\to\infty]{} \Phi\left(z_{1-\alpha/2}\right) - \Phi\left(z_{\alpha/2}\right) = 1 - \frac{\alpha}{2} - \frac{\alpha}{2} = 1-\alpha\,.$$

   Daraus folgt das Intervall (3.3.1) nach der Auflösung der Ungleichung

   $$z_{\alpha/2} \le \sqrt{n}\frac{\bar{X}_n - \mu}{S_n} \le z_{1-\alpha/2}$$

   bzgl. μ.

□

Betrachten wir weiterhin den wichtigen Spezialfall der normalverteilten Stichprobenvariablen $X_i\,, \quad i = 1, \ldots, n$, also $X \sim N(\mu, \sigma^2)$.

Satz 3.3.5
Falls $X_1, \ldots, X_n$ normalverteilt sind mit Parametern μ und σ^2, dann gilt

1. $$\frac{(n-1)S_n^2}{\sigma^2} \sim \chi^2_{n-1}\,,$$
2. $$\frac{n\tilde{S}_n^2}{\sigma^2} \sim \chi^2_n\,.$$

Beweis Beweisen wir den schwierigeren Fall 1, der Beweis im Fall 2 verläuft analog.

Da $X_i \sim N(\mu, \sigma^2)$, gilt, dass $\frac{X_i - \mu}{\sigma} \sim N(0,1)$ unabhängige identisch verteilte Zufallsvariablen für $i = 1, \ldots, n$ sind. Nach Lemma 2.2.1 gilt

$$\sum_{i=1}^n (X_i - \mu)^2 = \sum_{i=1}^n (X_i - \bar{X}_n)^2 + n(\bar{X}_n - \mu)^2$$

und somit

$$T_1 = \underbrace{\sum_{i=1}^n \left(\frac{X_i - \mu}{\sigma}\right)^2}_{\sim \chi^2_n} = \frac{n-1}{\sigma^2} S_n^2 + \underbrace{\left(\frac{\sqrt{n}(\bar{X}_n - \mu)}{\sigma}\right)^2}_{=T_2 \sim \chi^2_1 \text{ aus S. 3.3.1, 4)}}\,.$$

In Lemma 3.3.2 wird bewiesen, dass S_n^2 und $\bar{X}_n$ unabhängig sind. Somit gilt

$$\varphi_{T_1}(s) = \varphi_{\frac{n-1}{\sigma^2} S_n^2}(s) \cdot \varphi_{T_2}(s)\,, \qquad \forall s \in \mathbb{R}\,,$$

wobei $\varphi_Z(s)$ die charakteristische Funktion einer Zufallsvariablen Z ist. Da nach dem Satz 3.2.1

$$\varphi_{T_1}(s) = \frac{1}{(1-2is)^{n/2}}\,, \qquad \varphi_{T_2}(s) = \frac{1}{(1-2is)^{1/2}}\,,$$

folgt

$$\varphi_{\frac{n-1}{\sigma^2} S_n^2}(s) = \frac{\varphi_{T_1}(s)}{\varphi_{T_2}(s)} = \frac{1}{(1-2is)^{(n-1)/2}} = \varphi_{\chi^2_{n-1}}(s)\,.$$

Aus dem Satz 3.2.1 und dem Eindeutigkeitssatz für charakteristische Funktionen (vgl. Folgerung 5.1 aus dem WR-Skript) folgt

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

□

Lemma 3.3.2
Falls $X \sim N(\mu, \sigma^2)$, $X_1, \ldots, X_n$ unabhängige identisch verteilte Zufallsvariablen, $X_i \stackrel{d}{=} X$, dann sind $\bar{X}_n$ und S_n^2 unabhängig.

Dieses Lemma wird unter Anderem gebraucht, um folgendes Ergebnis zu beweisen:

Satz 3.3.6
Unter den Voraussetzungen von Lemma 3.3.2 gilt

$$\frac{\sqrt{n}(\bar{X}_n - \mu)}{S_n} \sim t_{n-1}\,.$$

Beweis von Lemma 3.3.2 Es folgt aus Lemma 2.2.1, dass

$$\bar{X}_n = \frac{1}{n}\sum_{i=1}^n X_i' + \mu \quad \text{und} \quad S_n^2 = \frac{1}{n-1}\sum_{i=1}^n (X_i' - \bar{X}_n')^2$$

für $X_i' = X_i - \mu$, $i = 1, \ldots, n$. Somit kann wegen des Satzes 3.2.5 o.B.d.A. $\mu = 0$ und $\sigma^2 = 1$ angenommen werden. Um die Unabhängigkeit von $\bar{X}_n$ und S_n^2 zu zeigen, stellen wir S_n^2 in alternativer Form dar:

$$S_n^2 = \frac{1}{n-1}\left((X_1 - \bar{X}_n)^2 + \sum_{i=2}^n (X_i - \bar{X}_n)^2\right) = \frac{1}{n-1}\left(\left(\sum_{i=2}^n (X_i - \bar{X}_n)\right)^2 + \sum_{i=2}^n (X_i - \bar{X}_n)^2\right),$$

weil $\sum_{i=1}^n (X_i - \bar{X}_n) = 0$ nach Abschnitt 2.2.1. Somit gilt

$$S_n^2 = \tilde{\varphi}(X_2 - \bar{X}_n, \ldots, X_n - \bar{X}_n),$$

wobei

$$\tilde{\varphi}(x_2, \ldots, x_n) = \frac{1}{n-1}\left(\left(\sum_{i=2}^n x_i\right)^2 + \sum_{i=2}^n x_i^2\right), \qquad (x_2, \ldots, x_n) \in \mathbb{R}^{n-1}.$$

Es genügt (nach Satz 3.2.5) zu zeigen, dass der Zufallsvektor $(X_2 - \bar{X}_n, \ldots, X_n - \bar{X}_n)$ unabhängig von $\bar{X}_n$ ist. Sei $X = (X_1, \ldots, X_n)$, X_i unabhängige identisch verteilte Zufallsvariablen mit $X_i \sim N(0,1)$ nach unserer Annahme. Dann gilt

$$f_X(x_1, \ldots, x_n) = \frac{1}{(2\pi)^{n/2}} \exp\left(-\frac{1}{2}\sum_{i=1}^n x_i^2\right), \qquad (x_1, \ldots, x_n) \in \mathbb{R}^n$$

für die Dichte von X. Sei $\varphi = (\varphi_1, \ldots, \varphi_n) : \mathbb{R}^n \to \mathbb{R}^n$ die lineare Abbildung mit

$$\begin{cases} \varphi_1(x) = \bar{x}_n, \\ \varphi_2(x) = x_2 - \bar{x}_n, \\ \quad\vdots \\ \varphi_n(x) = x_n - \bar{x}_n, \end{cases} \qquad x = (x_1, \ldots, x_n) \in \mathbb{R}^n.$$

Um die Umkehrabbildung $\varphi^{-1} : (y_1, \ldots, y_n) \mapsto (x_1, \ldots, x_n)$ zu finden, setzen wir $y_i = \varphi_i(x)$, $i = 1, \ldots, n$ und schreiben

$$\begin{cases} y_1 = \bar{x}_n \\ y_2 = x_2 - \bar{x}_n = x_2 - y_1 \\ \quad\vdots \\ y_n = x_n - y_1 \end{cases}, \text{ woraus } \begin{cases} x_2 = y_1 + y_2 \\ \quad\vdots \\ x_n = y_1 + y_m \\ x_2 + \ldots + x_n = (n-1)y_1 + y_2 + \ldots + y_n \\ x_1 + \ldots + x_n = ny_1 = x_1 + (n-1)y_1 + y_2 + \ldots + y_n \end{cases}$$

folgt und somit $x_1 = y_1 - \sum_{i=2}^n y_i$. Es gilt insgesamt

$$\begin{cases} \varphi_1^{-1}(y) = y_1 - \sum_{i=2}^n y_i, \\ \varphi_2^{-1}(y) = y_1 + y_2, \\ \quad\vdots \\ \varphi_n^{-1}(y) = y_1 + y_n, \end{cases} \qquad y = (y_1, \ldots, y_n) \in \mathbb{R}^n.$$

Um den Dichtetransformationssatz 3.2.6 für $\varphi(X)$ zu verwenden, brauchen wir die Determinante der Jacobi-Matrix

$$J = \det\left(\frac{\partial \varphi_i^{-1}}{\partial y_j}\right)_{i,j=1,\dots,n} = \left|\begin{pmatrix} 1 & -1 & -1 & -1 & \dots & -1 \\ 1 & 1 & 0 & 0 & \dots & 0 \\ 1 & 0 & 1 & 0 & \dots & 0 \\ \vdots & \vdots & \ddots & \ddots & \ddots & \vdots \\ 1 & 0 & \dots & 0 & 1 & 0 \\ 1 & 0 & \dots\dots\dots & & 0 & 1 \end{pmatrix}\right|$$

$$= 1\cdot 1 - (-1)\cdot 1 + (-1)\cdot(-1) - (-1)\cdot 1 + \dots = \underbrace{1+\dots+1}_{n} = n\,.$$

Somit gilt für die Dichte von $Y = \varphi(X) = (\bar{X}_n, X_2 - \bar{X}_n, \dots, X_n - \bar{X}_n)$

$$f_{\varphi(Y)}(y_1,\dots,y_n) = f_X(\varphi^{-1}(y))\cdot|J| = \frac{n}{(2\pi)^{n/2}}\exp\left\{-\frac{1}{2}\left(y_1 - \sum_{i=2}^n y_i\right)^2 - \frac{1}{2}\sum_{i=2}^n (y_1+y_i)^2\right\}$$

$$= \frac{n}{(2\pi)^{n/2}}\exp\left\{-\frac{1}{2}\left(y_1^2 - 2y_1\sum_{i=2}^n y_i + \left(\sum_{i=2}^n y_i\right)^2 + \sum_{i=2}^n y_i^2 + 2y_1\sum_{i=2}^n y_i + (n-1)y_1^2\right)\right\}$$

$$= \frac{n}{(2\pi)^{n/2}}\exp\left\{-\frac{1}{2}\left(ny_1^2 + \left(\sum_{i=2}^n y_i\right)^2 + \sum_{i=2}^n y_i^2\right)\right\}$$

$$= \underbrace{\left(\frac{n}{2\pi}\right)^{1/2}\exp\left\{-\frac{1}{2}ny_1^2\right\}}_{=f_{\varphi_1(X)}(y_1)}\cdot\underbrace{\left(\frac{n}{(2\pi)^{n-1}}\right)^{1/2}\exp\left\{-\frac{1}{2}\left(\sum_{i=2}^n y_i^2 + \left(\sum_{i=2}^n y_i\right)^2\right)\right\}}_{f_{(\varphi_2(X),\dots,\varphi_n(X))}(y_2,\dots,y_n)},$$

woraus die Unabhängigkeit von

$$\varphi_1(X) = \bar{X}_n \sim N\left(\mu, \frac{\sigma^2}{n}\right) \underset{\mu=0,\,\sigma^2=1}{=} N\left(0, \frac{1}{n}\right)$$

und

$$(\varphi_2(X),\dots,\varphi_n(X)) = (X_2 - \bar{X}_n, \dots, X_n - \bar{X}_n)$$

folgt. Somit sind auch $\bar{X}_n$ und $S_n^2 = \tilde{\varphi}(X_2 - \bar{X}_n, \dots, X_n - \bar{X}_n)$ unabhängig. □

Beweis des Satzes 3.3.6 Aus den Sätzen 3.3.1, 4) und 3.3.5 folgt

$$\bar{X}_n \sim N\left(\mu, \frac{\sigma^2}{n}\right) \quad \text{und} \quad \frac{(n-1)S_n^2}{\sigma^2} \sim \chi^2_{n-1}\,,$$

also

$$Y_1 = \sqrt{n}\frac{\bar{X}_n - \mu}{\sigma} \sim N(0,1) \quad \text{und} \quad Y_2 = \frac{(n-1)S_n^2}{\sigma^2} \sim \chi^2_{n-1}\,.$$

Nach dem Lemma 3.3.2 und Satz 3.2.5 sind Y_1 und Y_2 unabhängig. Dann gilt

$$T = \frac{Y_1}{\sqrt{\frac{Y_2}{n-1}}} \sim t_{n-1}$$

nach der Definition einer t-Verteilung, wobei

$$T = \frac{\sqrt{n}\frac{\bar{X}_n-\mu}{\sigma}}{\sqrt{\frac{(n-1)S_n^2}{\sigma^2(n-1)}}} = \sqrt{n}\frac{\bar{X}_n-\mu}{S_n}\,.$$

Somit gilt

$$\sqrt{n}\frac{\bar{X}_n-\mu}{S_n} \sim t_{n-1}\,.$$

□

Bemerkung 3.3.4
Mit Hilfe des Satzes 3.3.6 kann folgendes Konfidenzintervall für den Erwartungswert μ einer normalverteilten Stichprobe $(X_1,\ldots,X_n)$ bei unbekannter Varianz σ^2 $(X_i \sim N(\mu,\sigma^2)\,,\quad i = 1,\ldots,n)$ konstruiert werden:

$$\mathbb{P}\left(\mu \in \left[\bar{X}_n - \frac{t_{n-1,1-\alpha/2}}{\sqrt{n}}S_n\,,\ \bar{X}_n + \frac{t_{n-1,1-\alpha/2}}{\sqrt{n}}S_n\right]\right) = 1-\alpha$$

für $\alpha \in (0,1)$, denn

$$\begin{aligned}\mathbb{P}\Bigg(\sqrt{n}\frac{\bar{X}_n-\mu}{S_n} \in \Big[\underbrace{t_{n-1,\alpha/2}}_{=-t_{n-1,1-\alpha/2}\text{ wg. Sym. } t\text{-Vert.}},\ t_{n-1,1-\alpha/2}\Big]\Bigg) &= F_{t_{n-1}}(t_{n-1,1-\alpha/2}) - F_{t_{n-1}}(t_{n-1,\alpha/2}) \\ &= 1-\frac{\alpha}{2}-\frac{\alpha}{2} = 1-\alpha\,,\end{aligned} \tag{3.3.2}$$

wobei $t_{n-1,\alpha}$ das α-Quantil der t_{n-1}-Verteilung darstellt. Der Rest folgt aus (3.3.2) durch das Auflösen bzgl. μ.

3.3.4 Eigenschaften der Ordnungsstatistiken

In Abschnitt 2.2.2 haben wir bereits die Ordnungsstatistiken $x_{(1)},\ldots,x_{(n)}$ einer konkreten Stichprobe $(x_1,\ldots,x_n)$ betrachtet. Wenn wir nun auf der Modellebene arbeiten, also eine Zufallsstichprobe $(X_1,\ldots,X_n)$ von unabhängigen identisch verteilten Zufallsvariablen X_i mit Verteilungsfunktion $F(x)$ haben, welche Eigenschaften haben dann ihre Ordnungsstatistiken

$$X_{(1)},\ldots,X_{(n)}\,?$$

Satz 3.3.7

1. Die Verteilungsfunktion der Ordnungsstatistik $X_{(i)}\,,\quad i = 1,\ldots,n$ ist gegeben durch

$$F_{X_{(i)}}(x) = \sum_{k=i}^{n}\binom{n}{k}F^k(x)(1-F(x))^{n-k}\,,\qquad x \in \mathbb{R}\,. \tag{3.3.3}$$

2. Falls X_i eine diskrete Verteilung mit Wertebereich $E = \{\ldots,a_{j-1},a_j,a_{j+1},\ldots\}$ haben, $i = 1,\ldots,n$, $a_i < a_j$ für $i < j$, dann gilt für die Zähldichte von $X_{(i)}$, $i = 1,\ldots,n$:

$$\mathbb{P}(X_{(i)} = a_j) = \sum_{k=i}^{n}\binom{n}{k}\left(F^k(a_j)\big(1-F(a_j)\big)^{n-k} - F^k(a_{j-1})\big(1-F(a_{j-1})\big)^{n-k}\right)\,,$$

wobei

$$F(a_j) = \sum_{a_k \in E,\, k \le j} \mathbb{P}(X_i = a_k)\,.$$

3. Falls X_i absolut stetig verteilt sind mit Dichte f, die stückweise stetig ist, dann ist auch $X_{(i)}$, $i = 1, \dots, n$ absolut stetig verteilt mit der Dichte

$$f_{X_{(i)}}(x) = \frac{n!}{(i-1)!(n-i)!} f(x) F^{i-1}(x)(1 - F(x))^{n-i}\,, \quad x \in \mathbb{R}\,.$$

Beweis

1. Führen wir die Zufallsvariable

$$Y = \#\{i : X_i \le x\} = \sum_{i=1}^{n} \mathbb{I}(X_i \le x)\,, \quad x \in \mathbb{R}$$

ein. Da $X_1, \dots, X_n$ unabhängig identisch verteilt mit Verteilungsfunktion F sind, gilt $Y \sim Bin(n, F(x))$. Weiterhin gilt

$$F_{X_{(i)}}(x) = \mathbb{P}(X_{(i)} \le x) = \mathbb{P}(Y \ge i) = \sum_{k=i}^{n} \binom{n}{k} F^k(x)(1 - F(x))^{n-k}\,, \qquad x \in \mathbb{R}\,.$$

2. folgt aus 1) durch

$$\mathbb{P}(X_{(i)} = a_j) = \mathbb{P}(a_{j-1} < X_{(i)} \le a_j) = F_{X_{(i)}}(a_j) - F_{X_{(i)}}(a_{j-1}) \qquad \forall j, i\,.$$

3. Beweisen Sie 3) als Übungsaufgabe.

□

Bemerkung 3.3.5

1. Für $i = 1$ und $i = n$ sieht die Formel (3.3.3) besonders einfach aus:

$$F_{X_{(1)}}(x) = 1 - (1 - F(x))^n\,, \qquad x \in \mathbb{R}$$
$$F_{X_{(n)}}(x) = F^n(x)\,, \qquad x \in \mathbb{R}\,.$$

Diese Formeln lassen sich auch direkt herleiten:

$$\begin{aligned}
F_{X_{(1)}}(x) &= \mathbb{P}(\min_{i=1,\dots,n} X_i \le x) = 1 - \mathbb{P}(\min_{i=1,\dots,n} X_i > x) = 1 - \mathbb{P}(X_i \ge x\,, \quad \forall i = 1, \dots, n) \\
&\underset{X_i \text{ uiv}}{=} 1 - \prod_{i=1}^{n} \mathbb{P}(X_i > x) = 1 - (1 - F(x))^n\,, \\
F_{X_{(n)}}(x) &= \mathbb{P}(\max_{i=1,\dots,n} X_i \le x) = \mathbb{P}(X_i \le x\,, \quad \forall i = 1, \dots, n) \\
&\underset{X_i \text{ uiv}}{=} \prod_{i=1}^{n} \mathbb{P}(X_i \le x) = F^n(x)\,, \qquad x \in \mathbb{R}\,.
\end{aligned}$$

2. Falls X_i absolut stetig sind mit einer stückweise stetigen Dichte f, so lassen sich Formeln für die gemeinsame Dichte der Verteilung von $(X_{(i_1)}, \ldots, X_{(i_k)})\,, \quad i \leq k \leq n$ herleiten. Insbesondere gilt für $k = n$

$$f_{(X_{(1)},\ldots,X_{(n)})}(x_1, \ldots, x_n) = \begin{cases} n! \cdot f(x_1) \cdot \ldots \cdot f(x_n)\,, & \text{falls} - \infty < x_1 < \ldots < x_n < \infty\,, \\ 0\,, & \text{sonst.} \end{cases}$$

Übungsaufgabe 3.3.1
Zeigen Sie für $X_1, \ldots, X_n$ unabhängig identisch verteilt, $X_i \sim U[0, \theta]$, $\theta > 0$, $i = 1, \ldots, n$, dass

1. die Dichte von $X_{(i)}$ gleich

$$f_{X_{(i)}}(x) = \begin{cases} \frac{n!}{(i-1)!(n-i)!}\theta^{-n}x^{i-1}(\theta - x)^{n-i}\,, & x \in (0, \theta) \\ 0\,, & \text{sonst} \end{cases}$$

   und

2. $$\mathbb{E}X_{(i)}^k = \frac{\theta^k n!(i+k-1)!}{(n+k)!(i-1)!}\,, \qquad k \in \mathbb{N}\,, \quad i = 1, \ldots, n$$

sind. Insbesondere gilt $\mathbb{E}X_{(i)} = \frac{i}{n+1}\theta$ und $\operatorname{Var} X_{(i)} = \frac{i(n-i+1)\theta^2}{(n+1)^2(n+2)}$.

3.3.5 Empirische Verteilungsfunktion

Im Folgenden betrachten wir die statistischen Eigenschaften der in Abschnitt 2.1.2 eingeführten empirischen Verteilungsfunktion $\hat{F}_n(x)$ einer Zufallsstichprobe $(X_1, \ldots, X_n)$, wobei $X_i \stackrel{d}{=} X$ unabhängige identisch verteilte Zufallsvariablen mit Verteilungsfunktion $F(\cdot)$ sind.

Satz 3.3.8
Es gilt

1. $n\hat{F}_n(x) \sim Bin(n, F(x))\,, \quad x \in \mathbb{R}$.
2. $\hat{F}_n(x)$ ist ein erwartungstreuer Schätzer für $F(x)$, $x \in \mathbb{R}$ mit
$$\operatorname{Var} \hat{F}_n(x) = \frac{F(x)(1 - F(x))}{n}\,.$$
3. $\hat{F}_n(x)$ ist stark konsistent.
4. $\hat{F}_n(x)$ ist asymptotisch normalverteilt:
$$\sqrt{n}\frac{\hat{F}_n(x) - F(x)}{\sqrt{F(x)(1 - F(x))}} \xrightarrow{d} Y \sim N(0, 1)\,, \qquad \forall x : F(x) \in (0, 1)\,.$$

Beweis 1. folgt aus der Darstellung

$$\hat{F}_n(x) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{I}(X_i \leq x)\,, \qquad x \in \mathbb{R}\,,$$

weil $\mathbb{I}(X_i \leq x) \sim Bernoulli(F(x))\,, \quad \forall i = 1, \ldots, n$. Somit ist

$$\sum_{i=1}^{n} \mathbb{I}(X_i \leq x) \sim Bin(n, F(x))\,.$$

2. Es folgt aus 1)

$$\begin{cases} \mathbb{E}(n\hat{F}_n(x)) = nF(x)\,, & x \in \mathbb{R}\,, \\ \mathrm{Var}(n\hat{F}_n(x)) = nF(x) \cdot (1 - F(x))\,, & x \in \mathbb{R}\,, \end{cases}$$

woraus $\mathbb{E}\hat{F}_n(x) = F(x)$ und $\mathrm{Var}\,\hat{F}_n(x) = F(x)(1 - F(x))/n$ folgen.

3. Da $Y_i = \mathbb{I}(X_i \leq x)$, $i = 1, \ldots, n$, $x \in \mathbb{R}$ unabhängige identisch verteilte Zufallsvariablen sind, gilt nach dem starken Gesetz der großen Zahlen

$$\hat{F}_n(x) = \frac{1}{n} \sum_{i=1}^{n} Y_i \xrightarrow[n\to\infty]{\text{f.s.}} \mathbb{E}Y_i = F(x)\,.$$

4. folgt aus der Anwendung des zentralen Grenzwertsatzes auf die oben genannte Folge $\{Y_i\}_{i\in\mathbb{N}}$.

□

In Satz 3.3.8, 3) wird behauptet, dass

$$\hat{F}_n(x) \xrightarrow[n\to\infty]{\text{f.s.}} F(x)\,, \qquad \forall x \in \mathbb{R}\,.$$

Der nachfolgende Satz von Gliwenko-Cantelli behauptet, dass diese Konvergenz gleichmäßig in $x \in \mathbb{R}$ stattfindet. Um diesen Satz formulieren zu können, betrachten wir den *gleichmäßigen Abstand* zwischen $\hat{F}_n$ und F

$$D_n = \sup_{x\in\mathbb{R}} |\hat{F}_n(x) - F(x)|\,.$$

Dieser Abstand ist eine Zufallsvariable, die auch *Kolmogorow-Abstand* genannt wird. Er gibt den maximalen Fehler an, den man bei der Schätzung von $F(x)$ durch $\hat{F}_n(x)$ macht.

Übungsaufgabe 3.3.2
Zeigen Sie, dass

$$D_n = \max_{i\in\{1,\ldots,n\}} \max\left\{F\left(X_{(i)} - 0\right) - \frac{i-1}{n}\,, \frac{i}{n} - F\left(X_{(i)}\right)\right\}\,. \tag{3.3.4}$$

Beachten Sie dabei die Tatsache, dass $\hat{F}_n(x)$ eine Treppenfunktion mit Sprungstellen $X_{(i)}$, $i = 1, \ldots, n$ ist.

Satz 3.3.9 (Gliwenko-Cantelli):
Es gilt $D_n \xrightarrow[n\to\infty]{\text{f.s.}} 0$.

Beweis Für alle $m \in \mathbb{N}$ wähle beliebige Zahlen $-\infty = z_0 < z_1 < \ldots < z_{m-1} < z_m = \infty$. Dann gilt

$$D_n = \sup_{t\in\mathbb{R}} |\hat{F}_n(z) - F(z)| = \sup_{j=0,\ldots,m-1} \sup_{z\in[z_j, z_{j+1})} |\hat{F}_n(z) - F(z)|\,.$$

Zeigen wir, dass $\forall m \in \mathbb{N}$ $z_0, \dots, z_m$ existieren, für die gilt

$$F(z_{j+1} - 0) - F(z_j) \le \varepsilon = \frac{1}{m}\,. \tag{3.3.5}$$

Falls F stetig ist, genügt es, $z_j = F^{-1}(j/m)\,, \quad j = 1 \dots m-1$ gleichzusetzen. Im allgemeinen Fall existieren $n < m/2$ Punkte x_j mit der Eigenschaft

$$F(x_j) - F(x_j - 0) > 2\varepsilon = 2/m$$

(weil $n \cdot 2\varepsilon \le 1$ sein muss) und $k+1$ Punkte y_j zwischen diesen Punkten x_j mit Eigenschaft (3.3.5), wobei für k gilt:

$$n \cdot 2\varepsilon + (k+1)\varepsilon \le 1 \quad \Longrightarrow \quad 2n + k + 1 \le m \quad \Longrightarrow \quad k \le m - 2n - 1\,.$$

Setzen wir $\{z_j\} = \{x_j\} \cup \{y_j\}$. Für alle $z \in [z_j, z_{j+1})$ gilt

$$\hat{F}_n(z) - F(z) \le \hat{F}_n(z_{j+1} - 0) - F(z_j) \le \hat{F}_n(z_{j+1} - 0) - F(z_{j+1} - 0) + \varepsilon\,,$$

weil aus (3.3.5) folgt, dass $-F(z_j) \le \varepsilon - F(z_{j+1} - 0)\,, \quad \forall j$.

Genauso gilt

$$\hat{F}_n(z) - F(z) \ge \hat{F}_n(z_j) - F(z_{j+1} - 0) \ge \hat{F}_n(z_j) - F(z_j) - \varepsilon\,,$$

weil aus (3.3.5) für alle j folgt, dass $-F(z_{j+1} - 0) \ge -F(z_j) - \varepsilon$ gilt. Für alle $m \in \mathbb{N}\,, \quad j \in \{0, 1, \dots, m\}$ sei

$$\begin{aligned} A_{m,j} &= \{\omega \in \Omega : \lim_{n\to\infty} \hat{F}_n(z_j) = F(z_j)\}\,, \\ A'_{m,j} &= \{\omega \in \Omega : \lim_{n\to\infty} \hat{F}_n(z_j - 0) = F(z_j - 0)\}\,. \end{aligned}$$

Nach dem Satz 3.3.8, 3) gilt $\mathbb{P}(A_{m,j}) = 1$. Um $\mathbb{P}(A'_{m,j}) = 1$ zu zeigen, kann man die Verallgemeinerung von Aussage 3.3.8, 3) auf das Maß

$$\hat{F}_n(B) := \frac{1}{n}\sum_{i=1}^{n} \mathbb{I}(X_i \in B)\,, \qquad B \in \mathcal{B}_{\mathbb{R}}$$

benutzen: nach dem starken Gesetz der großen Zahlen gilt nämlich

$$\hat{F}_n(B) \xrightarrow[n\to\infty]{\text{f.s.}} F(B) = \mathbb{P}(X \in B)\,, \qquad B \in \mathcal{B}_{\mathbb{R}}\,.$$

Da $(-\infty, z_j) \in \mathcal{B}_{\mathbb{R}} \quad \forall j$, ist $\mathbb{P}(A'_{m,j}) = 1$ bewiesen $\forall m\ \forall j$. Für

$$A'_m = \bigcap_{j=0}^{m} \left(A_{m,j} \cap A'_{m,j}\right)$$

gilt $\mathbb{P}(A'_m) = 1\ \forall m$, weil

$$\mathbb{P}(A'_m) = 1 - \mathbb{P}(\bar{A}'_m) = 1 - \mathbb{P}\left(\bigcup_{j=0}^{m} \left(\bar{A}_{m,j} \cup \bar{A}'_{m,j}\right)\right) \ge 1 - \sum_{j=0}^{m} \left(\underbrace{\mathbb{P}(\bar{A}_{m,j})}_{=0} + \underbrace{\mathbb{P}(\bar{A}'_{m,j})}_{=0}\right) = 1\,.$$

Weiterhin: für $\varepsilon = 1/m$ $\forall \omega \in A'_m$ $\exists n(\omega, m) : \forall n > n(\omega, m)$ $\forall\, j \in \{0, \dots, m-1\}$ $\forall z \in [z_j, z_{j+1})$

$$\left.\begin{aligned} \hat{F}_n(z) - F(z) &\leq \underbrace{\hat{F}_n(z_{j+1} - 0) - F(z_{j+1} - 0)}_{<\varepsilon \text{ aus } A'_{m,j}} + \varepsilon < 2\varepsilon\,, \\ \hat{F}_n(z) - F(z) &\geq \underbrace{\hat{F}_n(z_j) - F(z_j)}_{>-\varepsilon \text{ aus } A_{m,j}} - \varepsilon > -2\varepsilon\,, \end{aligned}\right\} \Longrightarrow |\hat{F}_n(z) - F(z)| < 2\varepsilon\,.$$

$$\Longrightarrow D_n = \sup_{j=0,\dots,m-1} \sup_{z \in [z_j, z_{j+1})} |\hat{F}_n(z) - F(z)| < 2\varepsilon\,.$$

Nun wählen wir ein beliebiges $m \in \mathbb{N}$ und betrachten $A' = \bigcap_{m=1}^{\infty} A'_m$. Es folgt, dass $\mathbb{P}(A') = 1$ und $\forall\, \omega \in A'$ $\exists n_0 : \forall n \geq n_0$

$$D_n < 2\varepsilon = \frac{2}{m} \qquad \forall m \in \mathbb{N} \qquad \Longrightarrow \qquad D_n \xrightarrow[n\to\infty]{\text{f.s.}} 0\,.$$

□

Satz 3.3.10 (Ungleichung von Dvoretzky-Kiefer-Wolfowitz):
Seien $X_1, \dots X_n$ unabhängige identisch verteilte Zufallsvariablen mit Verteilungsfunktion F. Für alle $\varepsilon > 0$ gilt

$$\mathbb{P}(D_n > \varepsilon) \leq 2e^{-2n\varepsilon^2}\,.$$

(ohne Beweis)

Folgerung 3.3.3 (Konfidenzband für F):
Führen wir Statistiken

$$L(x) = \max\{\hat{F}_n(x) - \varepsilon_n, 0\} \text{ und } U(x) = \min\{\hat{F}_n(x) + \varepsilon_n, 1\}\,, \quad \varepsilon_n = \sqrt{\frac{1}{2n} \log\left(\frac{2}{\alpha}\right)}\,,\ \alpha \in (0,1)$$

ein. Dann gilt

$$\mathbb{P}\big(L(x) \leq F(x) \leq U(x) \quad \forall\, x \in \mathbb{R}\big) \geq 1 - \alpha \tag{3.3.6}$$

Beweis Beweisen Sie dieses Korollar als Übungsaufgabe! □

Bemerkung 3.3.6
Das simultane Konfidenzintervall $\{L(x) \leq F(x) \leq U(x)\,, \quad x \in \mathbb{R}\}$ aus (3.3.6) heißt *Konfidenzband* für F zum Konfidenzniveau $1 - \alpha$ (vgl. Abb. 3.5).

Falls die Verteilungsfunktion F stetig ist, kann man zeigen, dass die Zufallsvariable D_n nicht von F abhängt, also *verteilungsfrei* ist.

Satz 3.3.11
Für jede stetige Verteilungsfunktion F gilt

$$D_n \stackrel{d}{=} \sup_{y \in [0,1]} \left|\hat{G}_n(y) - y\right|\,, \text{ wobei } \hat{G}_n(y) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{I}\,(Y_i \leq y)\,, \quad y \in \mathbb{R}$$

die empirische Verteilungsfunktion der Zufallsstichprobe $(Y_1, \dots, Y_n)$ mit unabhängigen identisch verteilten Zufallsvariablen $Y_i \sim U[0,1]\,, \quad i = 1, \dots, n$ ist.

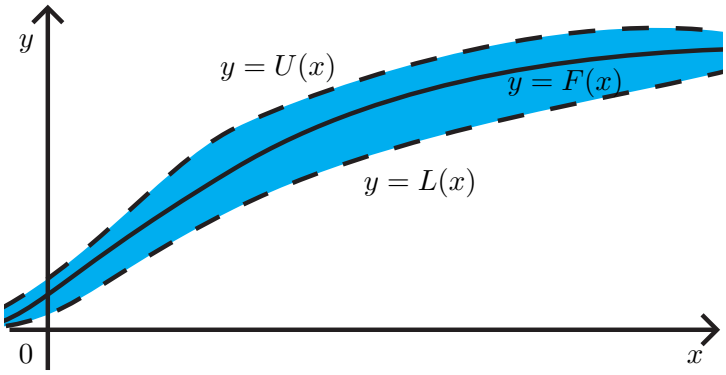


Abb. 3.5: Konfidenzband für F.

Beweis Zunächst definieren wir einen sogenannten *Konstanzbereich* $(a, b] \subset \mathbb{R}$ einer Verteilungsfunktion F als maximales Intervall mit der Eigenschaft $F(a) = F(b)$. Sei B die Vereinigung aller Konstanzbereiche von F. Auf B^C ist F eine monoton steigende eineindeutige Funktion. Damit folgt die Existenz ihrer Inversen $F^{-1} : [0, 1] \to B^C$. Gleichzeitig gilt

$$D_n = \sup_{x \in B^C} |\hat{F}_n(x) - F(x)| \,.$$

Führen wir $Y_i = F(X_i)\,, \quad i = 1, \ldots, n$ ein. Y_i sind unabhängig identisch verteilt und $Y_i \sim U[0, 1]$, denn

$$\mathbb{P}(Y_i \leq y) = \mathbb{P}(F(x_i) \leq y) = \mathbb{P}(X_i \leq F^{-1}(y)) = F(F^{-1}(y)) = y\,, \qquad y \in (0, 1)\,.$$

Somit gilt auch

$$\hat{F}_n(x) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{I}(X_i \leq x) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{I}(\underbrace{F(X_i)}_{Y_i} \leq F(x)) = \hat{G}_n(F(x))\,, \qquad x \in B^C\,.$$

Hieraus folgt

$$\begin{aligned} D_n &= \sup_{x \in B^C} \left|\hat{F}_n(x) - F(x)\right| = \sup_{x \in B^C} \left|\hat{G}_n(F(x)) - F(x)\right| = \sup_{x \in \mathbb{R}} \left|\hat{G}_n(F(x)) - F(x)\right| \\ &= \sup_{y \in [0,1]} \left|\hat{G}_n(y) - y\right|, \end{aligned}$$

wobei die letzte Gleichheit die Stetigkeit von F ausnützt. □

Folgerung 3.3.4
Falls F eine stetige Verteilungsfunktion ist, dann gilt

$$D_n \stackrel{d}{=} \max_{i=1,\ldots,n} \max \left\{ Y_{(i)} - \frac{i-1}{n}\,,\ \frac{i}{n} - Y_{(i)} \right\}\,,$$

wobei $Y_{(1)}, \ldots, Y_{(n)}$ die Ordnungsstatistiken der auf $[0, 1]$ gleichverteilten Stichprobenvariablen $Y_1, \ldots, Y_n$ sind.

Beweis Benutze dazu die Darstellung (3.3.4), den Satz 3.3.11 sowie die Tatsache, dass

$$F(x) = x\,, \quad x \in [0,1]$$

für die Verteilungsfunktion der $U[0,1]$-Verteilung ist. □

Folgende Ergebnisse werden ohne Beweis angegeben:

Bemerkung 3.3.7

1. Für die Zwecke des statistischen Testens (vgl. den Anpassungstest von Kolmogorow-Smirnow, Bemerkung 3.3.8, 3)) ist es notwendig, die Quantile der Verteilung von D_n zu nennen. Auf Grund der Komplexität der Verteilung von D_n ist es jedoch unmöglich, sie explizit anzugeben. Mit Hilfe des Satzes 3.3.11 ist es möglich, diese Quantile durch Monte-Carlo-Simulationen numerisch zu berechnen. Dazu simuliert man mehrere Stichproben $(Y_1, \ldots, Y_n)$ von $U[0,1]$-verteilten Pseudozufallszahlen, bildet $\hat{G}_n(x)$ und berechnet D_n nach Folgerung 3.3.4.

2. Für stetige Verteilungsfunktionen F kann folgende Integraldarstellung von Verteilungsfunktion von D_n bewiesen werden:

$$\mathbb{P}\left(D_n \le x + \frac{1}{2n}\right) = \begin{cases} 0\,, & x \le 0\,, \\ \int_{\frac{1}{2n}-x}^{\frac{1}{2n}+x} \int_{\frac{3}{2n}-x}^{\frac{3}{2n}+x} \cdots \int_{\frac{2n-1}{2n}-x}^{\frac{2n-1}{2n}+x} g(y_1, \ldots, y_n)\, dy_n \ldots dy_1, & 0 < x < \frac{2n-1}{2n}, \\ 1\,, & x \ge \frac{2n-1}{2n}\,. \end{cases}$$

   wobei

$$g(y_1, \ldots, y_n) = \begin{cases} n!\,, & 0 < y_1 < \ldots < y_n < 1\,, \\ 0\,, & \text{sonst} \end{cases}$$

   die Dichte der Ordnungsstatistiken $(Y_{(1)}, \ldots, Y_{(n)})$ von $U[0,1]$-verteilten Stichprobenvariablen $(Y_1, \ldots, Y_n)$ sind.

Satz 3.3.12 (Kolmogorow):
Falls die Verteilungsfunktion F der Stichprobenvariablen X_i, $i = 1, \ldots, n$ stetig ist, dann gilt

$$\sqrt{n} D_n \xrightarrow[n\to\infty]{d} Y\,,$$

wobei Y eine Zufallsvariable mit der Verteilungsfunktion

$$K(x) = \begin{cases} \sum_{k=-\infty}^{\infty} (-1)^k e^{-2k^2x^2} = 1 + 2\sum_{k=1}^{\infty} (-1)^k e^{-2k^2x^2}\,, & x > 0\,, \\ 0\,, & \text{sonst} \end{cases}$$

(Kolmogorow-Verteilung) ist.

Bemerkung 3.3.8

1. Die Verteilung von Kolmogorow ist die Verteilung des Maximums einer Brownschen Brücke, denn es gilt

$$Y \stackrel{d}{=} \sup_{t\in[0,1]} |w^\circ(t)|\,,$$

wobei $\{w^\circ(t)\,, \quad t \in [0,1]\}$ ein stochastischer Prozess ist, der die *Brownsche Brücke* genannt wird. Er wird als $w^\circ(t) = w(t) - w(1)t\,, \quad t \in [0,1]$ definiert, wobei $\{w(t)\,, t \in [0,1]\}$ die Brownsche Bewegung ist (für die unter anderem $w(t) \sim N(0,t)$ gilt). Der Name „Brücke“ ist der Tatsache $w^\circ(0) = w^\circ(1) = 0$ zu verdanken.

2. Aus Satz 3.3.12 folgt
$$\mathbb{P}(\sqrt{n}D_n \le x) \underset{n\to\infty}{\approx} K(x)\,, \quad x \in \mathbb{R}\,.$$
Die daraus resultierende Näherungsformel
$$\mathbb{P}(D_n \le x) \approx K(x\sqrt{n})$$
ist ab $n > 40$ praktisch brauchbar.

3. *Kolmogorow-Smirnow-Anpassungstest:* Mit Hilfe der Aussage des Satzes 3.3.12 ist es möglich, folgenden *asymptotischen Anpassungstest von Komogorow-Smirnow* zu entwickeln. Es wird die Haupthypothese $H_0 : F = F_0$ (die unbekannte Verteilungsfunktion der Stichprobenvariablen $X_1, \ldots, X_n$ ist gleich F_0) gegen die Alternative $H_1 : F \neq F_0$ getestet. Dabei wird H_0 verworfen, falls
$$\sqrt{n}D_n \notin [k_{\alpha/2}\,, k_{1-\alpha/2}]$$
ist, wobei
$$D_n = \sup_{x\in\mathbb{R}} |\hat{F}_n(x) - F_0(x)|$$
und k_α das α-Quantil der Kolmogorow-Verteilung ist. Somit ist die Wahrscheinlichkeit, die richtige Hypothese H_0 zu verwerfen (Wahrscheinlichkeit des *Fehlers 1. Art*) asymptotisch gleich
$$\mathbb{P}\left(\sqrt{n}D_n \notin [k_{\alpha/2}\,, k_{1-\alpha/2}] \,|\, H_0\right) \underset{n\to\infty}{\longrightarrow} 1 - K(k_{1-\alpha/2}) + K(k_{\alpha/2}) = 1 - (1-\alpha/2) + \alpha/2 = \alpha\,.$$
In der Praxis wird α klein gewählt, z.B. $\alpha \approx 0,05$. Somit ist im Fall, dass H_0 stimmt, die Wahrscheinlichkeit einer Fehlentscheidung in Folge des Testens klein.

   Dieser Test ist nur ein Beispiel dessen, wie der Satz von Kolmogorow in der statistischen Testtheorie verwendet wird. Die allgemeine Philosophie des Testens wird in Statistik II erläutert.

Mit Hilfe von $\hat{F}_n$ lassen sich sehr viele Schätzer durch die sogenannte *Plug-in-Methode* konstruieren. Dies werden wir jetzt näher erläutern: Sei $M = \{\text{Menge aller Verteilungsfunktionen}\}$.

Definition 3.3.9
Sei ein Parameter θ der Verteilungsfunktion F als Funktional $T : M \to \mathbb{R}$ von F gegeben: $\theta = T(F)$. Dann heißt $\hat{\theta} = T(\hat{F}_n)$ der *Plug-in-Schätzer* für θ.

Definition 3.3.10
Sei F eine beliebige Verteilungsfunktion. Das Funktional $T : M \to \mathbb{R}$ heißt *linear*, falls
$$T(aF_1 + bF_2) = aT(F_1) + bT(F_2) \qquad \forall a, b \in \mathbb{R}\,, \quad F_1, F_2 \in M\,.$$
Betrachten wir eine spezielle Klasse der linearen Funktionale
$$T(F) = \int_{\mathbb{R}} r(x)\, dF(x)\,,$$

wobei $r(x)$ eine beliebige stetige Funktion ist. Beispiele für solche T sind

$$\mathbb{E}X^k = \int_{\mathbb{R}} x^k \, dF(x)\,, \qquad k \in \mathbb{N}\,.$$

Lemma 3.3.3
Der Plug-in Schätzer für $\theta = \int_{\mathbb{R}} r(x)\, dF(x)$ ist durch

$$\hat{\theta} = \int_{\mathbb{R}} r(x)\, d\hat{F}_n(x) = \frac{1}{n}\sum_{i=1}^{n} r(x_i)$$

gegeben.

Übungsaufgabe 3.3.3
Beweisen Sie Lemma 3.3.3!

Beispiel 3.3.1 (Plug-in-Schätzer):

1. $\bar{X}_n$ ist ein Plug-in Schätzer für den Erwartungswert μ.
2. *Plug-in Schätzer* für $\sigma^2 = \operatorname{Var} X$: Es gilt $\operatorname{Var} X = \mathbb{E}X^2 - (\mathbb{E}X)^2$ und somit folgt

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n} X_i^2 - \left(\frac{1}{n}\sum_{i=1}^{n} X_i\right)^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2 = \frac{n-1}{n} S_n^2\,.$$

3. *Schätzer für Schiefe und Wölbung* $\hat{\gamma}_1$ und $\hat{\gamma}_2$ (vgl. Abschnitt 2.2.4) sind Plug-in Schätzer: Da der Koeffizient der Schiefe als

$$\gamma_1 = \mathbb{E}\left(\frac{X-\mu}{\sigma}\right)^3$$

definiert ist, wobei $\mu = \mathbb{E}X$, $\sigma^2 = \operatorname{Var} X$, folgt

$$\hat{\gamma}_1 \underset{\sigma^2 \mapsto \hat{\sigma}^2}{\overset{\mu \mapsto \bar{X}_n}{=}} \frac{\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X}_n)^3}{(\hat{\sigma}_n^2)^{3/2}} = \frac{\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X}_n)^3}{\left(\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2\right)^{3/2}}\,.$$

Die Konstruktion von $\hat{\gamma}_2$ erfolgt analog.

4. Der *empirische Korrelationskoeffizient* ϱ_{XY} ist ein Plug-in Schätzer:

$$\hat{\varrho}_{XY} = \frac{S_{XY}^2}{\sqrt{S_{XX}^2}\sqrt{S_{YY}^2}} = \frac{\sum_{i=1}^{n}(X_i - \bar{X}_n)(Y_i - \bar{Y}_n)}{\sqrt{\sum_{i=1}^{n}(X_i - \bar{X}_n)^2 \sum_{i=1}^{n}(Y_i - \bar{Y}_n)^2}}\,;$$

in der Tat ist

$$\varrho_{XY} = \frac{\mathbb{E}(X - \mathbb{E}X)(Y - \mathbb{E}Y)}{\sqrt{\operatorname{Var} X \cdot \operatorname{Var} Y}} = \frac{\mathbb{E}(XY) - \mathbb{E}X \cdot \mathbb{E}Y}{\sqrt{(\mathbb{E}X^2 - (\mathbb{E}X)^2)(\mathbb{E}Y^2 - (\mathbb{E}Y)^2)}}$$

und somit gilt für die linearen Funktionale

$$T_1(F) = \int x\, dF(x)\,, \qquad T_2(F) = \int x^2\, dF(x)\,, \qquad T_{12}(F,G) = \int xy\, dF(x) dG(y)\,,$$

$$\varrho_{XY} = \frac{T_{12}(F_X, F_Y) - T_1(F_X)\cdot T_1(F_Y)}{\sqrt{(T_2(F_X) - (T_1(F_X))^2)\,(T_2(F_Y) - (T_1(F_Y))^2)}}\,.$$

$\hat{\varrho}_{XY}$ bekommt man, in dem man T_1, T_2 und T_{12} durch Plug-in Schätzer ersetzt:

$$\hat{\varrho}_{XY} = \frac{T_{12}\left(\hat{F}_{n,X}, \hat{F}_{n,Y}\right) - T_1\left(\hat{F}_{n,X} \cdot T_1(\hat{F}_{n,Y})\right)}{\sqrt{\left(T_2(\hat{F}_{n,X}) - \left(T_1(\hat{F}_{n,X})\right)^2\right)\left(T_2(\hat{F}_{n,Y}) - \left(T_1(\hat{F}_{n,Y})\right)^2\right)}}\,.$$

3.4 Methoden zur Gewinnung von Punktschätzern

Sei $(X_1, \ldots, X_n)$ eine Stichprobe von unabhängigen identisch verteilten Zufallsvariablen X_i mit Verteilungsfunktion $F \in \{F_\theta : \theta \in \Theta\}$, $\Theta \subset \mathbb{R}^m$ (Parametrisches Modell). Sei die Parametrisierung $\theta \mapsto F_\theta$ unterscheidbar, d.h. $F_\theta \neq F_{\theta'} \iff \theta \neq \theta'$.

Zielstellung: Konstruiere einen Schätzer $\hat{\theta}(X_1, \ldots X_n)$ für $\theta = (\theta_1, \ldots, \theta_m)$.

3.4.1 Momentenschätzer

Aus der Wahrscheinlichkeitsrechung (Satz 4.8) folgt, dass unter gewissen Voraussetzungen (z.B. Gleichverteilung auf einem kompakten Intervall) an die Verteilung F diese Verteilung aus der Kenntnis von Momenten $\mathbb{E}X^k\,, \quad k \in \mathbb{N}$ wiedergewonnen werden kann. Auf dieser Idee der Schätzung von F aus den Momenten basiert die von Karl Pearson am Ende des XIX. Jh. vorgeschlagene *Momentenmethode*.

Annahme: Es existiert ein $r \geq m$, so dass $\mathbb{E}_\theta |X_i|^r < \infty$. Seien die Momente $\mathbb{E}_\theta X_i^k = g_k(\theta)$, $k = 1, \ldots, r$ als Funktionen des Parametervektors $\theta = (\theta_1, \ldots, \theta_m) \in \Theta$ gegeben.

Momenten-Gleichungssystem: $\hat{\mu}_k = g_k(\theta)\,, \quad k = 1, \ldots, r$, wobei $\hat{\mu}_k = \frac{1}{n}\sum_{k=}^n X_i^k$ die k-ten empirischen Momente sind.

Definition 3.4.1
Falls das obige Gleichungssystem eindeutig lösbar bzgl. θ ist, so heißt die Lösung $\hat{\theta}(X_1, \ldots, X_n)$ *Momentenschätzer (M-Schätzer)* von θ.

Lemma 3.4.1
Falls die Funktion $g = (g_1, \ldots, g_r) : \Theta \to C \subset \mathbb{R}^r$ eineindeutig und ihre Inverse $g^{-1} : C \to \Theta$ stetig ist, dann ist der M-Schätzer $\hat{\theta}(X_1, \ldots, X_n)$ von θ stark konsistent.

Beweis Es gilt $\hat{\theta}(X_1, \ldots, X_n) = g^{-1}(\hat{\mu}_1, \ldots, \hat{\mu}_r) \xrightarrow[n\to\infty]{\text{f.s.}} \theta$, weil $\hat{\mu}_k \xrightarrow[n\to\infty]{\text{f.s.}} g_k(\theta)\,, \quad k = 1, \ldots, r$ (starke Konsistenz der empirischen Momente) und g^{-1} stetig. □

Bemerkung 3.4.1

1. Unter gewissen Regularitätsbedingungen an F_θ ist der M-Schätzer $\hat{\theta}(X_1, \ldots, X_n)$ für θ asymptotisch normalverteilt:

$$\sqrt{n}\left(\hat{\theta}(X_1, \ldots, X_n) - \theta\right) \xrightarrow[n\to\infty]{d} N(0, \Sigma)\,,$$

wobei $N(0, \Sigma)$ die multivariate Normalverteilung mit Kovarianzmatrix

$$\Sigma = G\mathbb{E}(YY^T)G^T \qquad \text{ist mit} \quad Y = (X, X^2, \ldots, X^r)\,, \quad X \stackrel{d}{=} X_i\,,$$

und

$$G = \left(\frac{\partial g_i^{-1}}{\partial \theta_j}\right)_{\substack{i=1\ldots r\,,\\ j=1\ldots m}}\,.$$

2. Andere Eigenschaften gelten für M-Schätzer im Allgemeinen nicht. Zum Beispiel sind nicht alle M-Schätzer erwartungstreu (vgl. Beispiel 3.4.1, 1)).

3. Manchmal sind $r > m$ Gleichungen im Momentensystem notwendig, um einen M-Schätzer zu bekommen. Dies ist zum Beispiel dann der Fall, wenn manche Funktionen $g_i = const$ sind, d.h. sie enthalten keine Information über θ (vgl. Beispiel 3.4.1, 2)).

Beispiel 3.4.1

1. *Normalverteilung:* $X_i \overset{d}{=} X\,, \quad i = 1, \ldots, n\,, \quad X \sim N(\mu, \sigma^2)$; Gesucht ist ein M-Schätzer für μ und σ^2, also $\theta = (\mu, \sigma^2)$. Es gilt

$$\begin{aligned} g_1(\mu, \sigma^2) &= \mathbb{E}_\theta X = \mu\,, \\ g_2(\mu, \sigma^2) &= \mathbb{E}_\theta X^2 = \operatorname{Var}_\theta X + (\mathbb{E}_\theta X)^2 = \sigma^2 + \mu^2\,. \end{aligned}$$

Somit ergibt sich das Gleichungssystem

$$\begin{cases} \frac{1}{n}\sum_{i=1}^n X_i = \mu\,, \\ \frac{1}{n}\sum_{i=1}^n X_i^2 = \mu^2 + \sigma^2\,. \end{cases}$$

Damit folgt

$$\begin{aligned} \hat{\mu} &= \frac{1}{n}\sum_{i=1}^n X_i = \bar{X}_n\,, \\ \hat{\sigma}^2 &= \frac{1}{n}\sum_{i=1}^n X_i^2 - \hat{\mu}^2 = \frac{1}{n}\sum_{i=1}^n X_i^2 - \bar{X}_n^2 = \frac{1}{n}\sum_{i=1}^n \left(X_i^2 - n\bar{X}_n^2\right) \\ &= \frac{1}{n}\sum_{i=1}^n \left(X_i - \bar{X}_n\right)^2 = \frac{n-1}{n} S_n^2\,. \end{aligned}$$

Das heißt, das die M-Schätzer $\hat{\mu} = \bar{X}_n\,, \quad \hat{\sigma}^2 = \frac{n-1}{n} S_n^2$ sind. Dabei ist $\hat{\sigma}^2$ nicht erwartungstreu:

$$\mathbb{E}_\theta \hat{\sigma}^2 = \frac{n-1}{n} \cdot \mathbb{E}_\theta S_n^2 = \frac{n-1}{n}\sigma^2\,.$$

2. *Gleichverteilung:* $X_i \overset{d}{=} X\,, \quad i = 1, \ldots, n\,, \quad X \sim U[-\theta, \theta]\,, \quad \theta > 0$. Gesucht ist ein Momentenschätzer für θ. Es gilt

$$\begin{aligned} g_1(\theta) &= \mathbb{E}_\theta X = 0\,, \\ g_2(\theta) &= \mathbb{E}_\theta X^2 = \operatorname{Var}_\theta X = \frac{(\theta - (-\theta))^2}{12} = \frac{(2\theta)^2}{12} = \frac{\theta^2}{3}\,. \end{aligned}$$

Damit ergibt sich das Gleichungssystem

$$\begin{cases} \frac{1}{n}\sum_{i=1}^n X_i = 0 & \text{unbrauchbar}\,, \\ \frac{1}{n}\sum_{i=1}^n X_i^2 = \frac{\theta^2}{3}\,. \end{cases}$$

Es folgt, dass $\hat{\theta} = \sqrt{\frac{3}{n}\sum_{i=1}^n X_i^2}$ der Momentenschätzer für θ ist. Wir haben somit 2 Gleichungen für die Schätzung eines einzigen Parameters θ benötigt, d.h. $r = 2 > m = 1$.

3.4.2 Maximum-Likelihood-Schätzer

Diese wurden von Carl Friedrich Gauss (Anfang des XIX. Jh.) und Sir Ronald Fisher (1922) entdeckt. Seien alle Verteilungen aus der parametrischen Familie $\{F_\theta : \theta \in \Theta\}$ entweder diskret oder absolut stetig.

Definition 3.4.2

1. Falls die Stichprobenvariablen X_i, $i = 1, \ldots, n$ absolut stetig verteilt mit Dichte $f_\theta(x)$ sind, dann heißt

$$L(x_1, \ldots, x_n, \theta) = \prod_{i=1}^{n} f_\theta(x_i)\,, \qquad (x_1, \ldots, x_n) \in \mathbb{R}^n\,, \quad \theta \in \Theta$$

   die *Likelihood-Funktion* der Stichprobe $(x_1, \ldots, x_n)$.

2. Falls die Stichprobenvariablen X_i, $i = 1, \ldots, n$ diskret verteilt mit Zähldichte $p_\theta(x) = \mathbb{P}_\theta(X_i = x)\,, \quad x \in C$ sind (C ist der Wertebereich von X), dann heißt

$$L(x_1, \ldots, x_n, \theta) = \prod_{i=1}^{n} p_\theta(x_i)\,, \qquad (x_1, \ldots, x_n) \in C^n\,, \quad \theta \in \Theta$$

   die *Likelihood-Funktion* der Stichprobe $(x_1, \ldots, x_n)$.

Nach dieser Definition gilt im

- *diskreten Fall* $L(x_1, \ldots, x_n, \theta) = \mathbb{P}_\theta(X_1 = x_1, \ldots, X_n = x_n)$
- *absolut stetigen Fall*

$$\begin{aligned} L(x_1, \ldots, x_n, \theta)\Delta x_1 \cdot \ldots \cdot \Delta x_n &= f_{(X_1,\ldots,X_n),\theta}(x_1, \ldots, x_n)\Delta x_1 \cdot \ldots \cdot \Delta x_n \\ &\approx \mathbb{P}_\theta(X_1 \in [x_1\,,\, x_1 + \Delta x_1], \ldots, X_n \in [x_n\,,\, x_n + \Delta x_n])\,, \qquad \Delta x_i \to 0\,, \quad i = 1, \ldots, n\,. \end{aligned}$$

Nun wird ein Schätzer für θ so gewählt, dass die Wahrscheinlichkeit

$$\mathbb{P}_\theta(X_1 = x_1, \ldots, X_n = x_n) \quad \text{bzw.} \quad \mathbb{P}_\theta(X_i \in [x_i, x_i + \Delta x_i]\,, \quad i = 1, \ldots, n)$$

maximal wird. $\Longrightarrow$ Maximum-Likelihoodmethode:

Definition 3.4.3
Sei das Maximierungsproblem $L(x_1, \ldots, x_n, \theta) \mapsto \max_{\theta \in \Theta}$ eindeutig lösbar. Dann heißt

$$\hat{\theta}(x_1, \ldots, x_n) = \arg\max_{\theta \in \Theta} L(x_1, \ldots, x_n, \theta)$$

der *Maximum-Likelihood-Schätzer* von θ (*ML-Schätzer*).

Bemerkung 3.4.2

1. In relativ wenigen Fällen ist ein ML-Schätzer $\hat{\theta}$ für θ explizit auffindbar. In diesen Fällen wird meistens der konstante Faktor von $L(x_1, \ldots, x_n, \theta)$ weggeworfen und vom Rest der Logarithmus gebildet:

$$\log L(x_1, \ldots, x_n, \theta) \qquad \text{(die sog. Loglikelihood-Funktion).}$$

Dadurch wird

$$\prod_{i=1}^{n} f_\theta(x_i) \quad \text{bzw.} \quad \prod_{i=1}^{n} p_\theta(x_i)$$

zu einer Summe

$$\sum_{i=1}^{n} \log f_\theta(x_i) \quad \text{bzw.} \quad \sum_{i=1}^{n} \log p_\theta(x_i)\,,$$

die leichter bzgl. θ zu differenzieren ist. Danach betrachtet man

$$\frac{\partial \log L(x_1,\ldots,x_n,\theta)}{\partial \theta_j} = 0\,, \qquad j = 1\ldots m\,.$$

Dies ist die notwendige Bedingung eines Extremums von $\log L$ (und somit von L, weil $\log \nearrow$). Falls dieses System eindeutig lösbar ist, und die Lösung eine Maximum-Stelle ist, dann wird sie zum ML-Schätzer $\hat{\theta}(X_1,\ldots,X_n)$ erklärt.

2. In den meisten praxisrelevanten Fällen sind ML-Schätzer jedoch nur numerisch auffindbar.

Beispiel 3.4.2

1. *Bernoulli-Verteilung:* $X_i \overset{d}{=} X\,, \quad i = 1,\ldots,n,\ X \sim \mathit{Bernoulli}(p)$, für ein $p \in [0,1]$. Da

$$X = \begin{cases} 1\,, & \text{mit Wkt. } p \\ 0\,, & \text{sonst} \end{cases}$$

mit Zähldichte

$$p_\theta(x) = p^x(1-p)^{1-x}\,, \qquad x \in \{0,1\}\,,$$

ist die *Likelihood-Funktion der Stichprobe* $(X_1,\ldots,X_n)$ gegeben durch

$$L(x_1,\ldots,x_n,\theta) = \prod_{i=1}^{n} p^{x_i}(1-p)^{1-x_i} = p^{\sum_{i=1}^n x_i}(1-p)^{n-\sum_{i=1}^n x_i} \overset{\text{def.}}{=} h(p)\,.$$

a) Falls $\sum_{i=1}^n x_i = 0$ ($\Longleftrightarrow x_1 = x_2 = \ldots = x_n = 0$), es folgt $h(p) = (1-p)^n \to \max_{p\in[0,1]}$ bei $p = 0$. Dann ist der ML-Schätzer $\hat{p}(0,\ldots,0) = 0$.

b) Falls $\sum_{i=1}^n x_i = n$ ($\Longleftrightarrow x_1 = x_2 = \ldots = x_n = 1$), es folgt $h(p) = p^n \to \max_{p\in[0,1]}$ bei $p = 1$. Dann ist der ML-Schätzer $\hat{p}(1,1,\ldots,1) = 1$.

c) Falls $0 < \sum_{i=1}^n x_i < n$, dann gilt

$$\log L(x_1,\ldots,x_n,p) = n\bar{x}_n \log p + n(1-\bar{x}_n)\log(1-p) = n \cdot g(p)\,.$$

Da $g(p) \underset{p\to 0,1}{\longrightarrow} -\infty$ und

$$\frac{\partial \log L(x_1,\ldots,x_n,p)}{\partial p} = \frac{\bar{x}_n}{p} + \frac{1-\bar{x}_n}{1-p}\cdot(-1) = \frac{\bar{x}_n}{p} + \frac{\bar{x}_n - 1}{1-p} = 0$$
$$\Longleftrightarrow (1-p)\bar{x}_n + (\bar{x}_n - 1)p = 0 \quad \Longrightarrow \quad p = \bar{x}_n\,,$$

folgt aufgrund der Stetigkeit von g, dass g genau ein Extremum $\arg\max_p g(p) = \bar{x}_n$ besitzt.

Der ML-Schätzer ist also gegeben durch $\hat{p}(X_1, \ldots, X_n) = \bar{X}_n$.

2. *Gleichverteilung:* $X \sim U[0, \theta]\,, \quad \theta > 0\,, \quad (X_1, \ldots, X_n)$ unabhängig identisch verteilt, gesucht ist ein ML-Schätzer für θ. Es gilt

$$f_{X_i}(x) = 1/\theta \cdot \mathbb{I}(x \in [0, \theta])\,, \qquad i = 1, \ldots, n\,.$$

Somit ist die Likelihood-Funktion durch

$$\begin{aligned} L(x_1, \ldots, x_n, \theta) &= \begin{cases} (1/\theta)^n\,, & 0 \le x_1, \ldots, x_n \le \theta \\ 0\,, & \text{sonst} \end{cases} \\ &= \begin{cases} (1/\theta)^n\,, & \text{falls } \min\{x_1, \ldots, x_n\} \ge 0\,, \quad \max\{x_1, \ldots, x_n\} \le \theta \\ 0\,, & \text{sonst} \end{cases} \\ &= g(\theta)\,, \qquad \theta > 0 \end{aligned}$$

gegeben. Damit folgt $\hat{\theta} = \arg\max_{\theta > 0} g(\theta) = \max\{x_1, \ldots, x_n\} = x_{(n)}$, wodurch der ML-

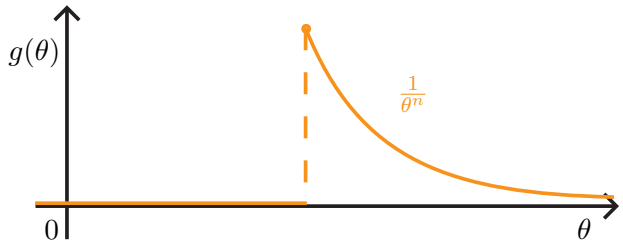


Abb. 3.6: Illustration der Funktion g.

Schätzer durch $\hat{\theta}(X_1, \ldots, X_n) = X_{(n)}$ gegeben ist.

Nun wollen wir zeigen, dass ML-Schätzer unter gewissen Voraussetzungen schwach konsistent und asymptotisch normalverteilt sind.

Definition 3.4.4
Sei

$$L(x, \theta) = \begin{cases} f_\theta(x)\,, & \text{im absolut stetigen Fall,} \\ p_\theta(x)\,, & \text{im diskreten Fall} \end{cases}$$

die Likelihood-Funktion von x. Für $\theta, \theta' \in \Theta\,, \quad \mathbb{P}_\theta(x \in \mathbb{R} : L(x, \theta') = 0) = 0$ definieren wir die *Information* (Abstand) $H(\mathbb{P}_\theta, \mathbb{P}_{\theta'})$ von *Kullback-Leibler* im absolut stetigen Fall als

$$H(\mathbb{P}_\theta, \mathbb{P}_{\theta'}) = \mathbb{E}_\theta \log L(X, \theta) - \mathbb{E}_{\theta'} \log L(X, \theta') = \int_{\mathbb{R}} \log \frac{L(x, \theta)}{L(x, \theta')} \cdot L(x, \theta)\, dx\,.$$

Für den Fall $\mathbb{P}_\theta(x \in \mathbb{R} : L(x, \theta') = 0) > 0$ setzen wir $H(\mathbb{P}_\theta, \mathbb{P}_{\theta'}) = \infty$. Im diskreten Fall betrachte statt des Integrals die Summe über die nicht trivialen $p_\theta(x)$.

Wir werden gleich zeigen, dass $H(\cdot\,,\,\cdot)$ die Eigenschaften $H(\mathbb{P}_\theta, \mathbb{P}_{\theta'}) = 0 \iff \theta = \theta'$ und $H(\mathbb{P}_\theta, \mathbb{P}_{\theta'}) \ge 0 \quad \forall\, \theta, \theta' \in \Theta$ besitzt. Es ist allerdings offensichtlich, dass $H(\mathbb{P}_\theta, \mathbb{P}_{\theta'})$ nicht symmetrisch bzgl. θ und θ' ist. Somit ist $H(\cdot\,,\,\cdot)$ keine Metrik.

Lemma 3.4.2
Es gilt

1. $H(\mathbb{P}_\theta, \mathbb{P}_{\theta'})$ ist wohldefiniert und ≥ 0.
2. Falls $H(\mathbb{P}_\theta, \mathbb{P}_{\theta'}) = 0$, dann gilt $\theta = \theta'$.

Beweis Wir betrachten zum Beispiel den Fall absolut stetiger $\mathbb{P}_\theta\,, \quad \theta \in \Theta$ (diskreter Fall folgt analog).

1. Definieren wir
$$f(x) = \begin{cases} \frac{L(x,\theta)}{L(x,\theta')}\,, & \text{falls } L(x,\theta') > 0\,, \\ 1\,, & \text{sonst.} \end{cases}$$
Betrachten wir den Fall $\mathbb{P}_\theta(x \in \mathbb{R} : L(x,\theta') = 0) = 0$, so folgt $\mathbb{P}_\theta(x \in \mathbb{R} : L(x,\theta') > 0) = 1$. Ansonsten ist $H(\mathbb{P}_\theta, \mathbb{P}_{\theta'}) = \infty > 0$, also positiv und wohldefiniert. Dann folgt mit Wahrscheinlichkeit 1, dass $L(x,\theta) = f(x) \cdot L(x,\theta')$. Sei $g(x) = 1 - x + x \log x\,, \quad x > 0$. Man kann zeigen, dass g konvex mit $g(x) \geq 0$ ist. Tatsächlich, es gilt
$$g'(x) = -1 + \log x + 1 = \log x\,, g''(x) = 1/x > 0\,.$$
Somit besitzt g genau eine Nullstelle bei $x = 1$, die gleichzeitig ihr Minimum ist. Betrachten wir $g(f(X))\,, \quad X \sim L(x,\theta')$. Dann gilt
$$\begin{aligned} 0 &\leq \mathbb{E}_{\theta'} g(f(X)) = 1 - \mathbb{E}_{\theta'} f(X) + \mathbb{E}_{\theta'}\left(f(X) \log f(X)\right) \\ &= 1 - \int \frac{L(x,\theta)}{L(x,\theta')} \cdot L(x,\theta')\,dx + \int \frac{L(x,\theta)}{L(x,\theta')} \cdot \log \frac{L(x,\theta)}{L(x,\theta')} \cdot L(x,\theta')\,dx = H(\mathbb{P}_\theta, \mathbb{P}'_\theta)\,. \end{aligned}$$
Somit gilt $H(\mathbb{P}_\theta, \mathbb{P}_{\theta'}) \geq 0$, was zu zeigen war.
2. Falls $H(\mathbb{P}_\theta, \mathbb{P}_{\theta'}) = 0 \quad \Longrightarrow \quad \mathbb{E}_{\theta'} g(f(X)) = 0\,, \quad g(f(X)) \geq 0$. Somit folgt fast sicher $g(f(X)) = 0 \quad \Longrightarrow f(X) \overset{\text{f.s.}}{=} 1$, damit entweder $L(x,\theta') = 0$ oder $L(x,\theta) = L(x,\theta')$ und daher $\mathbb{P}_\theta = \mathbb{P}_{\theta'}$.

□

Satz 3.4.1 (Schwache Konsistenz von ML-Schätzern):
Sei $m = 1$ und Θ ein offenes Intervall aus $\mathbb{R}$. Sei $L(x_1, \ldots, x_n, \theta)$ *unimodal*, d.h. für $\hat{\theta}$ ML-Schätzer für θ gilt
$$\begin{cases} \forall\, \theta < \hat{\theta}(x_1, \ldots, x_n) & \Longrightarrow L(x_1, \ldots, x_n, \theta) \text{ ist steigend} \\ \forall\, \theta > \hat{\theta}(x_1, \ldots, x_n) & \Longrightarrow L(x_1, \ldots, x_n, \theta) \text{ ist fallend} \end{cases}$$
(d.h. es existiert genau ein $\max_{\theta \in \Theta} L(x_1, \ldots, x_n, \theta)$). Dann gilt $\hat{\theta}(X_1, \ldots, X_n) \underset{n\to\infty}{\overset{P}{\longrightarrow}} \theta$.

Beweis Es ist zu zeigen, dass
$$\mathbb{P}_\theta\left(\left|\hat{\theta}(X_1, \ldots, X_n) - \theta\right| > \varepsilon\right) \underset{n\to\infty}{\longrightarrow} 0\,, \qquad \varepsilon > 0\,. \tag{3.4.1}$$
Wählen wir beliebiges $\varepsilon > 0 : \quad \theta \pm \varepsilon \in \Theta$. Dann gilt $H(\mathbb{P}_\theta, \mathbb{P}_{\theta\pm\varepsilon}) > \delta > 0$, wegen der Unterscheidbarkeit der Parametrisierung von $\mathbb{P}_\theta$ und Lemma 3.4.2. Betrachten wir $\{|\hat{\theta} - \theta| \leq \varepsilon\}$.

Um (3.4.1) zu zeigen, ist es hinreichend, eine untere Schranke für $\mathbb{P}_\theta(|\hat\theta-\theta|\le\varepsilon)$ zu konstruieren, die für $n\to\infty$ gegen 1 konvergiert. Es gilt

$$\begin{aligned}\Big\{|\hat\theta-\theta|<\varepsilon\Big\} &\overset{\text{Unimod}}{\supseteq} \{L(X_1,\dots,X_n,\theta-\varepsilon)<L(X_1,\dots,X_n,\theta)<L(X_1,\dots,X_n,\theta+\varepsilon)\}\\ &= \left\{\frac{L(X_1,\dots,X_n,\theta)}{L(X_1,\dots,X_n,\theta\pm\varepsilon)}>1\right\} \overset{\delta>0\Longrightarrow e^{n\delta}>1}{\supseteq} \left\{\frac{L(X_1,\dots,X_n,\theta)}{L(X_1,\dots,X_n,\theta\pm\varepsilon)}>e^{n\delta}\right\}\\ &= \left\{\frac{1}{n}\log\frac{L(X_1,\dots,X_n,\theta)}{L(X_1,\dots,X_n,\theta\pm\varepsilon)}>\delta\right\} = A_+\cap A_-\,,\end{aligned}$$

wobei

$$A_\pm = \left\{\frac{1}{n}\log\frac{L(X_1,\dots,X_n,\theta)}{L(X_1,\dots,X_n,\theta\pm\varepsilon)}>\delta\right\}.$$

Somit gilt also

$$\mathbb{P}_\theta\Big(|\hat\theta-\theta|<\varepsilon\Big)\ge\mathbb{P}_\theta(A_+\cap A_-)=\mathbb{P}_\theta(A_+)+\mathbb{P}_\theta(A_-)-\mathbb{P}_\theta(A_+\cup A_-)\,.$$

Wenn wir zeigen können, dass

$$\lim_{n\to\infty}\mathbb{P}_\theta(A_\pm)=1\,, \tag{3.4.2}$$

dann folgt daraus

$$1\ge\lim_{n\to\infty}\mathbb{P}_\theta(A_+\cup A_-)\ge\lim_{n\to\infty}\mathbb{P}_\theta(A_\pm)=1 \quad\Longrightarrow\quad \lim_{n\to\infty}\mathbb{P}_\theta(A_+\cup A_-)=1$$

und

$$1\ge\lim_{n\to\infty}\mathbb{P}_\theta\Big(|\hat\theta-\theta|<\varepsilon\Big)\ge 1+1-1=1\,,$$

womit folgt, dass

$$\lim_{n\to\infty}\mathbb{P}_\theta\Big(|\hat\theta-\theta|>\varepsilon\Big)=1-\underbrace{\lim_{n\to\infty}\mathbb{P}_\theta\Big(|\hat\theta-\theta|<\varepsilon\Big)}_{=1}=0$$

und somit $\lim_{n\to\infty}\mathbb{P}(|\hat\theta-\theta|>\varepsilon)\le\lim_{n\to\infty}\mathbb{P}(|\hat\theta-\theta|\ge\varepsilon)=0$, d.h., $\hat\theta\xrightarrow[n\to\infty]{P}\theta$.

Jetzt zeigen wir, dass $\mathbb{P}_\theta(A_+)\xrightarrow[n\to\infty]{}1$ (für $\mathbb{P}_\theta(A_-)$ ist es analog).

1. Sei $H(\mathbb{P}_\theta,\mathbb{P}_{\theta+\varepsilon})<\infty$. Sei

$$f(x)=\begin{cases}\dfrac{L(x,\theta)}{L(x,\theta+\varepsilon)}\,, & \text{falls } L(x,\theta+\varepsilon)>0\,,\\ 1\,, & \text{sonst.}\end{cases}$$

Dann folgt aus Definition 3.4.4, dass $\mathbb{P}_\theta(x:\,L(x,\theta+\varepsilon)>0)=1$. Weiter gilt

$$\begin{aligned}\frac{1}{n}\log\frac{L(X_1,\dots,X_n,\theta)}{L(X_1,\dots,X_n,\theta+\varepsilon)} &= \frac{1}{n}\sum_{i=1}^n\log\frac{L(X_i,\theta)}{L(X_i,\theta+\varepsilon)}=\frac{1}{n}\sum_{i=1}^n\log f(X_i)\\ &\xrightarrow[n\to\infty]{\text{f.s.}}\mathbb{E}_\theta\log f(X_1)=\int L(x,\theta)\cdot\log\frac{L(x,\theta)}{L(x,\theta+\varepsilon)}\,dx=H(\mathbb{P}_\theta,\mathbb{P}_{\theta+\varepsilon})>\delta>0\end{aligned}$$

nach dem starken Gesetz der großen Zahlen, weil $\log f(X_1)\in L^1(\Omega,\mathcal{F},\mathbb{P})$ wegen

$$\mathbb{E}_\theta\log f(X_1)=H(\mathbb{P}_\theta,\mathbb{P}_{\theta+\varepsilon})<\infty \quad\Longrightarrow\quad \mathbb{P}(A_+)\xrightarrow[n\to\infty]{}1\,.$$

2. Sei $H(\mathbb{P}_\theta, \mathbb{P}_{\theta+\varepsilon}) = \infty$ und $\mathbb{P}_\theta(x \in \mathbb{R} : L(x, \theta+\varepsilon) = 0) = 0$, dann folgt

$$f(x) \overset{\text{f.s.}}{=} \frac{L(x,\theta)}{L(x,\theta+\varepsilon)}\,.$$

Es gilt $\log\min\{f(X_1), c\} \in L^1(\Omega, \mathcal{F}, \mathbb{P})$ für alle $c > 0$. Somit folgt wie in Punkt 1:

$$\frac{1}{n}\sum_{i=1}^n \log\min\{f(X_i), c\} \xrightarrow[n\to\infty]{\text{f.s.}} \mathbb{E}_\theta \log\min\{f(X_1), c\} \in (0,\infty) \xrightarrow[c\to\infty]{} H(\mathbb{P}_\theta, \mathbb{P}_{\theta+\varepsilon}) = \infty$$

und damit

$$\begin{aligned} A_+ &\supset \left\{\frac{1}{n}\sum_{i=1}^n \log\min\{f(X_i), c\} > \delta\right\} \\ \Longrightarrow \mathbb{P}(A_+) &\geq \mathbb{P}\left(\frac{1}{n}\sum_{i=1}^n \log\min\{f(X_i), c\} > \delta\right) \xrightarrow[n\to\infty]{} 1\,. \end{aligned}$$

3. Sei $H(\mathbb{P}_\theta, \mathbb{P}_{\theta+\varepsilon}) = \infty$ und $\mathbb{P}_\theta(x \in \mathbb{R} : L(x, \theta+\varepsilon) = 0) = a > 0$, dann folgt

$$\begin{aligned} \mathbb{P}_\theta\left(\frac{1}{n}\log\frac{L(X_1,\dots,X_n,\theta)}{L(X_1,\dots,X_n,\theta+\varepsilon)} = \infty\right) &= 1 - \mathbb{P}(\frac{1}{n}\log\frac{L(X_1,\dots,X_n,\theta)}{L(X_1,\dots,X_n,\theta+\varepsilon)} < \infty) \\ &= 1 - \mathbb{P}\left(\bigcap_{i=1}^n \{L(X_i,\theta+\varepsilon) > 0\}\right) \overset{X_i \text{ u.i.v.}}{=} 1 - (1-a)^n \xrightarrow[n\to\infty]{} 1 \end{aligned}$$

Insgesamt also $\mathbb{P}(A_+) \xrightarrow[n\to\infty]{} 1$.

□

Definition 3.4.5
Sei $X = (X_1, \dots, X_n)$ eine Zufallsstichprobe von unabhängigen identisch verteilten Zufallsvariablen $X_i \sim F_\theta\,, \quad \theta \in \Theta$. Sei $L(x,\theta)$ die Likelihood-Funktion von X_i. Dann heißt der Ausdruck

$$I(\theta) = \mathbb{E}_\theta\left(\frac{\partial}{\partial\theta}\log L(X,\theta)\right)^2, \quad \theta \in \Theta \tag{3.4.3}$$

die *Fisher-Information* der Stichprobe $(X_1, \dots, X_n)$.

Es wird in Zukunft vorausgesetzt, dass $0 < I(\theta) < \infty$. Wir stellen nun einige Bedingungen auf, die für die asymptotische Normalverteiltheit von ML-Schätzern notwendig sind.

1. $\Theta \subset \mathbb{R}$ ist ein offenes Intervall ($m = 1$).

2. Es gelte $\mathbb{P}_\theta \neq \mathbb{P}_{\theta'}$ genau dann, wenn $\theta \neq \theta'$.

3. Die Familie $\{\mathbb{P}_\theta,\, \theta \in \Theta\}$, $\theta \in \Theta$ bestehe nur aus diskreten oder nur aus absolut stetigen Verteilungen, also nicht aus Mischungen von diskreten und absolut stetigen Verteilungen.

4. $B = \operatorname{supp} L(x,\theta) = \{x \in \mathbb{R} : L(x,\theta) > 0\}$ hängt nicht von $\theta \in \Theta$ ab. Dabei heißt supp (von englisch „support“) der „Träger“ einer Funktion f und ist definiert als

$$\operatorname{supp} f = \{x \in \mathbb{R} : \ f(x) \neq 0\}$$

und die Likelihood-Funktion $L(x,\theta)$ ist durch

$$L(x,\theta) = \begin{cases} p(x,\theta)\,, & \text{im diskreten Fall,} \\ f(x,\theta)\,, & \text{im absolut stetigen Fall} \end{cases} \tag{3.4.4}$$

gegeben, wobei $p(x,\theta)$ bzw. $f(x,\theta)$ die Wahrscheinlichkeitsfunktion bzw. Dichte von $\mathbb{P}_\theta$ ist.

5. Die Abbildung $L(x,\theta)$ ist dreimal stetig differenzierbar und es gilt

$$0 = \frac{d^k}{d\theta^k} \int_B L(x,\theta)\,dx = \int_B \frac{\partial^k}{\partial\theta^k} L(x,\theta)\,dx\,, \quad k = 1,2,\ \theta \in \Theta\,.$$

Da das Integral über die Dichte $L(x,\theta)$ gleich 1 ist, ist die Ableitung gleich 0. Dabei sind im diskreten die Integrale durch Summen zu ersetzen.

6. Für alle $\theta_0 \in \Theta$ existiert eine Konstante $\delta_{\theta_0} > 0$ und eine messbare Funktion $g_{\theta_0} : B \to [0,\infty)$, so dass

$$\left| \frac{\partial^3 \log L(x,\theta)}{\partial\theta^3} \right| \leq g_{\theta_0}(x)\,, \qquad \forall x \in B\,, \quad |\theta - \theta_0| < \delta_{\theta_0}\,,$$

wobei $\mathbb{E}_{\theta_0}\, g_{\theta_0}(X_1) < \infty$.

Bemerkung 3.4.3
Es gilt folgende Relation:

$$n \cdot I(\theta) = \operatorname{Var}_\theta \left(\frac{\partial}{\partial\theta} \log L(X_1,\ldots,X_n,\theta) \right)\,,$$

wobei

$$L(X_1,\ldots,X_n,\theta) = \prod_{i=1}^n L(X_i,\theta) \tag{3.4.5}$$

die Likelihood-Funktion der Stichprobe $(X_1,\ldots,X_n)$ ist mit $L(X_i,\theta)$ nach (3.4.4).

Beweis Es gilt

$$\frac{\partial}{\partial\theta} \log L(X_1,\ldots,X_n,\theta) = \frac{\partial}{\partial\theta} \sum_{i=1}^n \log L(X_i,\theta) = \sum_{i=1}^n \frac{\partial}{\partial\theta} \log L(X_i,\theta) = \sum_{i=1}^n \frac{L'(X_i,\theta)}{L(X_i,\theta)}\,.$$

Ferner

$$\mathbb{E}_\theta \left(\frac{\partial}{\partial\theta} L(X_1,\ldots,X_n,\theta) \right) = \sum_{i=1}^n \mathbb{E}_\theta \frac{L'(X_i,\theta)}{L(X_i,\theta)} = \sum_{i=1}^n \int_B \frac{L'(X,\theta)}{L(X,\theta)} \cdot L(X,\theta)\,dx \overset{5)}{=} 0\,.$$

Insgesamt gilt also

$$\begin{aligned}\mathrm{Var}_\theta\left(\frac{\partial}{\partial\theta}\log L(X_1,\dots,X_n,\theta)\right) &= \mathrm{Var}_\theta\left(\sum_{i=1}^n \frac{\partial}{\partial\theta}\log L(X_i,\theta)\right)\\ &\overset{X_i \text{ unabhg.}}{=} \sum_{i=1}^n \mathrm{Var}_\theta\left(\frac{\partial}{\partial\theta}\log L(X_i,\theta)\right) \overset{X_i \text{ ident.}}{\underset{\text{vert.}}{=}} n\cdot \mathrm{Var}_\theta\left(\frac{\partial}{\partial\theta}\log L(X_1,\theta)\right)\\ &= n\cdot \mathbb{E}_\theta\left(\frac{\partial}{\partial\theta}\log L(X_1,\theta)\right)^2 = n\cdot I(\theta)\,.\end{aligned}$$

□

Satz 3.4.2
Sei $(X_1,\dots,X_n)$ eine Stichprobe von Zufallsvariablen, für die die Bedingungen 1) bis 6) erfüllt sind und $0 < I(\theta) < \infty\,, \quad \theta\in\Theta$. Falls $\hat\theta(X_1,\dots,X_n)$ ein schwach konsistenter ML-Schätzer für θ ist, dann ist $\hat\theta(X_1,\dots,X_n)$ asymptotisch normalverteilt:

$$\sqrt{n\cdot I(\theta)}\left(\hat\theta(X_1,\dots,X_n)-\theta\right) \xrightarrow[n\to\infty]{d} Y \sim N(0,1)\,.$$

Beweis Führen wir die Bezeichnung $l_n(\theta) = \log L(X_1,\dots,X_n,\theta)\,, \quad \theta\in\Theta$ ein. Sei

$$l_n^{(k)}(\theta) = \frac{d^k}{d\theta^k} l_n(\theta)\,, \qquad k=1,2,3\,.$$

Ist $\hat\theta$ ein ML-Schätzer, so folgt $l_n^{(1)}(\hat\theta)=0$. Schreiben wir die Taylor-Entwicklung von $l_n^{(1)}(\hat\theta)$ in der Umgebung von θ auf:

$$0 = l_n^{(1)}(\hat\theta) = l_n^{(1)}(\theta) + (\hat\theta-\theta)\cdot l_n^{(2)}(\theta) + (\hat\theta-\theta)^2\cdot\frac{l_n^{(3)}(\theta^*)}{2}\,,$$

wobei θ^* zwischen θ und $\hat\theta$ liegt. Dabei ist

$$-(\hat\theta-\theta)\left(l_n^{(2)}(\theta) + (\hat\theta-\theta)\frac{l_n^{(3)}(\theta^*)}{2}\right) = l_n^{(1)}(\theta) \Longrightarrow \sqrt{n}(\hat\theta-\theta) = \frac{\frac{l_n^{(1)}(\theta)}{\sqrt{n}}}{-\frac{l_n^{(2)}(\theta)}{n} - (\hat\theta-\theta)\frac{l_n^{(3)}(\theta^*)}{2n}}$$

Falls wir zeigen können, dass

1. $$\frac{l_n^{(1)}(\theta)}{\sqrt{n}} \xrightarrow[n\to\infty]{d} N(0,I(\theta))\,,$$

2. $$-\frac{l_n^{(2)}(\theta)}{n} \xrightarrow[n\to\infty]{\text{f.s.}} I(\theta)\,,$$

3. $$(\hat\theta-\theta) \xrightarrow[n\to\infty]{P} 0 \quad \text{und} \quad \frac{l_n^{(3)}(\theta^*)}{2n}$$

   beschränkt ist, das heißt

$$\exists c>0: \qquad \lim_{n\to\infty}\mathbb{P}_\theta\left(\left|\frac{l_n^{(3)}(\theta^*)}{2n}\right| < c\right) = 1\,,$$

dann konvergiert der Ausdruck

$$(\hat{\theta}-\theta)\cdot\frac{l_n^{(3)}(\theta^*)}{2n}\xrightarrow[n\to\infty]{P}0\,,\ \text{weil}\ \left|\frac{l_n^{(3)}(\theta^*)}{n}\right|\le g_\theta(X_1)\ \text{integrierbar}$$

und somit gilt

$$\sqrt{n}(\hat{\theta}-\theta)=\frac{\frac{l_n^{(1)}(\theta)}{\sqrt{n}}}{-\frac{l_n^{(2)}(\theta)}{n}-(\hat{\theta}-\theta)\frac{l_n^{(3)}(\theta^*)}{2n}}\xrightarrow[n\to\infty]{d}Z_1\sim N\left(0,\frac{1}{I(\theta)}\right)$$

nach dem Satz von Slutsky. Damit folgt $\sqrt{n}\sqrt{I(\theta)}(\hat{\theta}-\theta)\xrightarrow[n\to\infty]{d}Y\sim N(0,1)$

1. Es gilt
$$\frac{l_n^{(1)}}{\sqrt{n}}=\frac{\sum_{i=1}^n\frac{\partial}{\partial\theta}\log L(X_i,\theta)}{\sqrt{n}}\xrightarrow[n\to\infty]{}Y_1\sim N\Bigg(0,\underbrace{\mathrm{Var}_\theta\Big(\frac{\partial}{\partial\theta}L(X_i,\theta)\Big)}_{=I(\theta)}\Bigg)$$
nach dem zentralen Grenzwertsatz, weil $\frac{\partial}{\partial\theta}\log L(X_i,\theta)$ unabhängig identisch verteilte Zufallsvariablen mit Erwartungswert 0 (siehe Bemerkung 3.4.3) sind.

2.
$$\begin{aligned}-\frac{1}{n}l_n^{(2)}(\theta)&=-\frac{1}{n}\sum_{i=1}^n\frac{\partial^2}{\partial\theta^2}\log L(X_i,\theta)=\frac{1}{n}\sum_{i=1}^n\frac{\left(L^{(1)}(X_i,\theta)\right)^2-L(X_i,\theta)\cdot L^{(2)}(X_i,\theta)}{(L(X_i,\theta))^2}\\&=\frac{1}{n}\sum_{i=1}^n\left(\frac{L^{(1)}(X_i,\theta)}{L(X_i,\theta)}\right)^2-\frac{1}{n}\sum_{i=1}^n\frac{L^{(2)}(X_i,\theta)}{L(X_i,\theta)}\\&\xrightarrow[n\to\infty]{\text{f.s.}}\mathbb{E}_\theta\left(\frac{L^{(1)}(X_1,\theta)}{L(X_1,\theta)}\right)^2-\mathbb{E}_\theta\left(\frac{L^{(2)}(X_1,\theta)}{L(X_1,\theta)}\right)=I(\theta)-0=I(\theta)\end{aligned}$$
nach dem Gesetz der großen Zahlen, wobei
$$L^{(k)}(X_i,\theta)=\frac{\partial^k}{\partial\theta^k}L(X_i,\theta)$$
und
$$\mathbb{E}_\theta\left(\frac{L^{(2)}(X_1,\theta)}{L(X_1,\theta)}\right)=\int_B\frac{\partial^2}{\partial\theta^2}L(x,\theta)\,dx\overset{5)}{=}\frac{d^2}{d\theta^2}\int_B L(x,\theta)\,dx=0\,.$$

3. $\hat{\theta}\xrightarrow[n\to\infty]{P}\theta$, weil $\hat{\theta}$ schwach konsistent ist. Zeigen wir, dass
$$\frac{l_n^{(3)}(\theta^*)}{n}(\hat{\theta}-\theta)\xrightarrow[n\to\infty]{P}0\,.$$
Aus $\hat{\theta}\xrightarrow[n\to\infty]{P}\theta$ folgt für alle $\varepsilon>0$
$$\mathbb{P}\left(|\hat{\theta}-\theta|\le\varepsilon\right)\xrightarrow[n\to\infty]{}1\,.$$

Damit folgt, dass mit asymptotisch großer Wahrscheinlichkeit $|\hat{\theta} - \theta| \leq \delta\,, \quad \delta > 0$ gilt, welches aus der Bedingung 6) folgt. Damit gilt, dass für alle $\theta : |\hat{\theta} - \theta| < \delta$

$$\left|\frac{l_n^{(3)}(\theta^*)}{n}\right| \leq \frac{1}{n}\sum_{i=1}^{n} \underbrace{\left|\frac{\partial^3}{\partial\theta^3}\log L(X_i,\theta)\right|}_{\leq g_\theta(X_i)} \leq \frac{1}{n}\sum_{i=1}^{n} g_\theta(X_i) \xrightarrow[n\to\infty]{\text{f.s.}} \mathbb{E}_\theta\, g_\theta(X_1) < \infty\,.$$

So folgt, dass eine Konstante $c > 0$ existiert, sodass

$$\mathbb{P}_\theta\left(\left|\frac{l_n^{(3)}(\theta^*)}{n}\right| < c\right) \xrightarrow[n\to\infty]{} 1 \quad \text{und somit} \quad \frac{l_n^{(3)}(\theta^*)}{n}(\hat{\theta} - \theta) \xrightarrow[n\to\infty]{P} 0\,.$$

Der Beweis ist beendet.

□

3.4.3 Bayes-Schätzer

Sei $(X_1, \ldots, X_n)$ eine Zufallsstichprobe, wobei X_i unabhängige identisch verteilte Zufallsvariablen mit Verteilungsfunktion $F_\theta\,, \quad \theta \in \Theta$ sind. Sei F_θ entweder eine diskrete oder eine absolut stetige Verteilung. Sei aber auch θ eine Zufallsvariable $\tilde{\theta}$ mit Verteilung $Q(\cdot)$ auf dem Messraum $(\Theta, \mathcal{B}_\Theta)$, die entweder diskret mit Zähldichte $q(\cdot)$ oder absolut stetig mit Dichte $q(\cdot)$ ist. Nach wie vor werden beide Fälle gemeinsam betrachtet, dabei entsprechen sich die Summation und Integration im diskreten bzw. absolut stetigen Fall.

Definition 3.4.6
Die Verteilung $Q(\cdot)$ heißt *a-priori-Verteilung* des Parameters θ (von $\tilde{\theta}$) (a-priori bedeutet hier „vor dem Experiment $(X_1, \ldots, X_n)$“).

Definition 3.4.7
Die *a-posteriori-Verteilung* des Parameters θ (von $\tilde{\theta}$) ist gegeben durch die (Zähl-)Dichte

$$q_{X_1,\ldots,X_n}(\theta) = \begin{cases} \mathbb{P}(\tilde{\theta} = \theta \,|\, X_1 = x_1, \ldots, X_n = x_n)\,, & \text{falls die Verteilung } Q \text{ diskret ist,} \\ f_{\tilde{\theta}|X_1,\ldots,X_n}(\theta, x_1, \ldots, x_n)\,, & \text{falls die Verteilung } Q \text{ absolut stetig ist.} \end{cases}$$

Dabei ist

$$\begin{aligned} \mathbb{P}(\tilde{\theta} = \theta \,|\, X = x_1, \ldots, X = x_n) &= \frac{\mathbb{P}(\tilde{\theta} = \theta, X_1 = x_1, \ldots, X_n = x_n)}{\mathbb{P}(X_1 = x_1, \ldots, X_n = x_n)} \\ &= \frac{\mathbb{P}_\theta(X_i = x_i\,,\, i = 1, \ldots, n) \cdot q(\theta)}{\sum_{\theta_1\in\Theta} \mathbb{P}_{\theta_1}(X_i = x_i,\, i = 1, \ldots, n) \cdot q(\theta_1)} \end{aligned}$$

die *Bayesche Formel*, bzw.

$$f_{\tilde{\theta}|X_1,\ldots,X_n}(\theta, x_1, \ldots, x_n) = \frac{f_{(\tilde{\theta},X_1,\ldots,X_n)}(\theta, x_1, \ldots, x_n)}{f_{X_1,\ldots,X_n}(x_1, \ldots, x_n)} = \frac{L(x_1, \ldots, x_n, \theta) \cdot q(\theta)}{\int_\Theta L(x_1, \ldots, x_n, \theta_1) \cdot q(\theta_1)\, d\theta_1}\,,$$

mit $L(x_1, \ldots, x_n, \theta)$ nach (3.4.5).

Definition 3.4.8
Eine *Verlustfunktion* $V : \Theta^2 \to \mathbb{R}_+$ ist eine Θ^2-messbare Funktion.

Verlustfunktionen spielen in unseren Betrachtungen folgende Rolle: $\mathbb{E}_* V(\tilde{\theta}, a)$ stellt den *erwarteten Verlust* (mittleres Risiko) dar, der bei der Schätzung des Parameters θ durch a entsteht. Dabei stellt $\mathbb{E}_*$ den Erwartungswert bezüglich der *a-posteriori-Verteilung* von $\tilde{\theta}$ dar. Es sind offensichtlich die konkreten Stichprobenwerte $x_1, \ldots, x_n$ in die a-posteriori-Verteilung eingegangen, deshalb ist $\mathbb{E}_* V(\tilde{\theta}, a)$ eine Funktion von a und $x_1, \ldots, x_n$:

$$\mathbb{E}_* V(\tilde{\theta}, a) = \varphi(x_1, \ldots, x_n, a)\,.$$

Definition 3.4.9
Ein Schätzer $\hat{\theta}$ heißt *Bayes-Schätzer* des Parameters θ, falls

$$\hat{\theta}(x_1, \ldots, x_n) = \arg\min_a \mathbb{E}_* V(\tilde{\theta}, a) \tag{3.4.6}$$

existiert und eindeutig ist.

Bemerkung 3.4.4

1. Manchmal gilt $\hat{\theta} \notin \Theta$, was mit der Existenz des Minimums von $\varphi(x_1, \ldots, x_n, a)$ auf Θ zu tun hat.

2. Der Name „Bayesscher Ansatz“ stammt von dem englischen Mathematiker Thomas Bayes (1702–1761), der die Bayessche Formel

$$\mathbb{P}(B_i|A) = \frac{\mathbb{P}(A|B_i) \cdot \mathbb{P}(B_i)}{\sum_j \mathbb{P}(A|B_j) \cdot \mathbb{P}(B_j)} \tag{3.4.7}$$

   nur ideenhaft eingeführt hat. Der eigentliche Entdecker der Formel (3.4.7) ist Pierre-Simon Laplace (1749–1827) (Ende des XVIII. Jahrhunderts). Diese Formel wurde bei der Herleitung der *a-posteriori-Verteilung* von $\tilde{\theta}$ implizit benutzt.

3. Die Vorgehensweise in Definition 3.4.9 ist in konkreten praxisrelevanten Fällen meistens nur numerisch möglich. Es gibt sehr wenige Beispiele für analytische Lösungen des in (3.4.6) gestellten Minimierungsproblems.

Beispiel 3.4.3 (Quadratische Verlustfunktion):
Ist $V(\theta_1, \theta_2) = (\theta_1 - \theta_2)^2$, so ist

$$\min_a (\varphi(x_1, \ldots, x_n, a)) = \min_a \left(\mathbb{E}_*(\tilde{\theta} - a)^2\right) = \min_a \left(\mathbb{E}_* \tilde{\theta}^2 - 2a\mathbb{E}_* \tilde{\theta} + a^2\right) = \mathbb{E}_* \tilde{\theta}$$

und daher der *Bayes-Schätzer* $\hat{\theta}(x_1, \ldots, x_n)$ für θ durch $\mathbb{E}_* \tilde{\theta}$ gegeben.

Beispiel 3.4.4 (Bernoulli-Verteilung):
Sei $(X_1, \ldots, X_n)$ eine unabhängig identisch verteilte Stichprobe von $X_i \sim Bernoulli(p)$, $p \in (0, 1)$. Weiter sei die a-priori-Verteilung

$$\tilde{p} \sim Beta(\alpha, \beta)\,, \quad \alpha, \beta > 0, \text{ mit Zähldichte} \quad q(p) = \frac{p^{\alpha-1}(1-p)^{\beta-1}}{B(\alpha, \beta)} \cdot \mathbb{I}(p \in [0, 1])\,,$$

die a-posteriori-Verteilung von $\tilde{p}$ ist dann gleich

$$q^*(p) = f_{\tilde{p}|X_1=x_1,\ldots,X_n=x_n}(p) = \frac{\mathbb{P}_p(X_1 = x_1, \ldots, X_n = x_n) \cdot q(p)}{\int_0^1 \mathbb{P}_{p_1}(X_1 = x_1, \ldots, X_n = x_n) \cdot q(p_1)\, dp_1}\,.$$

Es ist immer möglich die a-posteriori-Verteilung nicht bezüglich des Vektors $(X_1, \ldots, X_n)$, sondern bezüglich einer Funktion $g(X_1, \ldots, X_n)$, zu berechnen (*Komplexitätsreduktion*).

Hier ist $Y = g(X_1, \ldots, X_n) = \sum_{i=1}^n X_i$ die Gesamtanzahl aller Erfolge in n Experimenten, wobei

$$X_i = \begin{cases} 1\,, & \text{mit Wahrscheinlichkeit } p\,, \\ 0\,, & \text{sonst.} \end{cases}$$

Daher gilt für die a-posteriori-Verteilung bzgl. Y:

$$\begin{aligned} q^*(p) = f_{\tilde{p}|Y=k}(p) &= \frac{\mathbb{P}_p(Y=k)\cdot q(p)}{\int_0^1 \mathbb{P}_{p_1}(Y=k) q(p_1)\,dp_1} \\ &\underset{\substack{\text{falls } \tilde{p}=p}}{\overset{Y\sim Bin(n,p),}{=}} \frac{\binom{n}{k} p^k (1-p)^{n-k} \cdot (B(\alpha,\beta))^{-1} \cdot p^{\alpha-1}(1-p)^{\beta-1}}{\frac{\binom{n}{k}}{B(\alpha,\beta)} \cdot \int_0^1 p_1^{k+\alpha-1}(1-p_1)^{n-k+\beta-1}\,dp_1} \\ &= \frac{p^{k+\alpha-1}(1-p)^{n-k+\beta-1}}{B(k+\alpha, n-k+\beta)}\,, \qquad p \in [0,1]\,. \end{aligned}$$

Daher ist die a-posteriori-Verteilung von $\tilde{p}$ unter der Bedingung $Y = k$ durch

$$Beta(k+\alpha, n-k+\beta)$$

gegeben.

Für den *Bayes-Schätzer* gilt:

$$\begin{aligned} \hat{p}(x_1, \ldots, x_n) = \mathbb{E}_* \tilde{p} = \int_0^1 p \cdot q^*(p)\,dp &= \frac{\int_0^1 p^{k+\alpha}(1-p)^{n-k+\beta-1}\,dp}{B(k+\alpha, n-k+\beta)} \\ = \frac{B(k+\alpha+1, n-k+\beta)}{B(k+\alpha, n-k+\beta)} = \ldots = \frac{k+\alpha}{\alpha+\beta+n} &= \frac{\sum_{i=1}^n x_i + \alpha}{\alpha+\beta+n} = \frac{\alpha + n\bar{x}_n}{\alpha+\beta+n}\,. \end{aligned}$$

Interpretation:

$$\hat{p}(X_1, \ldots, X_n) = \underbrace{\frac{n}{\alpha+\beta+n}}_{=:c_1} \bar{X}_n + \underbrace{\frac{\alpha+\beta}{\alpha+\beta+n}}_{=:c_2} \cdot \frac{\alpha}{\alpha+\beta} = c_1 \cdot \bar{X}_n + c_2 \cdot \mathbb{E}_{apr}\tilde{\theta}\,,$$

wobei $c_1 + c_2 = 1$ ist. Dies heißt, dass die Bayessche Methode einen Mittelweg zwischen dem Schätzer $\mathbb{E}_{apr}\tilde{\theta}$ (in Abwesenheit der Information über die Stichprobe $(X_1, \ldots, X_n)$) und dem M-Schätzer $\bar{X}_n$ (in Abwesenheit der a-priori-Information über die Verteilung von $\tilde{p}$) für p einschlägt.

3.4.4 Resampling-Methoden zur Gewinnung von Punktschätzern

Sei $(X_1, \ldots, X_n)$ eine Stichprobe im parametrischen Modell. Gesucht ist ein Schätzer $\hat{\theta}$ für den Parameter θ. Um diesen Schätzer zu konstruieren, werden bei Resampling-Methoden neue Stichproben $(X_1^*, \ldots, X_n^*)$ durch das unabhängige Ziehen mit Zurücklegen aus der alten Stichprobe $(X_1, \ldots, X_n)$ generiert und auf ihrer Basis Mittelwerte, Stichprobenvarianzen und andere Schätzer gebildet. Dabei ist die Dimension m des Parameterraums Θ beliebig.

Wir werden im Folgenden die *Resampling*-Methoden

1. *Jackknife* (dt. „Taschenmesser“, weist auf Mittel, die jedem immer zur Hand sein sollten)

2. *Bootstrap* (engl. „self-sufficient“, dt. „mit eigenen Ressourcen“)

betrachten.

1. *Jackknife-Methoden zur Schätzung der Varianz bzw. der Verzerrung von Schätzern:*

   Als einführendes Beispiel betrachten wir $\theta = \mathbb{E}X = \mu$ bzw. $\theta = \operatorname{Var} X = \sigma^2$ und ihre (erwartungstreue) Schätzer $\hat{\mu} = \bar{X}_n$ bzw. $\hat{\sigma}^2 = S_n^2$.

   Wie wir bereits wissen, gilt

$$\operatorname{Var}\hat{\mu} = \frac{\sigma^2}{n}\,, \qquad \operatorname{Var}\hat{\sigma}^2 = \frac{1}{n}\left(\mu_4' - \frac{n-3}{n-1}\sigma^4\right)\,.$$

   Nun ist ein Schätzer für die Varianz von $\hat{\mu}$ bzw. $\hat{\sigma}^2$ gesucht. Dazu verwenden wir die Plug-in Methode

$$\widehat{\operatorname{Var}}\,\hat{\mu} = \frac{S_n^2}{n}\,, \qquad \widehat{\operatorname{Var}}\,\hat{\sigma}^2 = \frac{1}{n}\left(\hat{\mu}_4' - \frac{n-3}{n-1}\hat{\sigma}^4\right)\,,$$

   wobei $\hat{\mu}_4'$ das vierte zentrierte empirische Moment ist.

   Im Allgemeinen sind jedoch *keine* Formeln von $\operatorname{Var}\hat{\theta}$ bekannt. Hier kommt nun die *Jackknife*-Methode zum Einsatz:

   - Sei $X_{[i]}$ die Stichprobe $(X_1, \ldots, X_{i-1}, X_{i+1}, \ldots, X_n)\,, \quad i = 1, \ldots, n$. Falls

$$\hat{\theta}(X_1, \ldots, X_n) = \varphi_n(X_1, \ldots, X_n)\,,$$

     so bilden wir

$$\hat{\theta}_{[i]} = \varphi_{n-1}(X_{[i]})\,, \quad \bar{\theta}_{[\cdot]} = \frac{1}{n}\sum_{i=1}^n \hat{\theta}_{[i]}\,, \quad \widehat{\operatorname{Var}}_{jn}(\hat{\theta}) \overset{\text{def.}}{=} \frac{n-1}{n}\sum_{i=1}^n \left(\hat{\theta}_{[i]} - \bar{\theta}_{[\cdot]}\right)^2\,.$$

     Definition 3.4.10
     Der Schätzer $\bar{\theta}_{[\cdot]}$ bzw. $\widehat{\operatorname{Var}}_{jn}(\hat{\theta})$ heißt *Jackknife-Schätzer* für den Erwartungswert bzw. die Varianz des Schätzers $\hat{\theta}$ von θ.

     Beispiel 3.4.5
     Sei $\theta = \mu\,, \quad \hat{\theta} = \hat{\mu} = \bar{X}_n$, so gilt

$$\varphi_n(x_1, \ldots, x_n) = \frac{1}{n}\sum_{i=1}^n x_i\,,$$

     womit folgt, dass

$$\hat{\theta}_{[i]} = \frac{1}{n-1}\sum_{j\neq i} X_j = \frac{1}{n-1}\left(-X_i + \sum_{j=1}^n X_j\right) = \frac{n}{n-1}\bar{X}_n - \frac{1}{n-1}X_i\,, \quad \forall i = 1, \ldots, n\,,$$

$$\bar{\theta}_{[\cdot]} = \frac{1}{n}\sum_{i=1}^n \hat{\theta}_{[i]} = \frac{n}{n-1}\bar{X}_n - \frac{1}{n(n-1)}\sum_{i=1}^n X_i = \frac{n\cdot\bar{X}_n}{n-1} - \frac{\bar{X}_n}{n-1} = \frac{n-1}{n-1}\bar{X}_n = \bar{X}_n\,.$$

     Daher ist ein *Jackknife-Schätzer für* μ gleich $\bar{X}_n$.

Konstruieren wir nun einen *Jackknife-Schätzer der Varianz*:

$$\widehat{\mathrm{Var}}_{jn}(\hat{\theta}) = \frac{n-1}{n}\sum_{i=1}^{n}\left(\frac{n}{n-1}\bar{X}_n - \frac{1}{n-1}X_i - \bar{X}_n\right)^2 = \frac{n-1}{n}\sum_{i=1}^{n}\left(\frac{1}{n-1}(\bar{X}_n - X_i)\right)^2$$
$$= \frac{n-1}{n(n-1)^2}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2 = \frac{1}{n}S_n^2\,,$$

wobei dies genau der Plug-in Schätzer der Varianz von $\hat{\mu}$ ist.

- *Jackknife-Schätzer für die Verzerrung eines Schätzers*
  Sei $\hat{\theta}(X_1,\ldots,X_n)$ ein Schätzer für θ. Der Bias von $\hat{\theta}$ ist $\mathbb{E}_\theta\hat{\theta} - \theta = \mathrm{Bias}(\hat{\theta})$.

  Definition 3.4.11
  Ein *Jackknife-Schätzer der Verzerrung (Bias)* von $\hat{\theta}$ ist durch
  $$\widehat{\mathrm{Bias}}_{jn}(\hat{\theta}) = (n-1)(\bar{\theta}_{[\cdot]} - \hat{\theta})$$
  gegeben.

  An folgenden Beispielen wird klar, dass der oben beschriebene Vorgang zur Verringerung der Verzerrung beiträgt:
  Der *Schätzer*
  $$\tilde{\theta} = \hat{\theta} - \widehat{\mathrm{Bias}}_{jn}(\hat{\theta}) = n\hat{\theta} - (n-1)\bar{\theta}_{[\cdot]} \tag{3.4.8}$$
  hat in der Regel einen *kleineren Bias* als $\hat{\theta}$. Dabei ist wiederum
  $$\hat{\theta}_{[i]} = \varphi_{n-1}(X_{[i]}) \quad \text{und} \quad \bar{\theta}_{[\cdot]} = \frac{1}{n}\sum_{i=1}^{n}\hat{\theta}_{[i]} \quad \text{mit} \quad \hat{\theta}(X_1,\ldots,X_n) = \varphi_n(X_1,\ldots,X_n)\,.$$

  Beispiel 3.4.6

  a) Ist $\theta = \mathbb{E}X_i = \mu$, so ist $\hat{\theta} = \bar{X}_n$ ein unverzerrter Schätzer für μ. Was ist der Bias-korrigierte Schätzer $\tilde{\mu}$? (Dieser sollte schließlich nicht schlechter werden!)
  Es gilt $\bar{\theta}_{[\cdot]} = \bar{X}_n$, daher ist der Bias-Schätzer von Jackknife $\widehat{\mathrm{Bias}}_{jn}(\hat{\theta}) = (n-1)(\bar{X}_n - \bar{X}_n) = 0$ und somit $\tilde{\theta} = \hat{\theta} - 0 = \bar{X}_n$. Wir haben also gesehen, dass die Jackknife-Methode die unverzerrten Schätzer (zumindest in diesem Beispiel) richtig behandelt, indem sie keinen zusätzlichen Bias einbaut.

  b) $\theta = \sigma^2 = \mathrm{Var}X_i$, $\hat{\theta} = \hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2$ ein verzerrter M-Schätzer der Varianz. Was ist $\tilde{\theta}$ in diesem Fall?

  Übungsaufgabe 3.4.1
  Zeigen Sie, dass $\tilde{\theta} = S_n^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2 = \frac{n}{n-1}\hat{\sigma}^2$ ein erwartungstreuer Schätzer der Varianz ist. Somit wurde der Bias von $\hat{\sigma}^2$ durch die Anwendung der Jackknife-Methode vollständig beseitigt.
  Beweisidee: Zeigen Sie hierzu zunächst, dass
  $$\widehat{\mathrm{Bias}}_{jn}(\hat{\theta}) = -\frac{1}{n(n-1)}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2\,.$$

  Bemerkung 3.4.5
  Die Beispiele 3.4.6 a), b), in denen sich der Jackknife-Schätzer analytisch bestimmen ließ, sind eher eine Ausnahme als die Regel. In den meisten Fällen erfolgt die Bias-Reduktion mit Hilfe der Monte-Carlo-Methoden auf Basis der Formel (3.4.8).

2. *Bootstrap-Schätzer:*

Die Bootstrap-Methode besteht in dem Erzeugen einer neuen Stichprobe $(X_1^*, \ldots, X_n^*)$, die aus einer approximativen Verteilung $\hat{F}$ der Stichprobenvariablen X_i, $i = 1, \ldots, n$ gewonnen wird. Seien $\mathbb{E}_*$ und Var_* die wahrscheinlichkeitstheoretischen Größen, die auf dem Verteilungsgesetz $\mathbb{P}_*$ der neuen Stichprobe $(X_1^*, \ldots, X_n^*)$ beruhen. Dabei gibt es folgende Möglichkeiten, $\hat{F}$ zu konstruieren:

i) $\hat{F}(x) = \hat{F}_n(x)$ die empirische Verteilungsfunktion von X_i, falls X_i unabängig identisch verteilt sind.

ii) $\hat{F}$ ist ein parametrischer Schätzer von F, der parametrischen Verteilungsfunktion von X_i. Das heißt, falls $X_i \sim F_\theta\,, \quad i = 1, \ldots, n$ für ein $\theta \in \Theta$ und $\hat{\theta} = \hat{\theta}(X_1, \ldots, X_n)$ ein Schätzer für θ ist, so setzen wir $\hat{F} = F_{\hat{\theta}}$ (Plug-in Methode).

Definition 3.4.12
Ein *Bootstrap-Schätzer* für den *Erwartungswert* (bzw. *Bias* oder *Varianz*) von Schätzer $\hat{\theta}(X_1, \ldots, X_n)$ ist gegeben durch

a) $\hat{\mathbb{E}}_{boot}(\hat{\theta}) = \mathbb{E}_*\hat{\theta}(X_1^*, \ldots, X_n^*)$.
b) $\widehat{\mathrm{Bias}}_{boot}(\hat{\theta}) = \hat{\mathbb{E}}_{boot}\hat{\theta} - \hat{\theta}$.
c) $\widehat{\mathrm{Var}}_{boot}(\hat{\theta}) = \mathrm{Var}_*(\hat{\theta}(X_1^*, \ldots, X_n^*))$.

Beispiel 3.4.7
Sei $\theta = \mu = \mathbb{E}X_i$ und $\hat{F} = \hat{F}_n$ die empirische Verteilungsfunktion. Wie generiert man eine Stichprobe $X_1^*, \ldots, X_n^*$, wobei $X_i^* \sim \hat{F}_n$?

$\hat{F}_n$ gewichtet jede Beobachtung x_i der ursprünglichen Stichprobe mit dem Gewicht $1/n$, deshalb genügt es, einen der Einträge $(x_1, \ldots, x_n)$ auszuwählen (mit Wahrscheinlichkeit $1/n$, Urnenmodell „Ziehen mit Zurücklegen“), um X_j^*, $j = 1, \ldots, n$ zu generieren.

Bootstrap-Schätzer für den Erwartungswert von $\hat{\mu} = \bar{X}_n$:

$$\hat{\mathbb{E}}_{boot}\hat{\mu} = \mathbb{E}_*\left(\frac{1}{n}\sum_{i=1}^n X_i^*\right) \overset{X_i^* \text{ u.i.v.}}{=} \frac{1}{n} \cdot n\mathbb{E}_*(X_1^*) = \int x\, d\hat{F}_n(x) = \frac{1}{n}\sum_{i=1}^n X_i = \bar{X}_n\,.$$

Somit folgt $\widehat{\mathrm{Bias}}_{boot}\hat{\mu} = 0$.

$$\widehat{\mathrm{Var}}_{boot}(\hat{\mu}) = \mathrm{Var}_*\left(\frac{1}{n}\sum_{i=1}^n X_i^*\right) \overset{X_i^* \text{ u.i.v.}}{=} \frac{1}{n^2} \cdot n \cdot \mathrm{Var}_*(X_1^*) = \frac{1}{n} \cdot \frac{1}{n}\sum_{i=1}^n (X_i - \bar{X}_n)^2 = \frac{\hat{\sigma}^2}{n}\,,$$

ein Plug-in Schätzer für $\mathrm{Var}\bar{X}_n = \sigma^2/n$.

Monte-Carlo-Methoden zur numerischen Berechnung von Bootstrap-Schätzern:

Was kann man tun, wenn keine expliziten Formeln für z.B. $\widehat{\mathrm{Var}}_{Boot}(\hat{\theta})$ vorliegen (der Regelfall in der Statistik)?

Generiere M unabhängige Stichproben $(X_{i1}^*, \ldots, X_{in}^*)$, $i = 1, \ldots, M$ nach der Regel i) oder ii) mit Hilfe der Monte-Carlo-Simulation. Dann berechne

$$\hat{\theta}_i = \hat{\theta}(X_{i1}^*, \ldots, X_{in}^*)\,, \quad i = 1, \ldots, M \quad \text{und setze} \quad \hat{\mathbb{E}}_{boot}\hat{\theta} \approx \frac{1}{M}\sum_{i=1}^M \hat{\theta}_i\,.$$

Ähnlich gewinnt man approximative Bootstrap-Schätzer für $\operatorname{Bias}\hat{\theta}$ und $\operatorname{Var}\hat{\theta}$:

$$\widehat{\operatorname{Bias}}_{boot}\hat{\theta} \approx \hat{\mathbb{E}}_{boot}\hat{\theta} - \hat{\theta}\,, \quad \widehat{\operatorname{Var}}_{boot}\hat{\theta} \approx \frac{1}{M-1}\sum_{i=1}^{M}\left(\hat{\theta}_i - \hat{\mathbb{E}}_{boot}\hat{\theta}\right)^2.$$

Mehr sogar, man kann die Verteilungsfunktion von X_{ij}^* durch die empirische Verteilungsfunktion bestimmen:

$$\hat{F}_{boot}(x) = \frac{1}{M}\sum_{i=1}^{M}\frac{1}{n}\sum_{j=1}^{n}\mathbb{I}(X_{ij}^* \le x)\,, \qquad x \in \mathbb{R}\,.$$

Ferner lassen sich mit Hilfe von oben genannten Methoden *Bootstrap-Konfidenzintervalle* für $\hat{\theta}$ ableiten:

Dafür lassen sich Quantile von $\hat{F}_{boot}(x)$ empirisch bestimmen. Damit gilt

$$\mathbb{P}\left(\hat{F}_{boot}^{-1}(\alpha_1) \le \hat{\theta}(X_1^*, \ldots, X_n^*) \le \hat{F}_{boot}^{-1}(\alpha_2)\right) \approx 1 - \alpha_1 - \alpha_2 = 1 - \alpha\,,$$

wobei $\alpha = \alpha_1 + \alpha_2$ klein ist. Beachte dabei, dass man hofft, dass X_i^* sehr ähnlich verteilt ist wie X_i und somit

$$\mathbb{P}\left(\hat{F}_{boot}^{-1}(\alpha_1) \le \hat{\theta}(X_1, \ldots, X_n) \le \hat{F}_{boot}^{-1}(\alpha_2)\right) \approx 1 - \alpha_1 - \alpha_2 = 1 - \alpha$$

gilt.

3.5 Weitere Güteeigenschaften von Punktschätzern

3.5.1 Ungleichung von Cramér-Rao

Sei $(X_1, \ldots, X_n)$ eine Stichprobe von unabhängigen identisch verteilten Zufallsvariablen X_i mit Verteilungsfunktion F_θ, $\theta \in \Theta$. Sei $\hat{\theta}(X_1, \ldots, X_n)$ ein Schätzer für θ. Falls $\hat{\theta}$ erwartungstreu ist, dann misst man die Güte eines anderen erwartungstreuen Schätzers $\tilde{\theta}$ von θ am Wert seiner Varianz. Das bedeutet, falls $\operatorname{Var}_\theta \tilde{\theta} < \operatorname{Var}_\theta \hat{\theta}$, dann ist der Schätzer $\tilde{\theta}$ besser. Wir werden uns nun mit der Frage befassen, ob immer wieder neue, bessere Schätzer $\tilde{\theta}$ mit immer kleinerer Varianz konstruiert werden können. Die Antwort hierauf ist unter gewissen Voraussetzungen negativ. Die untere Schranke der Varianz $\operatorname{Var}_\theta \hat{\theta}$ hierzu liefert der Satz von Cramér-Rao.

Sei $L(x, \theta)$ die Likelihood-Funktion von X_i, d.h.

$$L(x, \theta) = \begin{cases} \mathbb{P}_\theta(x)\,, & \text{im diskreten Fall}\,, \\ f_\theta(x)\,, & \text{im stetigen Fall} \end{cases}$$

und $L(x_1, \ldots, x_n, \theta) = \prod_{i=1}^n L(x_i, \theta)$ die Likelihood-Funktion von der gesamten Stichprobe $(X_1, \ldots, X_n)$. Es gelten die Bedingungen 1) bis 5), die für die asymptotische Normalverteiltheit von ML-Schätzern auf Seite 78 gestellt wurden, wobei die Bedingung 5) für $k = 1$ gilt.

Satz 3.5.1 (Ungleichung von Cramér-Rao):
Sei $\hat{\theta}(X_1, \ldots, X_n)$ ein Schätzer für θ mit den folgenden Eigenschaften:

1. $\mathbb{E}_\theta \hat{\theta}^2(X_1, \ldots, X_n) < \infty \quad \forall\, \theta \in \Theta$.

2. Für alle $\theta \in \Theta$ existiert

$$\frac{d}{d\theta}\mathbb{E}_\theta\hat{\theta}(X_1,\dots,X_n) = \begin{cases} \int_{\mathbb{R}} \hat{\theta}(x_1,\dots,x_n)\frac{\partial}{\partial\theta}L(x_1,\dots,x_n,\theta)\,dx_1\dots dx_n\,, & \text{im abs. ste. Fall,} \\ \sum_{x_1,\dots,x_n} \hat{\theta}(x_1,\dots,x_n)\frac{\partial}{\partial\theta}L(x_1,\dots,x_n,\theta)\,, & \text{im diskr. Fall.} \end{cases}$$

Dann gilt

$$\operatorname{Var}_\theta \hat{\theta}(X_1,\dots,X_n) \geq \frac{\left(\frac{d}{d\theta}\mathbb{E}_\theta\,\hat{\theta}(X_1,\dots,X_n)\right)^2}{n\cdot I(\theta)}\,, \qquad \theta\in\Theta\,,$$

wobei $I(\theta)$ die Fisher-Information aus (3.4.3) ist.

Beweis Führen wir die Funktion

$$\varphi_\theta(x_1,\dots,x_n) = \frac{\partial}{\partial\theta}\log L(x_1,\dots,x_n,\theta)$$

ein. In Bemerkung 3.4.3 haben wir bewiesen, dass

$$\mathbb{E}_\theta\varphi_\theta(X_1,\dots,X_n) = 0\,, \quad \operatorname{Var}_\theta\varphi_\theta(X_1,\dots,X_n) = n\cdot\mathbb{I}(\theta)\,.$$

Wenden wir die Ungleichung von Cauchy-Schwarz auf $\operatorname{Cov}_\theta(\varphi_\theta(X_1,\dots,X_n),\hat{\theta}(X_1,\dots,X_n))$ an:

$$\begin{aligned}\operatorname{Cov}_\theta\Big(\varphi_\theta(X_1,\dots,X_n),\hat{\theta}(X_1,\dots,X_n)\Big) &= \mathbb{E}_\theta\Big(\varphi_\theta(X_1,\dots,X_n)\cdot\hat{\theta}(X_1,\dots,X_n)\Big) - 0 \\ &\leq \sqrt{\operatorname{Var}_\theta\varphi_\theta(X_1,\dots,X_n)}\sqrt{\operatorname{Var}_\theta\hat{\theta}(X_1,\dots,X_n)}\end{aligned}$$

Somit folgt

$$\operatorname{Var}_\theta\hat{\theta}(X_1,\dots,X_n) \geq \frac{\Big(\overbrace{\mathbb{E}_\theta\Big(\varphi_\theta(X_1,\dots,X_n)\cdot\hat{\theta}(X_1,\dots,X_n)\Big)}^{=:A}\Big)^2}{\operatorname{Var}_\theta\varphi_\theta(X_1,\dots,X_n)} = \frac{A^2}{n\cdot I(\theta)}\,.$$

Es bleibt zu zeigen, dass

$$A = \frac{d}{d\theta}\mathbb{E}_\theta\,\hat{\theta}(X_1,\dots,X_n)\,.$$

Wir zeigen die Aussage für den absolut stetigen Fall (im diskreten Fall sind die Integrale durch Summen zu ersetzen):

$$\begin{aligned}A &= \int \frac{\partial}{\partial\theta}\log L(x_1,\dots,x_n,\theta)\cdot\hat{\theta}(x_1,\dots,x_n)\cdot L(x_1,\dots,x_n,\theta)\,dx_1\dots dx_n \\ &= \int \frac{\partial}{\partial\theta}L(x_1,\dots,x_n,\theta)\cdot\hat{\theta}(x_1,\dots,x_n)\,dx_1\dots dx_n \overset{\text{Vor. 2)}}{=} \frac{d}{d\theta}\mathbb{E}_\theta\,\hat{\theta}(X_1,\dots,X_n)\,.\end{aligned}$$

□

Folgerung 3.5.1
Falls $\hat{\theta}$ ein erwartungstreuer Schätzer für θ ist und die Voraussetzungen des Satzes 3.5.1 erfüllt sind, so gilt

$$\operatorname{Var}_\theta\hat{\theta}(X_1,\dots,X_n) \geq \frac{1}{n\cdot I(\theta)}\,.$$

Beweis Wende die Ungleichung von Cramér-Rao an $\hat{\theta}$ mit

$$\frac{d}{d\theta}\left(\mathbb{E}_\theta\,\hat{\theta}(X_1,\ldots,X_n)\right) = \frac{d}{d\theta}\theta = 1$$

an. □

An folgenden Beispielen werden wir sehen, dass der Schätzer $\bar{X}_n$ des Erwartungswertes μ in der Klasse aller Schätzer für μ, die die Voraussetzungen des Satzes 3.5.1 erfüllen, die kleinste Varianz besitzt. Somit ist $\bar{X}_n$ der beste erwartungstreue Schätzer in dieser Klasse für mindestens zwei parametrische Familien von Verteilungen:

- Normalverteilung und
- Poisson-Verteilung.

Beispiel 3.5.1

1. $X_i \sim N(\mu,\sigma^2)$, $\hat{\mu} = \bar{X}_n$ als Schätzer für μ. Dabei ist $\hat{\mu}$ erwartungstreu mit $\mathrm{Var}\hat{\mu} = \sigma^2/n$. Zeigen wir, dass die Cramér-Rao-Schranke für die Varianz eines erwartungstreuen Schätzers $\hat{\theta}$ für μ ebenso gleich σ^2/n ist. Prüfen wir zunächst die Voraussetzungen des Satzes 3.5.1:

   Zeigen wir, dass

$$0 = \frac{d}{d\mu}\int_{\mathbb{R}} L(x,\mu)\,dx = \int_{\mathbb{R}} \frac{\partial}{\partial\mu}L(x,\mu)\,dx \quad \text{mit} \quad L(x,\mu) = \frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} :$$

$$\begin{aligned}\frac{\partial}{\partial\mu}L(x,\mu) &= \frac{2(x-\mu)}{2\sigma^2}\cdot\frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} = \frac{x-\mu}{\sigma^2}\cdot L(x,\mu)\,,\\ \mathbb{E}\frac{\partial}{\partial\mu}L(X,\mu) &= \int_{\mathbb{R}}\frac{\partial}{\partial\mu}L(x,\mu)\,dx = \mathbb{E}\left(\frac{X-\mu}{\sigma^2}\right) = 0\,.\end{aligned}$$

   Zeigen wir weiterhin die Gültikeit der Bedingung 2) des Satzes 3.5.1:

$$\frac{d}{d\mu}\mathbb{E}\bar{X}_n = \frac{d}{d\mu}(\mu) = 1 \stackrel{?}{=} \frac{1}{n}\int_{\mathbb{R}^n}(x_1+\ldots+x_n)\frac{\partial}{\partial\mu}\left(\prod_{i=1}^n \frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{1}{2}\left(\frac{x_i-\mu}{\sigma}\right)^2}\right)dx_1\ldots dx_n\,.$$

   Induktion bzgl. n:

   - Induktionsanfang $n=1$:

$$\int_{\mathbb{R}} x\frac{\partial}{\partial\mu}L(x,\mu)\,dx = \int_{\mathbb{R}}\frac{x(x-\mu)}{\sigma^2}L(x,\mu)\,dx = \frac{1}{\sigma^2}\left(\mathbb{E}_\mu X^2-\mu^2\right) = \frac{\mathrm{Var}_\mu X}{\sigma^2} = 1\,.$$

   - Induktionshypothese: Für n gilt

$$\int_{\mathbb{R}^n}(x_1+\ldots+x_n)\cdot\frac{\partial}{\partial\mu}L(x_1,\ldots,x_n,\mu)\,dx_1\ldots dx_n = n\,.$$

- Induktionsschritt $n \to n+1$:

$$A = \int_{\mathbb{R}^{n+1}} (x_1 + \ldots + x_{n+1}) \frac{\partial}{\partial \mu} \underbrace{L(x_1, \ldots, x_{n+1}, \mu)}_{= L(x_1, \ldots, x_n, \mu) \cdot L(x_{n+1}, \mu)} \, dx_1 \ldots dx_{n+1} \overset{?}{=} n+1 \,.$$

Dabei gilt für A:

$$\begin{aligned}
A &= \int_{\mathbb{R}^{n+1}} (x_1 + \ldots + x_n) \cdot \Bigg(\frac{\partial}{\partial \mu} L(x_1, \ldots, x_n, \mu) \cdot L(x_{n+1}, \mu) + L(x_1, \ldots, x_n, \mu) \cdot \\
&\quad \cdot \frac{\partial}{\partial \mu} L(x_{n+1}, \mu) \Bigg) dx_1 \ldots dx_n dx_{n+1} + \int_{\mathbb{R}^{n+1}} x_{n+1} \Bigg(\frac{\partial}{\partial \mu} L(x_1, \ldots, x_n, \mu) \cdot \\
&\quad \cdot L(x_{n+1}, \mu) + L(x_1, \ldots, x_n, \mu) \cdot \frac{\partial}{\partial \mu} L(x_{n+1}, \mu) \Bigg) dx_1 \ldots dx_n dx_{n+1} \\
&= n \cdot \underbrace{\int_{\mathbb{R}} L(x_{n+1}, \mu) \, dx_{n+1}}_{=1} + \int_{\mathbb{R}^n} (x_1 + \ldots + x_n) \cdot L(x_1, \ldots, x_n, \mu) \, dx_1 \ldots dx_n \cdot \\
&\quad \cdot \underbrace{\int \frac{\partial}{\partial \mu} L(x_{n+1}, \mu) \, dx_{n+1}}_{=0} + \int_{\mathbb{R}} x_{n+1} L(x_{n+1}, \mu) \, dx_{n+1} \cdot \\
&\quad \cdot \underbrace{\int_{\mathbb{R}^n} \frac{\partial}{\partial \mu} L(x_1, \ldots, x_n, \mu) \, dx_1 \ldots dx_n}_{=0} + \underbrace{\int_{\mathbb{R}} x_{n+1} \frac{\partial}{\partial \mu} L(x_{n+1}, \mu) \, dx_{n+1}}_{= \frac{d}{d\mu} \mathbb{E}_\mu X = \frac{d}{d\mu} \mu = 1} \cdot \\
&\quad \cdot \underbrace{\int_{\mathbb{R}^n} L(x_1, \ldots, x_n, \mu) \, dx_1 \ldots dx_n}_{=1} = n+1 \,.
\end{aligned}$$

Nachdem alle Voraussetzungen erfüllt sind, berechnen wir die Schranke

$$\frac{1}{n \cdot I(\mu)} \quad \text{mit} \quad I(\mu) = \mathbb{E}_\mu \left(\frac{\partial}{\partial \mu} \log L(X, \mu) \right)^2 .$$

Es gilt

$$\frac{\partial}{\partial \mu} \log L(x, \mu) = \frac{\partial}{\partial \mu} \left(-\log \sqrt{2\pi\sigma^2} - \frac{1}{2} \left(\frac{x-\mu}{\sigma} \right)^2 \right) = -\frac{2(x-\mu)}{2\sigma^2} \cdot (-1) = \frac{x-\mu}{\sigma^2} \,,$$

woraus folgt, dass

$$I(\mu) = \frac{1}{\sigma^4} \mathbb{E}_\mu (X-\mu)^2 = \frac{1}{\sigma^4} \cdot \operatorname{Var}_\mu X = \frac{\sigma^2}{\sigma^4} = \frac{1}{\sigma^2} \quad \Longrightarrow \quad n \cdot I(\mu) = \frac{n}{\sigma^2} \,.$$

Insgesamt gilt also

$$\operatorname{Var}_\mu \hat{\theta} \geq \frac{1}{\frac{n}{\sigma^2}} = \frac{\sigma^2}{n} = \operatorname{Var}_\mu \bar{X}_n$$

für einen beliebigen erwartungstreuen Schätzer $\hat{\theta}$ für μ, der die Voraussetzungen des Satzes 3.5.1 erfüllt.

2. Das zweite Beispiel sei folgende Übungsaufgabe:

   Übungsaufgabe 3.5.1
   Seien $X_i \sim Poisson(\lambda)\,, \quad i = 1, \ldots, n$. Zeigen Sie, dass die Schranke von Cramér-Rao

$$\frac{1}{n \cdot I(\lambda)} = \frac{\lambda}{n} = \operatorname{Var}_\lambda \bar{X}_n$$

   ist. Dies bedeutet, dass auch hier $\bar{X}_n$ der beste erwartungstreue Schätzer ist, der die Voraussetzungen des Satzes 3.5.1 erfüllt.

An Hand des nächsten Beispiels wollen wir zeigen, dass die Konstruktion von Schätzern mit einer Varianz, die kleiner als die Cramér-Rao-Schranke ist, möglich ist, falls die Voraussetzungen von Satz 3.5.1 nicht erfüllt sind.

Beispiel 3.5.2
Seien $X_i \sim U[0, \theta]$, $\theta > 0$. Dann ist die Bedingung „supp$f_\theta(x) = [0, \theta]$ unabhängig von θ“ verletzt und auch eine weitere Bedingung:

$$0 \neq \int_{\mathbb{R}} \frac{\partial}{\partial\theta} L(x, \theta)\, dx = \int_0^\theta \left(\frac{1}{\theta}\right)' dx = -\frac{1}{\theta^2} \cdot \theta = -\frac{1}{\theta}\,.$$

Sei $\hat{\theta}$ ein erwartungstreuer Schätzer für θ, so würde nach der Ungleichung von Cramér-Rao folgen, dass $\operatorname{Var}_\theta \hat{\theta} \geq (n \cdot I(\theta))^{-1}$, wobei

$$I(\theta) = \mathbb{E}\left(\frac{\partial}{\partial\theta} \log L(X, \theta)\right)^2 = \int_0^\theta \frac{1}{\theta}\left(\frac{\partial}{\partial\theta} \log\left(\frac{1}{\theta}\right)\right)^2 dx = \frac{1}{\theta}\int_0^\theta dx \cdot \left(-\frac{1}{\theta}\right)^2 = \frac{1}{\theta^2}\,.$$

Damit hätten wir

$$\operatorname{Var}_\theta \hat{\theta} \geq \frac{\theta^2}{n}\,.$$

Betrachten wir

$$\hat{\theta}(X_1, \ldots, X_n) = \frac{n+1}{n} \max\{X_1, \ldots, X_n\} = \frac{n+1}{n} X_{(n)}\,.$$

Zeigen wir, dass

$$\mathbb{E}_\theta\, \hat{\theta}(X_1, \ldots, X_n) = \theta \quad \text{und} \quad \operatorname{Var}_\theta \hat{\theta}(X_1, \ldots, X_n) < \frac{\theta^2}{n}\,.$$

Berechnen wir dazu $\mathbb{E}_\theta X_{(n)}^k$, $k \in \mathbb{N}$. Es gilt

$$F_{X_{(n)}}(x) = F_{X_i}^n(x) = \begin{cases} \frac{x^n}{\theta^n}\,, & x \in [0, \theta]\,, \\ 1\,, & x \geq \theta\,, \\ 0\,, & x < 0\,, \end{cases}$$

$$f_{X_{(n)}}(x) = F'_{X_{(n)}}(x) = \frac{n x^{n-1}}{\theta^n} \cdot \mathbb{I}(x \in [0, \theta])\,,$$

$$\mathbb{E}_\theta X_{(n)}^k = \int_0^\theta x^k \frac{n x^{n-1}}{\theta^n}\, dx = \frac{n}{\theta^n}\int_0^\theta x^{n+k-1}\, dx = \frac{n \cdot \theta^{n+k}}{\theta^n \cdot (n+k)} = \frac{n\theta^k}{n+k}\,.$$

Damit folgt

$$\mathbb{E}_\theta \hat{\theta} = \frac{n+1}{n} \cdot \mathbb{E}_\theta X_{(n)} = \frac{n+1}{n} \cdot \frac{n\theta}{n+1} = \theta\,,$$

das heißt, $\hat{\theta}$ ist erwartungstreu. Weiterhin gilt

$$\begin{aligned}
\operatorname{Var}_\theta \hat{\theta} &= \left(\frac{n+1}{n}\right)^2 \cdot \operatorname{Var}_\theta X_{(n)} = \left(\frac{n+1}{n}\right)^2 \cdot \left(\frac{n\theta^2}{n+2} - \frac{n^2\theta^2}{(n+1)^2}\right)^2 \\
&= \frac{(n+1)^2}{n^2} \cdot \frac{n(n+1)^2 - n^2(n+2)}{(n+2)(n+1)^2} \cdot \theta^2 \\
&= \frac{\theta^2}{n(n+2)}(n^2 + 2n + 1 - n^2 - 2n) = \frac{\theta^2}{n(n+2)}
\end{aligned}$$

und somit

$$\operatorname{Var}_\theta \hat{\theta} = \frac{\theta^2}{n(n+2)} < \frac{\theta^2}{n}\,.$$

3.5.2 Bedingte Erwartung

Seien X und Y zwei Zufallsvariablen, wobei Y eine absolut stetige Verteilung besitzt. Dann folgt $\mathbb{P}(Y = y) = 0 \quad \forall\, y \in \mathbb{R}$. Deshalb kann die bedingte Wahrscheinlichkeit $\mathbb{P}(X \in B \,|\, Y = y)$ auf dem gewöhnlichen Wege

$$\mathbb{P}(X \in B | Y = y) = \frac{\mathbb{P}(X \in B,\, Y = y)}{\mathbb{P}(Y = y)}$$

nicht definiert werden. Aus der Praxis ist aber eine Reihe von Fragestellungen bekannt (z.B. Bayessche Analyse), in denen Wahrscheinlichkeiten $\mathbb{P}(X \in B | Y = y)$ ausgewertet werden müssen. Deswegen werden wir eine neue Definition der bedingten Wahrscheinlichkeit geben, die solche Situationen berücksichtigt. Diese Definition erfolgt durch die Definition der bedingten Erwartung.

Schema:

1. Es wird die bedingte Erwartung von der Zufallsvariablen X bzgl. der σ-Algebra $\mathcal{B}$ als Zufallsvariable $\mathbb{E}(X|\mathcal{B})$ eingeführt, wobei $\mathcal{B}$ eine Teil-σ-Algebra von $\mathcal{F}$ und $(\Omega, \mathcal{F}, \mathbb{P})$ der Wahrscheinlichkeitsraum ist.

2. Die bedingte Erwartung von X unter der Bedingung Y wird als $\mathbb{E}(X|Y) = \mathbb{E}(X|\sigma_Y)$ eingeführt, wobei σ_Y die von Y erzeugte σ-Algebra ist.

3. $\mathbb{P}(X \in B|Y = y)$ wird als Zufallsvariable $\mathbb{E}(\mathbb{I}(X \in B)|Y)$ auf der Menge $\{\omega \in \Omega : Y(\omega) = y\}$ eingeführt.

Gehen wir nun dieses Schema im Detail durch:

1. Sei $(\Omega, \mathcal{F}, \mathbb{P})$ ein Wahrscheinlichkeitsraum und $\mathcal{B}$ eine Teil-σ-Algebra von $\mathcal{F}$, d.h. $\mathcal{B} \subseteq \mathcal{F}$.

Definition 3.5.1
Der *bedingte Erwartungswert* einer Zufallsvariablen X definiert auf dem Wahrscheinlichkeitsraum $(\Omega, \mathcal{F}, \mathbb{P})$ bezüglich einer σ-Algebra $\mathcal{B} \subseteq \mathcal{F}$ ist in dem Fall $\mathbb{E}|X| < \infty$ als eine $\mathcal{B}$-messbare Zufallsvariable Y definiert, die die Eigenschaft

$$\int_B Y(\omega)\,\mathbb{P}(d\omega) = \int_B X\,\mathbb{P}(d\omega)\,, \qquad \forall\, B \in \mathcal{B}$$

besitzt. Dabei wird die Bezeichnung $Y = \mathbb{E}(X|\mathcal{B})$ verwendet.

Warum existiert diese Zufallsvariable Y?

- Zerlegen wir X in den positiven X_+ und negativen X_- Anteil $X = X_+ - X_-$ und beweisen die Existenz von $\mathbb{E}(X_\pm|\mathcal{B})$. Danach setzen wir $\mathbb{E}(X|\mathcal{B}) = \mathbb{E}(X_+|\mathcal{B}) - \mathbb{E}(X_-|\mathcal{B})$.
- Somit genügt es zu zeigen, dass der Erwartungswert $\mathbb{E}(X|\mathcal{B})$ einer nicht negativen Zufallsvariablen $X \geq 0$ fast sicher existiert.
- Sei $Q(B) = \int_B X(\omega)\,\mathbb{P}(d\omega)$. Man kann zeigen, dass $Q(\cdot)$ ein Maß auf $(\Omega, \mathcal{F})$ ist. Dabei folgt aus $\mathbb{P}(B) = 0$ die Gleichheit $Q(B) = 0$ für $B \in \mathcal{B}_\mathbb{R}$ (bzw. $B \in \mathcal{B}$). Somit ist Q absolut stetig bzgl. $\mathbb{P}$. Weiter existiert nach dem Satz von Radon-Nikodym eine Dichte $Y(\omega)$, die messbar bzgl. $\mathcal{B}$ ist und für die

  $$Q(B) = \int_B Y(\omega)\,\mathbb{P}(d\omega) \quad \Longrightarrow \quad Y(\omega) = \mathbb{E}(X|\mathcal{B})$$

  gilt.

Bemerkung 3.5.1
Aus der obigen Beweisskizze wird ersichtlich, dass $Y(\omega) = \mathbb{E}(X|\mathcal{B})$ nur $\mathbb{P}$-fast sicher definiert ist. Somit kann man mehrere Versionen von $Y(\omega)$ angeben, die sich auf einer Menge der Wahrscheinlichkeit 0 unterscheiden.

Satz 3.5.2 (Eigenschaften des bedingten Erwartungswertes):
Seien X und Y Zufallsvariablen auf dem Wahrscheinlichkeitsraum $(\Omega, \mathcal{F}, \mathbb{P})$ mit der Eigenschaft $\mathbb{E}|X| < \infty$, $\mathbb{E}|Y| < \infty$ und $\mathbb{E}|XY| < \infty$ (dies kann noch ein wenig abgeschwächt werden, ist hier allerdings ausreichend). Seien $\mathcal{B}$, $\mathcal{B}_1$ und $\mathcal{B}_2$ Teil-σ-Algebren von $\mathcal{F}$. Es gelten folgende Eigenschaften (im fast sicheren Sinne):

a) $\mathbb{E}(X|\{\emptyset, \Omega\}) = \mathbb{E}X$, $\mathbb{E}(X|\mathcal{F}) = X$ fast sicher.
b) Falls $X \leq Y$ fast sicher, dann gilt ebenso $\mathbb{E}(X|\mathcal{B}) \leq \mathbb{E}(Y|\mathcal{B})$ fast sicher.
c) Es gilt $\mathbb{E}(XY|\mathcal{B}) = X \cdot \mathbb{E}(Y|\mathcal{B})$, falls X $\mathcal{B}$-messbar ist.
d) $\mathbb{E}(c|\mathcal{B}) = c$ für $c = const.$
e) Es gilt $\mathbb{E}((X|\mathcal{B}_2)|\mathcal{B}_1) = \mathbb{E}(X|\mathcal{B}_1)$ und $\mathbb{E}(\mathbb{E}(X|\mathcal{B}_1)|\mathcal{B}_2) = \mathbb{E}(X|\mathcal{B}_1)$, falls $\mathcal{B}_1 \subseteq \mathcal{B}_2$.
f) Falls X unabhängig von $\mathcal{B}$ ist (d.h., die σ-Algebren $\sigma_X = X^{-1}(\mathcal{B}_\mathbb{R})$ und $\mathcal{B}$ sind unabhängig), dann gilt $\mathbb{E}(X|\mathcal{B}) = \mathbb{E}X$.

Ohne Beweis (siehe Beweis in [26]).

Beispiel 3.5.3
Sei $\mathcal{B} = \sigma(\{A_1, \ldots, A_n\})$, wobei $\{A_1, \ldots, A_n\}$ eine messbare Zerlegung des Wahrscheinlichkeitsraumes $(\Omega, \mathcal{F}, \mathbb{P})$ ist, d.h. $\bigcup_{i=1}^n A_i = \Omega$, $A_i \cap A_j = \emptyset\,, \quad i \neq j$, $\mathbb{P}(A_i) > 0\,, \quad i = 1, \ldots, n$. Was ist $\mathbb{E}(X|\mathcal{B})$? Da $\mathbb{E}(X|\mathcal{B})$ $\mathcal{B}$-messbar ist, können wir die allgemeine Form

der Funktionen ausnutzen, die messbar bzgl. einer endlich erzeugten σ-Algebra $\mathcal{B} = \sigma(\{A_1, \ldots, A_n\})$ sind: $\mathbb{E}(X(\omega)|\mathcal{B}) = \sum_{i=1}^n k_i \mathbb{I}(\omega \in A_i)$ (ohne Beweis).

Berechnen wir k_i: Aus der Definition 3.5.1 folgt für $B = A_j$

$$\begin{aligned}
\int_B \mathbb{E}(X|\mathcal{B})\,\mathbb{P}(d\omega) &= \int_{A_j} \sum_{i=1}^n k_i \cdot \mathbb{I}(\omega \in A_i)\,\mathbb{P}(d\omega) = k_j \cdot \mathbb{P}(A_j) \\
&= \int_B X\,\mathbb{P}(d\omega) = \int_{A_j} X\,\mathbb{P}(d\omega) = \mathbb{E}(X \cdot \mathbb{I}_{A_j}) \\
&\Longrightarrow k_j = \frac{\mathbb{E}(X \cdot \mathbb{I}_{A_j})}{\mathbb{P}(A_j)}\,, \quad j = 1, \ldots, n\,. \\
&\Longrightarrow \mathbb{E}(X(\omega)|\mathcal{B}) = \frac{\mathbb{E}(X \cdot \mathbb{I}_{A_j})}{\mathbb{P}(A_j)}\,, \quad \text{falls } \omega \in A_j\,, \quad j = 1, \ldots, n\,.
\end{aligned}$$

2. *Bedingte Erwartung bzgl. einer Zufallsvariablen Y:*

   Definition 3.5.2
   Seien X und Y zwei Zufallsvariablen definiert auf dem Wahrscheinlichkeitsraum $(\Omega, \mathcal{F}, \mathbb{P})$. Der *bedingte Erwartungswert von X unter der Bedingung Y* wird als $\mathbb{E}(X|Y) = \mathbb{E}(X|\sigma_Y)$ eingeführt, wobei σ_Y die von Y erzeugte σ-Algebra ist: $\sigma_Y = Y^{-1}(\mathcal{B}_\mathbb{R})$.

   Lemma 3.5.1
   Es existiert eine Borel-messbare Funktion $g : \mathbb{R} \to \mathbb{R}$, für die gilt, dass $\mathbb{E}(X|Y) = g(Y)$ fast sicher (Ohne Beweis).

   Daher wird die Schreibweise $\mathbb{E}(X|Y = y)$ als $g(y)$ verstanden: $\mathbb{E}(X|Y = y) = g(y)$ oder $\mathbb{E}(X|Y = y)$ ist der Wert von $\mathbb{E}(X|Y)$ auf der Menge $\{\omega \in \Omega : Y(\omega) = y\}$.

3. *Bedingte Wahrscheinlichkeit bzgl. einer σ-Algebra bzw. einer Zufallsvariable.*

   Definition 3.5.3
   Die *bedingte Wahrscheinlichkeit* von $A \in \mathcal{F}$ unter der Bedingung $\mathcal{B}$ ist gegeben durch $\mathbb{P}(A|\mathcal{B}) = \mathbb{E}(\mathbb{I}_A|\mathcal{B})$ fast sicher. Analog dazu definieren wir $\mathbb{P}(A|Y) = \mathbb{E}(\mathbb{I}_A|Y)$ für eine Zufallsvariable Y.

   Bemerkung 3.5.2
   Die so definierte Familie von Zufallsvariablen $\mathbb{P}(\cdot|\mathcal{B})$ erfüllen (fast sicher) nicht die Eigenschaften eines Maßes: Es gilt

   $$0 \leq \mathbb{P}(A|\mathcal{B}) \leq 1\,, \quad \forall\, A \in \mathcal{F} \text{ fast sicher,}$$

   aber die Eigenschaft der σ-Additivität

   $$\mathbb{P}(\bigcup_{i=1}^\infty A_i|\mathcal{B}) \overset{\text{f.s.}}{=} \sum_{i=1}^\infty \mathbb{P}(A_i|\mathcal{B})$$

   für disjunkte $\{A_i\}$ hängt von der Version $\mathbb{P}(\cdot|\mathcal{B})$ ab. Das bedeutet, es existiert kein $M \in \mathcal{F} : \mathbb{P}(M) = 0$, so dass die obige Eigenschaft für alle $\omega \in M^C$ gilt.

3.5.3 Suffizienz

Sei $(X_1, \ldots, X_n)$ eine Stichprobe von unabhängigen identisch verteilten Zufallsvariablen X_i mit Verteilungsfunktion F_θ, $\theta \in \Theta \subseteq \mathbb{R}^m$. Wenn man von der vollen Information $\{X_1 = x_1, \ldots, X_n = x_n\}$ zum Schätzer $\hat{\theta}(X_1, \ldots, X_n)$ des Parameters θ übergeht, dann entsteht durch die Abbildung

$$\hat{\theta} : \mathbb{R}^n \to \mathbb{R}^m, \qquad m \ll n$$

ein Informationsverlust, weil man normalerweise $(X_1, \ldots, X_n)$ nicht aus $\hat{\theta}(X_1, \ldots, X_n)$ zurückrechnen kann. Die sogenannten *suffizienten* Schätzer minimieren diesen Informationsverlust im stochastischen Sinne:

Definition 3.5.4

1. Seien Zufallsvariablen $X_1, \ldots, X_n$ und $\hat{\theta}(X_1, \ldots, X_n)$ diskret verteilt. Ein Schätzer $\hat{\theta}$ des Parameters θ heißt *suffizient*, falls

$$\mathbb{P}_\theta \Big(X_1 = x_1, \ldots, X_n = x_n \,|\, \hat{\theta}(X_1, \ldots, X_n) = t \Big)$$

   nicht von θ abhängt für beliebige $x_1, \ldots, x_n$ und t aus den Trägern der Zähldichten von $(X_1, \ldots, X_n)$ bzw. $\hat{\theta}(X_1, \ldots, X_n)$.

2. Falls $X_1, \ldots, X_n$ und $\hat{\theta}(X_1, \ldots, X_n)$ absolut stetig verteilt sind, dann heißt der Schätzer $\hat{\theta}$ *suffizient* für θ, falls die Wahrscheinlichkeit

$$\mathbb{P} \Big((X_1, \ldots, X_n) \in B \,|\, \hat{\theta}(X_1, \ldots, X_n) = t \Big)$$

   für beliebige $B \in \mathcal{B}_{\mathbb{R}^n}$ und $t \in \operatorname{supp} f_{\hat{\theta}}$ nicht von $\theta \in \Theta$ abhängt, wobei $f_{\hat{\theta}}$ die Dichte von $\hat{\theta}$ ist.

Bemerkung 3.5.3

1. Betrachten wir im diskreten Fall die bedingte Likelihood-Funktion

$$L_{\hat{\theta}}(x_1, \ldots, x_n, \theta) = \mathbb{P}_\theta \Big(X_1 = x_1, \ldots, X_n = x_n \,|\, \hat{\theta}(X_1, \ldots, X_n) = t \Big) .$$

   Aus Definition 3.5.4 folgt, dass wir keinen neuen ML-Schätzer für θ aus dieser bedingten Likelihood $L_\theta(x_1, \ldots, x_n, \theta)$ gewinnen werden können, da sie nicht von θ abhängt. Das heißt, der Schätzer $\hat{\theta}$ enthält bereits die volle Information über θ, die man aus der Stichprobe $(x_1, \ldots, x_n)$ gewinnen kann.

2. Falls $g : \mathbb{R}^m \to \mathbb{R}^m$ eine bijektive Borel-messbare Abbildung und $\hat{\theta}(X_1, \ldots, X_n)$ ein suffizienter Schätzer von $\theta \in \Theta \subset \mathbb{R}^m$ ist, dann ist der Schätzer $g(\hat{\theta}(X_1, \ldots, X_n))$ auch ein suffizienter Schätzer für θ. Dies wird aus der Tatsache ersichtlich, dass

$$\Big\{ \omega \in \Omega : \, g \Big(\hat{\theta}(X_1, \ldots, X_n) \Big) = t \Big\} = \Big\{ \omega \in \Omega : \, \hat{\theta}(X_1, \ldots, X_n) = g^{-1}(t) \Big\} , \quad \forall\, t .$$

Lemma 3.5.2 (Suffizienz):
Seien Zufallsvariablen $X_1, \ldots, X_n$ und $\hat{\theta}(X_1, \ldots, X_n)$ entweder alle diskret oder absolut stetig verteilt mit den Likelihood-Funktionen

$$L(x_1, \ldots, x_n, \theta) = \begin{cases} \mathbb{P}_\theta(X_1 = x_1, \ldots, X_n = x_n)\,, & \text{im diskreten Fall,} \\ f_{X_1,\ldots,X_n}(x_1, \ldots, x_n)\,, & \text{im absolut stetigen Fall,} \end{cases}$$

$$L_{\hat{\theta}}(t, \theta) = \begin{cases} \mathbb{P}_\theta(\hat{\theta}(X_1, \ldots, X_n) = t)\,, & \text{im diskreten Fall,} \\ f_{\hat{\theta}}(t, \theta)\,, & \text{im absolut stetigen Fall.} \end{cases}$$

Seien die Träger L bzw. $L_{\hat{\theta}}$ gegeben durch

$$\begin{aligned} \text{supp}L &= \{(x_1, \ldots, x_n) \in \mathbb{R}^n : \quad L(x_1, \ldots, x_n, \theta) > 0\}\,, \\ \text{supp}L_{\hat{\theta}} &= \{t \in \mathbb{R} : \quad L_{\hat{\theta}}(t, \theta) > 0\}\,. \end{aligned}$$

Der Schätzer $\hat{\theta}$ ist suffizient für θ genau dann, wenn

$$\frac{L(x_1, \ldots, x_n, \theta)}{L_{\hat{\theta}}(\hat{\theta}(x_1, \ldots, x_n), \theta)} \tag{3.5.1}$$

nicht von θ abhängig ist für alle $(x_1, \ldots, x_n) \in \text{supp}L$.

Beweis Wir beweisen lediglich den diskreten Fall:

„$\Longrightarrow$“ Ist $\hat{\theta}$ suffizient, so überprüfen wir, ob damit folgt, dass (3.5.1) von θ abhängt für alle $(x_1, \ldots, x_n) \in \text{supp}L$. Es gilt:

$$\begin{aligned} &\mathbb{P}_\theta(X_1 = x_1, \ldots, X_n = x_n \,|\, \hat{\theta}(X_1, \ldots, X_n) = t) \\ &\qquad = \frac{\mathbb{P}_\theta(X_1 = x_1, \ldots, X_n = x_n, \hat{\theta}(X_1, \ldots, X_n) = t)}{\mathbb{P}_\theta(\hat{\theta}(X_1, \ldots, X_n) = t)} \\ &\qquad = \begin{cases} 0\,, & \text{falls } \hat{\theta}(x_1, \ldots, x_n) \neq t \\ \frac{\mathbb{P}_\theta(X_1 = x_1, \ldots, X_n = x_n)}{\mathbb{P}_\theta(\hat{\theta}(X_1, \ldots, X_n) = \hat{\theta}(x_1, \ldots, x_n))}\,, & \text{falls } \hat{\theta}(x_1, \ldots, x_n) = t\,. \end{cases} \end{aligned}$$

Somit hängt (3.5.1) nicht von θ ab.

„$\Longleftarrow$“ Folgt aus dem 1.Fall durch Betrachtung von hinten.

□

Beispiel 3.5.4

1. *Bernoulli-Verteilung:* Seien $X_i \sim Bernoulli(p)$, $p \in [0, 1]$, $i = 1, \ldots, n$, $\hat{p} = \bar{X}_n$ ein erwartungstreuer Schätzer für p. Wir zeigen nun, dass $\hat{p}$ suffizient ist. Es gilt

$$\hat{p} = \bar{X}_n = \frac{1}{n} \sum_{i=1}^{n} X_i = \frac{1}{n} Y\,,$$

wobei $Y \sim Bin(n, p)$. Es genügt nach Bemerkung 3.5.3 2) zu zeigen, dass Y ein suffizienter Schätzer für p ist. Nach Lemma 3.5.2 gilt für $x_i \in \{0, 1\}\,, \quad i = 1, \ldots, n$

$$\mathbb{P}(X_1 = x_1, \ldots, X_n = x_n) = \prod_{i=1}^{n} p^{x_i}(1-p)^{1-x_i} = p^{\sum_{i=1}^n x_i}(1-p)^{n-\sum_{i=1}^n x_i}\,.$$

Definieren wir nun L_Y als

$$L_Y(y,p) = \binom{n}{y} p^y (1-p)^{1-y}\,, \qquad y = 0,\dots,n\,.$$

Setzen wir nun statt y die Summe $\sum_{i=1}^n x_i$ ein und betrachten

$$\frac{L(x_1,\dots,x_n,p)}{L_Y(\sum_{i=1}^n x_i,p)} = \frac{p^{\sum_{i=1}^n x_i}(1-p)^{n-\sum_{i=1}^n x_i}}{\binom{n}{\sum_{i=1}^n x_i} p^{\sum_{i=1}^n x_i}(1-p)^{n-\sum_{i=1}^n x_i}} = \frac{1}{\binom{n}{\sum_{i=1}^n x_i}}\,.$$

Dies hängt offensichtlich nicht von p ab, womit folgt, dass Y und somit $\hat{p}$ suffizient sind.

2. *Normalverteilung mit bekannter Varianz:* Seien $X_i \sim N(\mu,\sigma^2)\,,\quad i = 1,\dots,n,\ \sigma^2$ bekannt. So ist $\hat{\mu} = \bar{X}_n$ ein erwartungstreuer Schätzer für μ. Zeigen wir nun, dass $\hat{\mu}$ suffizient ist: Betrachten wir

$$\begin{aligned}
L(x_1,\dots,x_n,\mu) &= \prod_{i=1}^n \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2}\left(\frac{x_i-\mu}{\sigma}\right)^2\right)\\
&= \frac{1}{(2\pi\sigma^2)^{n/2}} \cdot \exp\left(-\frac{\sum_{i=1}^n (x_i-\mu)^2}{2\sigma^2}\right)\\
&= \frac{1}{(2\pi\sigma^2)^{n/2}} \cdot \exp\left(-\frac{1}{2\sigma^2}\left(\sum_{i=1}^n (x_i-\bar{x}_n+\bar{x}_n-\mu)^2\right)\right)
\end{aligned}$$

und mit

$$\begin{aligned}
\sum_{i=1}^n (x_i-\bar{x}_n+\bar{x}_n-\mu)^2 &= \sum_{i=1}^n (x_i-\bar{x}_n)^2 + 2\sum_{i=1}^n (x_i-\bar{x}_n)(x_i-\mu) + \sum_{i=1}^n (\bar{x}_n-\mu)^2\\
&= \sum_{i=1}^n (x_i-\bar{x}_n)^2 + 2(\bar{x}_n-\mu)\underbrace{\sum_{i=1}^n (x_i-\bar{x}_n)}_{=n\bar{x}_n-n\bar{x}_n=0} + n(\bar{x}_n-\mu)^2\\
&= \sum_{i=1}^n (x_i-\bar{x}_n)^2 + n(\bar{x}_n-\mu)^2
\end{aligned}$$

folgt weiter

$$= \frac{1}{(2\pi\sigma^2)^{n/2}} \cdot \exp\left(-\frac{\sum_{i=1}^n (x_i-\bar{x}_n)^2 + n(\bar{x}_n-\mu)^2}{2\sigma^2}\right)\,.$$

Ferner gilt bekanntermaßen $\hat{\mu} \sim N(\mu,\sigma^2/n)$, und somit

$$\begin{aligned}
L_{\hat{\mu}}(x,\mu) &= \frac{\sqrt{n}}{\sqrt{2\pi}\sigma} \cdot \exp\left(-\frac{n}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right)\,,\\
\frac{L(x_1,\dots,x_n,\mu)}{L_{\hat{\mu}}(\bar{x}_n,\mu)} &= \frac{\frac{1}{(2\pi\sigma^2)^{n/2}} \cdot \exp\left(-\frac{\sum_{i=1}^n (x_i-\bar{x}_n)^2 + n(\bar{x}_n-\mu)^2}{2\sigma^2}\right)}{\frac{\sqrt{n}}{\sqrt{2\pi}\sigma} \cdot \exp\left(\frac{-n(\bar{x}_n-\mu)^2}{2\sigma^2}\right)}\\
&= \frac{\sqrt{n}}{(2\pi\sigma^2)^{n/2-1}} \cdot \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^n (x_i-\bar{x}_n)\right)\,,
\end{aligned}$$

was von μ unabhängig ist. Somit folgt nach Lemma 3.5.2, dass $\hat{\mu} = \bar{X}_n$ ein suffizienter Schätzer für μ ist.

Mit Hilfe des nächsten Satzes von Neyman-Fisher wird es möglich sein zu zeigen, dass bei unbekannter Varianz der Schätzer $(\bar{X}_n, S_n^2)$ für (μ, σ^2) suffizient ist.

Satz 3.5.3 (Neyman-Fisher Faktorisierungssatz):
Unter den Voraussetzungen von Lemma 3.5.2 ist $\hat{\theta}(X_1, \ldots, X_n)$ ein suffizienter Schätzer für θ genau dann, wenn zwei messbare Funktionen $g : \mathbb{R}^m \times \Theta \to \mathbb{R}$ und $h : \mathbb{R}^n \to \mathbb{R}$ existieren, so dass folgende Faktorisierung der Likelihood-Funktion $L(x_1, \ldots, x_n, \theta)$ der Stichprobe $(X_1, \ldots, X_n)$ gilt:

$$L(x_1, \ldots, x_n, \theta) = g\left(\hat{\theta}(x_1, \ldots, x_n), \theta\right) \cdot h(x_1, \ldots, x_n)\,, \qquad \forall (x_1, \ldots, x_n) \in \operatorname{supp} L\,, \quad \theta \in \Theta\,.$$

Beweis Wir beweisen nur den diskreten Fall.

1. Falls $\hat{\theta}$ suffizient ist, dann hängt nach Lemma 3.5.2

$$\frac{L(x_1, \ldots, x_n, \theta)}{\underbrace{L_{\hat{\theta}}(\hat{\theta}(x_1, \ldots, x_n), \theta)}_{=g(\hat{\theta}(x_1,\ldots,x_n),\theta)}} = h(x_1, \ldots, x_n)$$

   nicht von θ ab. Somit bekommen wir die Faktorisierung von Neyman-Fisher.

2. Sei nun $L(x_1, \ldots, x_n, \theta) = g(\hat{\theta}(x_1, \ldots, x_n), \theta) \cdot h(x_1, \ldots, x_n)$ für alle $(x_1, \ldots, x_n) \in \operatorname{supp} L$, $\theta \in \Theta$. Führen wir eine Menge

$$C = \{(y_1, \ldots, y_n) \in \mathbb{R}^n : \hat{\theta}(y_1, \ldots, y_n) = \hat{\theta}(x_1, \ldots, x_n)\} = \hat{\theta}^{-1}\left(\hat{\theta}\left(x_1, \ldots, x_n\right)\right)$$

   ein. So gilt

$$\begin{aligned}
\frac{\mathbb{P}_\theta(X_1 = x_1, \ldots, X_n = x_n)}{\underbrace{L_\theta(\hat{\theta}(x_1, \ldots, x_n), \theta)}_{=\mathbb{P}_\theta(\hat{\theta}(X_1,\ldots,X_n)=\hat{\theta}(x_1,\ldots,x_n))}} &= \frac{g(\hat{\theta}(x_1, \ldots, x_n), \theta) \cdot h(x_1, \ldots, x_n)}{\sum_{(y_1,\ldots,y_n)\in C} \mathbb{P}_\theta(X_1 = y_1, \ldots, X_n = y_n)} \\
&= \frac{g(\hat{\theta}(x_1, \ldots, x_n), \theta) \cdot h(x_1, \ldots, x_n)}{\sum_{(y_1,\ldots,y_n)\in C} g(\underbrace{\hat{\theta}(y_1, \ldots, y_n)}_{=\hat{\theta}(x_1,\ldots,x_n)}, \theta) \cdot h(y_1, \ldots, y_n)} \\
&= \frac{h(x_1, \ldots, x_n)}{\sum_{(y_1,\ldots,y_n)\in C} h(y_1, \ldots, y_n)}\,,
\end{aligned}$$

   welches nicht von θ abhängt. Daher ist $\hat{\theta}$ nach Lemma 3.5.2 suffizient.

□

Beispiel 3.5.5

1. *Poisson-Verteilung:* Seien $X_i \sim Poisson(\lambda)$, $\lambda > 0$, $\hat{\lambda} = \bar{X}_n$ ein erwartungstreuer Schätzer für λ. Zeigen wir, dass $\hat{\lambda}$ suffizient ist. Es gilt für $x_i \in \{0, 1, 2, \ldots\}$, $i = 1, \ldots, n$

$$\begin{aligned} L(x_1, \ldots, x_n, \lambda) &= \prod_{i=1}^{n} e^{-\lambda} \frac{\lambda^{x_i}}{x_i!} = \frac{e^{-\lambda n} \cdot \lambda^{\sum_{i=1}^{n} x_i}}{x_1! \cdot \ldots \cdot x_n!} = \frac{e^{-n\lambda} \lambda^{n\bar{x}_n}}{x_1! \cdot \ldots \cdot x_n!} \\ &= g(\bar{x}_n, \lambda) \cdot h(x_1, \ldots, x_n)\,, \end{aligned}$$

wobei $g(\bar{x}_n, \lambda) = e^{-n\lambda} \cdot \lambda^{n\bar{x}_n}$, $\quad h(x_1, \ldots, x_n) = \frac{1}{x_1! \cdot \ldots \cdot x_n!}$ ist. Somit ist $\hat{\lambda} = \bar{X}_n$ nach Satz 3.5.3 suffizient.

2. *Exponentialverteilung:* Seien $X_i \sim Exp(\lambda)$, $\lambda > 0$, $\hat{\lambda} = \bar{X}_n^{-1}$ ein Momentenschätzer für λ, der zwar nicht erwartungstreu ist, jedoch stark konsistent, denn $\bar{X}_n \xrightarrow[n\to\infty]{\text{f.s.}} \mathbb{E}X_i = \frac{1}{\lambda}$ nach dem starken Gesetz der großen Zahlen. Zeigen wir, dass $\hat{\lambda}$ suffizient ist. Für $x_1 \geq 0, \ldots, x_n \geq 0$ gilt

$$\begin{aligned} L(x_1, \ldots, x_n, \lambda) &= \prod_{i=1}^{n} \lambda e^{-\lambda x_i} = \lambda^n e^{-\lambda \sum_{i=1}^{n} x_i} = \lambda^n e^{-\lambda n \bar{x}_n} \\ &= \lambda^n e^{-\frac{\lambda n}{\hat{\lambda}}} = g\left(\hat{\lambda}, \lambda\right) \cdot \underbrace{h(x_1, \ldots, x_n)}_{=1}\,, \end{aligned}$$

wobei $g(\hat{\lambda}, \lambda) = \lambda^n e^{-\frac{\lambda n}{\hat{\lambda}}}$ und $h(x_1, \ldots, x_n) \equiv 1$ ist. Somit ist $\hat{\lambda}$ nach dem Satz 3.5.3 suffizient.

Übungsaufgabe 3.5.2
Zeigen Sie mit Hilfe des Satzes 3.5.3, dass der Schätzer $(\bar{X}_n, S_n^2)$ suffizient für (μ, σ^2) im Falle der normal und unabhängig identisch verteilten Stichprobe $(X_1, \ldots, X_n)$, $X_i \sim N(\mu, \sigma^2)$ ist.

Bemerkung 3.5.4
Der Vorteil des Satzes von Neyman-Fisher ist, dass man für die Überprüfung der Suffizienzeigenschaft von $\hat{\theta}$ die Likelihood-Funktion von $\hat{\theta}$ nicht explizit zu kennen braucht. Dies ist insbesondere in den Fällen vorteilhaft, in denen der Schätzer $\hat{\theta}$ kompliziert ist und seine Likelihood-Funktion nicht analytisch angegeben werden kann (bzw. unbekannt ist).

3.5.4 Vollständigkeit

Definition 3.5.5
Ein Schätzer $\hat{\theta}(X_1, \ldots, X_n)$ des Parameters $\theta \in \Theta \subset \mathbb{R}^m$ heißt *vollständig*, falls für beliebige messbare Funktionen $g : \mathbb{R}^m \to \mathbb{R}$ mit der Eigenschaft $\mathbb{E}_\theta g(\hat{\theta}(X_1, \ldots, X_n)) = 0 \ \forall \theta \in \Theta$ folgt

$$g\left(\hat{\theta}\left(X_1, \ldots, X_n\right)\right) \overset{\text{f.s.}}{\equiv} 0\,.$$

Bemerkung 3.5.5

1. Seien $g_1, g_2 : \mathbb{R}^m \to \mathbb{R}$ Funktionen, für die $\forall \theta \in \Theta$ gilt

$$\mathbb{E}_\theta \left| g_i\left(\hat{\theta}\left(X_1, \ldots, X_n\right)\right)\right| < \infty\,, \quad \mathbb{E}_\theta g_1\left(\hat{\theta}\left(X_1, \ldots, X_n\right)\right) = \mathbb{E}_\theta g_2\left(\hat{\theta}\left(X_1, \ldots, X_n\right)\right)\,,$$

wobei $\hat{\theta}$ vollständig ist. So folgt aus der Definition 3.5.5

$$g_1\left(\hat{\theta}\left(X_1,\ldots,X_n\right)\right) = g_2\left(\hat{\theta}(X_1,\ldots,X_n)\right)$$

fast sicher (nehme $g = g_1 - g_2$).

Fazit: Die Eigenschaft der Vollständigkeit erlaubt aus dem Vergleich der Schätzer $g_1(\hat{\theta})$ und $g_2(\hat{\theta})$ im Mittel eine Aussage über ihre fast sichere Gleichheit zu machen.

2. Falls $\hat{\theta}$ ein vollständiger Schätzer für θ ist, dann ist auch $g(\hat{\theta})$ ein vollständiger Schätzer für θ für eine beliebige messbare Funktion $g : \mathbb{R}^m \to \mathbb{R}^n$.

Beispiel 3.5.6

1. *Bernoulli-Verteilung:* Seien $X_i \sim Bernoulli(p)$, $p \in [0,1]$. Zeigen wir, dass $\hat{p} = \bar{X}_n$ vollständig ist:

   Sei g eine beliebige Funktion $\mathbb{R} \to \mathbb{R}$. Es genügt zu zeigen, dass $Y = \sum_{i=1}^n X_i$ vollständig ist. Es gilt $Y \sim Bin(n,p)$, womit folgt, dass

$$\mathbb{E}_p g(Y) = \sum_{k=0}^n g(k)\binom{n}{k}p^k(1-p)^{n-k}\,.$$

   Weiter gilt $\mathbb{E}_p g(Y) = 0$ genau dann, wenn

$$\sum_{k=0}^n g(k)\binom{n}{k}\Bigg(\underbrace{\frac{p}{1-p}}_{=t}\Bigg)^k = p_n(t) = 0$$

   für $p \in (0,1)$, also $t \in (0,\infty)$. $p_n(t)$ ist ein Polynom des Grades n, womit folgt

$$g(k)\binom{n}{k} = 0 \quad \forall k \quad \Longrightarrow \quad g(k) = 0 \quad \forall k = 0,\ldots,n \quad \Longrightarrow \quad g(Y) = 0 \quad \mathbb{P}_p\text{-fast sicher.}$$

   Somit ist Y vollständig und daher auch $\hat{p} = \bar{X}_n$.

2. *Gleichverteilung:* Sei $X_i \sim U[0,\theta]\,, \quad i = 1,\ldots,n$. Wie wir bereits gezeigt haben, ist der Schätzer $\hat{\theta}(X_1,\ldots,X_n) = \frac{n+1}{n}X_{(n)}$ erwartungstreu. Zeigen wir nun, dass er ein vollständiger Schätzer ist. Es genügt zu zeigen, dass $X_{(n)} = \max_{i=1,\ldots,n} X_i$ vollständig ist. Es ist zu zeigen, dass für alle messbaren $g : \mathbb{R} \to \mathbb{R}$ aus $\mathbb{E}_\theta g(X_{(n)}) = 0$ folgt $g(X_{(n)}) = 0$ fast sicher. Die Dichte von $X_{(n)}$ ist nach Beispiel 3.5.2 gegeben durch $f_{X_{(n)}}(x) = \frac{nx^{n-1}}{\theta^n} \cdot \mathbb{I}_{[0,\theta]}(x)$.

$$\begin{aligned}
0 &= \frac{d}{d\theta}\mathbb{E}_\theta g(X_{(n)}) = \frac{d}{d\theta}\int_0^\theta g(x)f_{X_{(n)}}(x)\,dx = \frac{d}{d\theta}\frac{1}{\theta^n}\int_0^\theta nx^{n-1}g(x)\,dx \\
&= -n\frac{1}{\theta^{n+1}}\int_0^\theta g(x)nx^{n-1}\,dx + \frac{1}{\theta^n}n\theta^{n-1}g(\theta) = -\frac{n}{\theta}\underbrace{\mathbb{E}_\theta g(X_{(n)})}_{=0} + \frac{n}{\theta}g(\theta) \\
&= \frac{n}{\theta}g(\theta) = 0\,, \quad \forall\theta > 0 \Longrightarrow g(x) = 0\,, \quad \forall x > 0.
\end{aligned}$$

   Daher gilt $g(X_{(n)}) = 0$ fast sicher.

3.5.5 Bester erwartungstreuer Schätzer

Aus Definition 3.3.7 folgt: Sei $(X_1, \ldots, X_n)$ eine Zufallsstrichprobe, $X_i \sim F_\theta$, $\theta \in \Theta \subset \mathbb{R}$ $(m = 1)$, X_i unabhängig identisch verteilte Zufallsvariablen. Dann heißt $\hat{\theta}(X_1, \ldots, X_n)$ bester erwartungstreuer Schätzer, falls

$$\mathbb{E}_\theta \hat{\theta}^2(X_1, \ldots, X_n) < \infty \quad \forall \theta \in \Theta\,, \qquad \mathbb{E}_\theta \hat{\theta}(X_1, \ldots, X_n) = \theta \quad \forall \theta \in \Theta$$

und $\hat{\theta}$ die minimale Varianz unter allen erwartungstreuen Schätzern besitzt.

Lemma 3.5.3 (Eindeutigkeit der besten erwartungstreuen Schätzer):
Falls $\hat{\theta}$ ein bester erwartungstreuer Schätzer für θ ist, dann ist er eindeutig bestimmt.

Beweis Sei $\hat{\theta} = \hat{\theta}(X_1, \ldots, X_n)$ ein bester erwartungstreuer Schätzer für θ und $\tilde{\theta}$ ein weiterer fester erwartungstreuer Schätzer für θ. Zeigen wir, dass $\hat{\theta} = \tilde{\theta}$.

Ex adverso: Nehmen wir an, dass $\hat{\theta} \neq \tilde{\theta}$ ist und betrachten $\theta^* = 1/2(\hat{\theta} + \tilde{\theta})$. Offensichtlich ist θ^* erwartungstreu. Untersuchen wir

$$\mathrm{Var}_\theta \theta^* = \frac{1}{4}\mathrm{Var}_\theta(\hat{\theta} + \tilde{\theta}) = \frac{1}{4}\mathrm{Var}_\theta \hat{\theta} + \frac{1}{4}\mathrm{Var}_\theta \tilde{\theta} + \frac{1}{2}\mathrm{Cov}_\theta(\hat{\theta}, \tilde{\theta})\,.$$

Da $\hat{\theta}, \tilde{\theta}$ beste erwartungstreue Schätzer sind und mit der Ungleichung von Cauchy-Schwarz $|\mathrm{Cov}_\theta(\hat{\theta}, \tilde{\theta})| \leq \sqrt{\mathrm{Var}_\theta \hat{\theta} \cdot \mathrm{Var}_\theta \tilde{\theta}} = \mathrm{Var}_\theta \hat{\theta}$ gilt, folgt

$$\mathrm{Var}_\theta \theta^* \leq \frac{1}{2}\mathrm{Var}_\theta \hat{\theta} + \frac{1}{2}\mathrm{Var}_\theta \hat{\theta} = \mathrm{Var}_\theta \hat{\theta}\,.$$

Da $\hat{\theta}$ der beste erwartungstreue Schätzer ist folgt $\mathrm{Var}_\theta \theta^* = \mathrm{Var}_\theta \hat{\theta}$ und somit $\varrho(\hat{\theta}, \tilde{\theta}) = 1 \Longrightarrow \hat{\theta}$ und $\tilde{\theta}$ sind linear abhängig, d.h. es existieren Konstanten a und b, für die gilt $\hat{\theta} = a\tilde{\theta} + b$. Es folgt $a = 1$ aus $\mathrm{Var}_\theta \hat{\theta} = a^2 \mathrm{Var} \tilde{\theta} = \mathrm{Var}_\theta \hat{\theta}$ und $b = 0$, weil $\hat{\theta}$ und $\tilde{\theta}$ erwartungstreu sind: $\theta = \mathbb{E}_\theta \hat{\theta} = \mathbb{E}_\theta \tilde{\theta} + b = \theta + b$. Das bedeutet, dass $\hat{\theta} = \tilde{\theta}$. □

Lemma 3.5.4
Ein erwartungstreuer Schätzer $\hat{\theta}$, dessen zweites Moment endlich ist, ist genau dann der beste erwartungstreue Schätzer für θ, wenn $\mathrm{Cov}_\theta(\hat{\theta}, \varphi) = 0\,, \quad \theta \in \Theta$ für eine beliebige Stichprobenfunktion $\varphi : \mathbb{R}^n \to \mathbb{R}$ mit der Eigenschaft $\mathbb{E}_\theta \varphi(X_1, \ldots, X_n) = 0\,, \quad \forall \theta \in \Theta$.

Beweis Wir beweisen den Satz für beide Richtungen getrennt:

„$\Longrightarrow$“ Sei $\hat{\theta}$ der beste erwartungstreue Schätzer für θ, $\varphi(X_1, \ldots, X_n)$ eine Stichprobenfunktion mit $\mathbb{E}_\theta \varphi(X_1, \ldots, X_n) = 0\,, \quad \forall \theta \in \Theta$. So ist zu zeigen, dass $\mathrm{Cov}_\theta(\hat{\theta}, \varphi) = \mathbb{E}_\theta(\hat{\theta}\varphi) = 0\,, \quad \forall \theta \in \Theta$ gilt.

Definieren wir $\tilde{\theta} = \hat{\theta} + a\varphi\,, \quad a \in \mathbb{R}$. Berechnen wir

$$\mathrm{Var}_\theta \tilde{\theta} = \mathrm{Var}_\theta \hat{\theta} + a^2 \mathrm{Var}_\theta \varphi + 2a\mathrm{Cov}_\theta(\hat{\theta}, \varphi)$$

für $a \in \mathbb{R}$. Sei $g(a) = a^2 \mathrm{Var}_\theta \varphi + 2a\mathrm{Cov}_\theta(\varphi, \hat{\theta})$. Falls $\mathrm{Cov}_\theta(\varphi, \hat{\theta}) \neq 0$, dann existiert ein $a \in \mathbb{R}$ mit $g(a) < 0$. Da $\tilde{\theta}$ ein erwartungstreuer Schätzer für θ ist $(\mathbb{E}_\theta \tilde{\theta} = \mathbb{E}_\theta \hat{\theta} + a\mathbb{E}_\theta \varphi = \theta + 0 = \theta)$ folgt $\mathrm{Var}_\theta \tilde{\theta} \geq \mathrm{Var}_\theta \hat{\theta}$ für alle $a \in \mathbb{R}$. Dies ist jedoch ein Widerspruch mit $g(a) < 0$ für ein $a \in \mathbb{R}$. Damit folgt $\mathrm{Cov}_\theta(\varphi, \hat{\theta}) = 0\,, \quad \forall \theta \in \Theta$.

„$\Longleftarrow$“ Sei $\hat{\theta}$ erwartunsgtreu, $\mathbb{E}_\theta\hat{\theta}^2 < \infty \quad \forall\theta \in \Theta$, $\mathrm{Cov}_\theta(\varphi, \hat{\theta}) = 0\,, \quad \forall\theta \in \Theta$, falls $\mathbb{E}_\theta\varphi = 0 \quad \forall\theta \in \Theta$. Sei $\tilde{\theta}$ ein anderer erwartungstreuer Schätzer für θ. Zeigen wir, dass $\mathrm{Var}_\theta\tilde{\theta} \geq \mathrm{Var}_\theta\hat{\theta}$. Es gilt

$$\tilde{\theta} = \hat{\theta} + \underbrace{(\tilde{\theta} - \hat{\theta})}_{=:\varphi}\,, \quad \mathbb{E}_\theta\varphi = \mathbb{E}_\theta\tilde{\theta} - \mathbb{E}_\theta\hat{\theta} = \theta - \theta = 0\,, \quad \forall\theta \in \Theta\,.$$

Somit

$$\mathrm{Var}_\theta\tilde{\theta} = \mathrm{Var}_\theta\hat{\theta} + \underbrace{\mathrm{Var}_\theta\varphi}_{\geq 0} + 2\underbrace{\mathrm{Cov}_\theta(\hat{\theta}, \varphi)}_{=0} \geq \mathrm{Var}_\theta\hat{\theta}\,,$$

woraus folgt, dass $\hat{\theta}$ der beste Erwartungstreuer Schätzer für θ ist.

□

Satz 3.5.4 (Lehmann-Scheffé):
Sei $\hat{\theta}$ ein erwartungstreuer vollständiger und suffizienter Schätzer für θ, $\mathbb{E}_\theta\hat{\theta}^2 < \infty\,, \quad \forall\theta \in \Theta$. Dann ist $\hat{\theta}$ der beste erwartungstreue Schätzer für θ.

Beweis Nach Lemma 3.5.3 ist zu zeigen, dass $\mathrm{Cov}_\theta(\hat{\theta}, \varphi) = \mathbb{E}_\theta(\hat{\theta}\varphi) = 0\,, \quad \forall\theta \in \Theta$, falls $\mathbb{E}_\theta\varphi = 0 \quad \forall\theta \in \Theta$. Es ist

$$\mathbb{E}_\theta(\hat{\theta}\varphi) = \mathbb{E}_\theta(\mathbb{E}(\hat{\theta}\varphi|\hat{\theta})) \overset{\hat{\theta}\ \sigma(\hat{\theta})\text{-messbar}}{=} \mathbb{E}_\theta(\hat{\theta} \cdot \mathbb{E}_\theta(\varphi|\hat{\theta})) = \mathbb{E}_\theta(\hat{\theta} \cdot g(\hat{\theta})) \overset{?}{=} 0\,,$$

falls $g(\hat{\theta}) = 0$ fast sicher. Da $\hat{\theta}$ suffizient ist, ist $g(t) = \mathbb{E}_\theta(\varphi \,|\, \hat{\theta} = t)$ unabhängig von θ. Betrachten wir $\mathbb{E}_\theta g(\hat{\theta})$. Wir wollen zeigen, dass $\mathbb{E}_\theta g(\hat{\theta}) = 0\,, \quad \forall\theta \in \Theta$. Daraus und aus der Vollständigkeit von $\hat{\theta}$ wird folgen, dass $g(\hat{\theta}) = 0$ fast sicher $\forall\theta \in \Theta$.

$$\mathbb{E}_\theta g(\hat{\theta}) = \mathbb{E}_\theta(\mathbb{E}_\theta(\varphi|\hat{\theta})) = \mathbb{E}_\theta\varphi = 0$$

nach Voraussetzung. Somit folgt $\mathbb{E}_\theta(\varphi\hat{\theta}) = 0$ und $\hat{\theta}$ ist unkorreliert mit $\varphi : \mathbb{E}_\theta\varphi = 0\,, \quad \forall\theta \in \Theta$, womit olgt, dass nach Lemma 3.5.3 $\hat{\theta}$ der beste erwartungstreue Schätzer ist. □

Satz 3.5.5
Sei $\hat{\theta}$ ein erwartungstreuer Schätzer für θ, $\mathbb{E}_\theta\hat{\theta}^2 < \infty \quad \forall\theta \in \Theta$. Sei $\tilde{\theta}$ ein vollständiger und suffizienter Schätzer für θ. Dann ist der Schätzer $\theta^* = \mathbb{E}(\hat{\theta} \,|\, \tilde{\theta})$ der beste erwartungstreue Schätzer für θ.

Beweis

1. Zeigen wir, dass $\mathbb{E}_\theta\theta^{*2} < \infty \quad \forall\theta \in \Theta$. Es gilt

$$\mathbb{E}_\theta\left(\theta^{*2}\right) = \mathbb{E}_\theta\left(\mathbb{E}\left(\hat{\theta} \,|\, \tilde{\theta}\right)\right)^2 \leq \mathbb{E}_\theta\left(\mathbb{E}\left(\hat{\theta}^2 \,|\, \tilde{\theta}\right)\right) = \mathbb{E}_\theta\hat{\theta}^2 < \infty\,,$$

da mit der Ungleichung von Jensen für bedingte Erwartung gilt

$$f(\mathbb{E}(X \,|\, \mathcal{B})) \overset{\text{f.s.}}{\leq} \mathbb{E}(f(X) \,|\, \mathcal{B})$$

für jede Zufallsvariable X, σ-Algebra $\mathcal{B}$ und konvexe Funktion f.

2. Zeigen wir, dass θ^* erwartungstreu ist: $\mathbb{E}_\theta\theta^* = \mathbb{E}_\theta(\mathbb{E}(\hat{\theta} \,|\, \tilde{\theta})) = \mathbb{E}_\theta\hat{\theta} = \theta \quad \forall\theta \in \Theta$, weil $\hat{\theta}$ erwartungstreu ist.

3. Nach Lemma 3.5.3 genügt es zu zeigen, dass $\mathbb{E}_\theta(\theta^*\varphi) = 0 \quad \forall\theta \in \Theta$, falls $\mathbb{E}_\theta\varphi = 0 \quad \theta \in \Theta$.

$$\mathbb{E}_\theta(\theta^*\varphi) = \mathbb{E}_\theta(\underbrace{\mathbb{E}(\hat{\theta}\,|\,\tilde{\theta})}_{=g(\tilde{\theta}),\,\tilde{\theta}\text{ suf.}}\varphi) = \mathbb{E}_\theta(g(\tilde{\theta})\varphi) = \mathbb{E}_\theta(\mathbb{E}(g(\tilde{\theta})\varphi\,|\,\tilde{\theta}))$$
$$\overset{g(\tilde{\theta})\ \tilde{\theta}\text{-messbar}}{=} \mathbb{E}_\theta(g(\tilde{\theta})\cdot\underbrace{\mathbb{E}(\varphi\,|\,\tilde{\theta})}_{=g_1(\tilde{\theta})}) = 0\,,$$

falls $g_1(\tilde{\theta}) \overset{\text{f.s.}}{=} 0 \quad \forall\theta \in \Theta$. Zeigen wir, dass $\mathbb{E}_\theta g_1(\tilde{\theta}) = 0$. Es gilt $\mathbb{E}_\theta g_1(\tilde{\theta}) = \mathbb{E}_\theta(\mathbb{E}(\varphi\,|\,\tilde{\theta})) = \mathbb{E}_\theta\varphi = 0$ nach Voraussetzung. Daraus und aus der Vollständigkeit von $\tilde{\theta}$ folgt genauso wie im Beweis des Satzes 3.5.4, dass $g_1(\tilde{\theta}) = 0$ fast sicher.

□

Lemma 3.5.5 (Ungleichung von Blackwell-Rao):
Sei $\hat{\theta}$ ein erwartungstreuer Schätzer für θ, $\mathbb{E}_\theta\hat{\theta}^2 < \infty\,, \quad \forall\theta \in \Theta$. Sei $\tilde{\theta}$ ein suffizienter Schätzer für θ. Dann besitzt der erwartungstreue Schätzer $\theta^* := \mathbb{E}(\hat{\theta}\,|\,\tilde{\theta})$ eine Varianz, die kleiner oder gleich als $\mathrm{Var}_\theta\hat{\theta}$ ist.

Beweis Siehe Beweis des Satzes 3.5.5. Dabei folgt die Erwartungstreue von θ^* aus Beweispunkt 2) des Satzes 3.5.5 und $\mathrm{Var}_\theta\theta^* = \mathbb{E}_\theta\theta^{*2} - \theta^2 \leq \mathbb{E}_\theta\hat{\theta}^2 - \theta^2 = \mathrm{Var}_\theta\hat{\theta}$ aus Beweispunkt 1) des Satzes 3.5.5. □

Folgerung 3.5.2
Falls $\hat{\theta}$ ein vollständiger und suffizienter Schätzer für θ ist und falls eine Funktion $g : \mathbb{R} \to \mathbb{R}$ so existiert, dass $\mathbb{E}_\theta g(\hat{\theta}) = \theta \quad \forall\theta \in \Theta$, dann ist $g(\hat{\theta})$ der beste erwartungstreue Schätzer für θ.

Beweis $g(\hat{\theta}) = \mathbb{E}(g(\hat{\theta})\,|\,\hat{\theta})$, welcher nach Satz 3.5.5 der beste erwartungstreue Schätzer ist. □

Literaturverzeichnis

[1] Bickel, P. ; Doksum, K.: *Mathematical Statistics: Basic Ideas and Selected Topics.* London : Prentice Hall, 2001. – 2nd ed., Vol. l

[2] Borovkov, A. A.: *Mathematical Statistics.* Gordon & Breach, 1998

[3] Burkschat, M. ; Cramer, E. ; Kamps, U.: *Beschreibende Statistik, Grundlegende Methoden.* Berlin : Springer, 2004

[4] Casella, G. ; Berger, R. L.: *Statistical Inference.* Duxbury : Pacific Grove (CA), 2002

[5] Cramer, E. ; Cramer, K. ; Kamps, U. ; Zuckschwerdt: *Beschreibende Statistik, Interaktive Grafiken.* Berlin : Springer, 2004

[6] Cramer, E. ; Kamps, U.: *Grundlagen der Wahrscheinlichkeitsrechnung und Statistik.* Berlin : Springer, 2007

[7] Dalgaard, P.: *Introductory Statistics with R.* Berlin : Springer, 2002

[8] Fahrmeir, L. ; Kneib, T. ; Lang, S.: *Regression. Modelle, Methoden und Anwendungen.* Berlin : Springer, 2007

[9] Fahrmeir, L. ; Künstler, R. ; Pigeot, I. ; Tutz, G.: *Statistik. Der Weg zur Datenanalyse.* Berlin : Springer, 2001

[10] Georgii, H. O.: *Stochastik: Einführung in die Wahrscheinlichkeitstheorie und Statistik.* Berlin : de Gruyter, 2002

[11] Hartung, J. ; Elpert, B. ; Klösener, K. H.: *Statistik.* München : R. Oldenbourg Verlag, 1993. – 9. Auflage

[12] Heyde, C. C. ; Seneta, E.: *Statisticians of the Centuries.* Berlin : Springer, 2001

[13] Irle, A.: *Wahrscheinlichkeitstheorie und Statistik, Grundlagen – Resultate – Anwendungen.* Teubner, 2001

[14] Kazmir, L. J.: *Wirtschaftsstatistik.* McGraw-Hill, 1996

[15] Koch, K. R.: *Parameter Estimation and Hypothesis Testing in Linear Models.* Berlin : Springer, 1999

[16] Krause, A. ; Olson, M.: *The Basics of S-PLUS.* Berlin : Springer, 2002. – Third Ed.

[17] Krengel, U.: *Einführung in die Wahrscheinlichkeitstheorie und Statistik.* Braunschweig : Vieweg, 2002. – 6. Auflage

[18] Lehmann, E. L.: *Elements of Large-Sample Theory.* New York : Springer, 1999

[19] Lehn, J. ; Wegmann, H.: *Einführung in die Statistik.* Stuttgart : Teubner, 2000. – 3. Auflage

[20] Maindonald, J. ; Braun, J.: *Data Analysis and Graphics Using R.* Cambridge University Press, 2003

[21] Overbeck-Larisch, M. ; Dolejsky, W.: *Stochastik mit Mathematica.* Braunschweig : Vieweg, 1998

[22] Pruscha, H.: *Angewandte Methoden der Mathematischen Statistik.* Stuttgart : Teubner, 1996

[23] Pruscha, H.: *Vorlesungen über Mathematische Statistik.* Stuttgart : Teubner, 2000

[24] Sachs, L.: *Angewandte Statistik.* Springer, 1992

[25] Sachs, L. ; Hedderich, J.: *Angewandte Statistik, Methodensammlung mit R.* Berlin : Springer, 2006

[26] Shiryaev, A. N.: *Probability.* New York : Springer, 1996

[27] Spiegel, M. R. ; Stephens, L. J.: *Statistik.* McGraw-Hill, 1999

[28] Stahel, W. A.: *Statistische Datenanalyse.* Vieweg, 1999

[29] Venables, W. ; Ripley, D.: *Modern applied statistics with S-PLUS.* Springer, 1999. – 3rd ed

[30] Wasserman, L.: *All of Statistics. A Concise Course in Statistical Inference.* Springer, 2004

Index

A
a-posteriori-Verteilung . . . 82
a-priori-Verteilung . . . 82
absolute Häufigkeit . . . 7
Abweichung, mittlere quadratische . . . 15
arithmetisches Mittel . . . 12
asymptotisch erwartungstreu . . . 50
asymptotisch normalverteilt . . . 51
Ausgangsvariable . . . 33

B
Balkendiagramm . . . 8
Bandbreite . . . 25
Bayes-Schätzer . . . 83
Bayesche Formel . . . 82
bedingte Wahrscheinlichkeit . . . 95
bedingter Erwartungswert . . . 94
besserer Schätzer . . . 51
bester erwartungstreuer Schätzer . . . 51
Bestimmtheitsmaß . . . 37
Bias . . . 50
bimodal . . . 10
Blackwell-Rao, Ungleichung von . . . 104
Bootstrap
  Konfidenzintervall . . . 88
  Schätzer . . . 87
Bootstrap-Schätzer
  Monte-Carlo-Methoden . . . 87
Box-Plot . . . 14
  modifizierter . . . 14
Bravais-Pearson-Koeffizient . . . 31
Bravais-Pearson-Korrelationskoeffizient . . . 29
Brownsche Brücke . . . 69

C
χ^2-Verteilung . . . 43
Cramér-Rao, Ungleichung von . . . 88

D
Daten-Stichproben . . . 1
Datenbereinigung . . . 1
Datenerhebung . . . 1
Dichteschätzung . . . 25
Dichtetransformationssatz für Zufallsvektoren . . . 47
gleichmäßiger Abstand D_n . . . 64
Dvoretzky-Kiefer-Wolfowitz, Ungleichung von . . . 66

E
Eindeutigkeit der besten erwartungstreuen Schätzer . . . 102
Einfache lineare Regression . . . 33
Einflussfaktor . . . 33
empirische(r)
  Kovarianz . . . 29
  Median . . . 14
  Standardabweichung . . . 15
  Varianz . . . 15
  Variationskoeffizient . . . 15, 17
  Verteilungsfunktion . . . 10
Erlangverteilung . . . 43
erwartungstreu . . . 49
Erwartungstreue . . . 16
Erwartungswert, bedingter . . . 94
Explorative Datenanalyse . . . 1

F
Faktorisierungssatz von Neyman-Fisher . . . 99
Fehler 1. Art . . . 69
Fisher
  Fisher-Information . . . 78
  Fisher-Snedecor-Verteilung, F-Verteilung . . . 47
  Wölbungsmaß von Fisher . . . 21

G
Gammaverteilung
  Faltungsstabilität . . . 43
  Momenterzeugende und charakteristische Funktion . . . 42
geometrisches Mittel . . . 12, 13
Gesamtstreuung . . . 37
getrimmtes Mittel . . . 12
Gini-Koeffizient . . . 17, 18
Gini-Koeffizient, Darstellung von . . . 18
Gliwenko-Cantelli, Satz von . . . 64
Grundgesamtheit . . . 1

H
Häufigkeit
  absolute . . . 7
  relative . . . 7

harmonisches Mittel . . . 13
Herfindahl-Index . . . 17
Histogramm . . . 7
    eindimensionales Histogramm . . . 7
    zweidimensionales Histogramm . . . 27

I
identifizierbar . . . 40
Information von Kullback-Leibler . . . 75
Invarianzeigenschaften . . . 32

J
Jackknife-Schätzer für die/den
    Erwartungswert . . . 85
    Varianz . . . 86
    Verzerrung (Bias) . . . 86

K
Kerndichteschätzer
    eindimensionaler Kerndichteschätzer . . . 25
    zweidimensionaler Kerndichteschätzer . . . 29
Kolmogorow, Satz von . . . 68
Kolmogorow-Abstand D_n . . . 64
Kolmogorow-Verteilung . . . 68
Konfidenzband . . . 66
Konfidenzintervall . . . 57
    Bootstrap . . . 88
konsistenter Schätzer . . . 50
Konstanzbereich . . . 67
Konzentrationsrate . . . 17, 20
Korrelationskoeffizient . . . 29
    Spearmans . . . 31
Kovarianz, empirische . . . 29
Kreisdiagramm . . . 7, 8
Kullback-Leibler, Information von . . . 75
Kurtosis . . . 21

L
Lagemaß . . . 12
Lehmann-Scheffé, Satz von . . . 103
Lernstichprobe . . . 3
Likelihood-Funktion . . . 73
linksschief . . . 10
linkssteil . . . 10
Lorenz-Münzner-Koeffizient . . . 20
Lorenzkurve . . . 17

M
maximale Streuung . . . 15
Maximum-Likelihood-Schätzer . . . 73
    schwache Konsistenz . . . 76
Median . . . 12, 15
    empirischer . . . 14
Mittel
    arithmetisches . . . 12
    geometrisches . . . 12, 13
    getrimmtes . . . 12
    harmonisches . . . 12, 13
Mittelwert . . . 12, 15
mittlere quadratische Abweichung . . . 15
mittlerer quadratischer Fehler . . . 50
Modalität . . . 10
Modellierung von Daten . . . 1
Modellvalidierung . . . 1
modifizierter Box-Plot . . . 14
Modus . . . 12, 15
Momentenmethode . . . 71
Momentenschätzer . . . 71
multimodal . . . 10

N
Neyman-Fisher, Faktorisierungssatz . . . 99

O
Ordnungsstatistik . . . 6, 12, 13

P
Parameterraum . . . 40
Parametervektor . . . 40
Plug-in-Methode . . . 69
Plug-in-Schätzer . . . 69, 70
Polynomiale Regression . . . 33
Punktschätzer . . . 40

Q
Quantil . . . 12, 13
Quantilplot . . . 21
Quartil . . . 12, 14

R
Rangkorrelationskoeffizient . . . 31
Realisierung . . . 5, 7
rechtsschief . . . 10
rechtssteil . . . 10
Regressand . . . 33
Regression . . . 33
    einfache lineare . . . 33
    polynomiale . . . 33
Regressionsgerade . . . 35
Regressionsgerade, Eigenschaften von . . . 37
Regressionskoeffizient . . . 34
Regressionskonstante . . . 34
Regressionsvarianz . . . 34
Regressor . . . 33
relative Häufigkeit . . . 7
Resampling-Methode . . . 84
Residualplot . . . 39
Residuen . . . 35

S
Säulendiagramm . . . 8
Satz
- χ^2-Verteilung, Spezialfall . . . 43
- Darstellung des Gini-Koeffizient . . . 18
- Dichte der t-Verteilung . . . 45
- Dichtetransformationssatz für Zufallsvektoren 47
- Eigenschaften der empirischen Momente . . . 51
- Eigenschaften des bedingten Erwartungswertes . . . 94
- Faktorisierungssatz von Neyman-Fisher . . . 99
- Gliwenko-Cantelli . . . 64
- Invarianzeigenschaften . . . 32
- Kolmogorow . . . 68
- Lehmann-Scheffé . . . 103
- Momenterzeugende und charakteristische Funktion der Gammaverteilung . . . 42
- Schwache Konsistenz von ML-Schätzern . . . 76
- Ungleichung von Cramér-Rao . . . 88
- Ungleichung von Dvoretzky-Kiefer-Wolfowitz 66

Schätzer . . . 49
- besserer . . . 51
- bester erwartungstreuer . . . 51
- konsistenter . . . 50
- suffizienter . . . 96
- Vergleich von . . . 51

Schiefe . . . 12, 20
Schließende Datenanalyse . . . 2
Spannweite . . . 15
Spearmans Korrelationskoeffizient . . . 31
Stabdiagramm . . . 7
Stamm-Blatt-Diagramm . . . 9
Standardabweichung . . . 16
Statistische Merkmale . . . 4
stem-leaf display . . . 9
Stichproben . . . 5
Stichprobenfunktion . . . 6
Stichprobenmittel . . . 6, 12
Stichprobenvarianz . . . 6, 15
Streudiagramm . . . 27, 39
suffizienter Schätzer . . . 96
sum of squared residuals . . . 37
sum of squares explained . . . 37
sum of squares total . . . 37
Symmetriekoeffizient . . . 20
symmetrisch . . . 10

T
t-Verteilung . . . 45
Tortendiagramm . . . 8
Transformationsregel . . . 16

U
unimodal . . . 10, 76
unverzerrt . . . 49

V
Varianz, empirische . . . 15
Verlustfunktion . . . 82
verteilungsfrei . . . 66
Verteilungsfunktion, empirische . . . 10
Vertrauensintervall . . . 57
Verzerrung . . . 50
Visualisierung . . . 1
Vollständigkeit . . . 100

W
Wölbung . . . 12
Wölbungsmaß von Fisher . . . 21

Z
Zielgröße . . . 33
Zufallsstichprobe . . . 5
Zufallsvektoren
- Dichtetransformationssatz . . . 47

zweidimensionaler Kerndichteschätzer . . . 29
zweidimensionales Histogramm . . . 27