Einführung in die Wahrscheinlichkeitstheorie und Statistik

Prof. Dr. Rainer Dahlhaus

20. August 2012

Falls Sie Fehler finden, schicken Sie bitte eine mail an dahlhaus@statlab.uni-heidelberg.de

Konzept der Vorlesung:

Eine einführende Vorlesung in die Wahrscheinlichkeitstheorie und/oder Statistik wird traditionell an fast jeder Universität gelesen. In so einer Vorlesung (wie auch in dieser "Grundvorlesung Wahrscheinlichkeitstheorie und Statistik") werden die Grundbegriffe des Gebietes wie Wahrscheinlichkeitsverteilungen, Zufallsvariable, Erwartungswerte, Varianz, Kovarianz, stochastische Unabhängigkeit, das schwache Gesetz der großen Zahlen u.a. behandelt. In zwei Punkten unterscheidet sich diese Vorlesung jedoch von den meisten anderen:

(i) Die Vorlesung enthält in wesentlich größerem Umfang als allgemein üblich auch Elemente der Statistik (dafür werden Elemente der Wahrscheinlichkeitstheorie wie z.B. eine elementare Behandlung der Markov-Ketten weggelassen). In dem "klassischen Zyklus" der mathematischen Statistik folgt auf eine Einführung in die Stochastik (oft ohne Statistik oder mit sehr wenig Statistik) eine Kursusvorlesung Wahrscheinlichkeitstheorie I und frühestens im Anschluss daran eine erste (dann oft sehr theoretische) Kursusvorlesung in Statistik. Bei so einem Zyklus lernen die Studierenden die einfachen Konzepte der Statistik sehr spät oder überhaupt nicht kennen. Aus diesem Grunde lese ich vor den notwendigen Kursusvorlesungen in Wahrscheinlichkeitstheorie zunächst eine Grundvorlesung mit möglichst vielen Elementen der Statistik, um bei den Studierenden das Interesse an Statistik zu wecken und um ihnen ein gewisses minimales Wissen in Statistik mit auf den Weg zu geben. Nach dem Hören dieser Grundvorlesung sollten Sie in der Lage sein, sich aus der Vielzahl der vorhandenen Bücher zu dem Thema weitere elementare Methoden der Statistik selbst anzueignen. Für anspruchsvollere Verfahren oder ein tieferes Verständnis sollten Sie später eine Kursusvorlesung in Statistik besuchen.

(ii) Ein klassisches Dilemma der Grundvorlesung besteht darin, dass man für die Behandlung stetiger Verteilungen wie der Normalverteilung die Maßtheorie benötigt. Die notwendigen Grundkenntnisse der Maßtheorie stehen den Studierenden in den Anfangssemestern aber leider nicht zur Verfügung. Trotzdem werden stetige Verteilungen als so wichtig erachtet, dass sie heute in keiner Grundvorlesung fehlen. Dieses führt zwangsläufig zu mathematischen Ungenauigkeiten oder künstlichen Konstruktionen, die auch in den meisten elementaren Büchern zu finden sind[1]. Nach mehreren eigenen Versuchen habe ich

mich entschlossen, in der vorliegenden Vorlesung eine saubere Konstruktion mit einem Minimum an maßtheoretischen Begriffen zu verwenden und dabei aber auf die meisten Beweise (die den Rahmen der Grundvorlesung sprengen würden) zu verzichten. Diese Beweise werden im Rahmen der Kursusvorlesung Wahrscheinlichkeitstheorie I nachgeholt. Der Vorteil diese Vorgehens ist, dass die Studierenden einen sauberen Zugang haben, der mit dem späteren Stoff deckungsgleich ist.

Die ersten Kapitel der Vorlesung beinhalten eine Einführung in die (im Vergleich zur Analysis und Linearen Algebra) neuen Techniken und Denkweisen der Stochastik. Studierende, die damit bereits von der Schule her vertraut sind, empfinden diesen ersten Teil der Vorlesung meistens als sehr elementar und "verschlafen" dann den Einstieg in den anspruchsvolleren zweiten Teil. Von der Mathematik her verwende ich in der Grundvorlesung vor allem Techniken der LA I wie Projektionen, Orthogonalität und die Diagonalisierung von Matrizen. Das ist Standardstoff der LA I und nicht schwierig - wird aber von den Studierenden erfahrungsgemäß als schwierig empfunden. In höheren Semestern bedient sich die Statistik vieler anspruchsvoller Techniken aus anderen Gebieten der Mathematik, wie z.B. aus der Maßtheorie, Funktionalanalysis, Topologie, den Differentialgleichungen etc. Beispielsweise sind stochastische Prozesse (wie die Brownsche Bewegung) dann Zufallsvariable mit Werten in metrischen Räumen, die zugehörigen Wahrscheinlichkeitsverteilungen dementsprechend Maße auf metrischen Räumen. Es ist eine spannende Herausforderung in der Statistik, diese abstrakten mathematischen Konzepte mit angewandten statistischen Fragestellungen der Datenanalyse zu verknüpfen.

Ich danke meiner Sekretärin Elke Carlow für das Tippen zahlreicher Kapitel, Herrn Jonas Peters für das Erstellen mehrerer Grafiken, sowie meinen Kollegen Jan Johannes, Matthias Katzfuß und Ilya Pavlyukevich sowie vielen Studierenden für zahlreiche Verbesserungsvorschläge.

Notation: Bemerkungen in eckigen Klammern sind mündliche Kommentare, die in der Vorlesung idR nicht angeschrieben werden, z.B. [bitte nachrechnen].

Inhaltsverzeichnis

1 Wahrscheinlichkeitsverteilungen

In diesem Kapitel werden Wahrscheinlichkeitsverteilungen definiert und elementare Eigenschaften hergeleitet. Das wesentliche Beispiel ist die Laplace-Verteilung. Ferner werden die Grundformeln der Kombinatorik bewiesen. An zwei Beispielen wird eine stochastische Modellierung demonstriert.

Bemerkung: Die Darstellung ist in sich abgeschlossen - trotzdem sollten fehlende Schulkenntnisse in Stochastik durch zusätzliches Studium der Literatur und zusätzliches Lösen von Übungsaufgaben ausgeglichen werden.

Definition 1.1 (Axiome von Kolmogorov, 1933) Sei Ω eine nichtleere Menge und $\mathcal{A} \subset \mathcal{P}(\Omega)$ eine σ-Algebra (s. unten). Eine Abbildung $\mathbf{P} : \mathcal{A} \to [0,1]$ heißt Wahrscheinlichkeitsverteilung auf $(\Omega, \mathcal{A})$, falls gilt
(i) $\mathbf{P}(\Omega) = 1$;
(ii) $\mathbf{P}(A) \geq 0$ für alle $A \in \mathcal{A}$;
(iii) $\mathbf{P}\big(\bigcup_{i=1}^{\infty} A_i\big) = \sum_{i=1}^{\infty} \mathbf{P}(A_i)$ für alle paarweise disjunkten $A_i \in \mathcal{A}$ (σ-Additivität).

Ω ist die Menge aller möglichen Ergebnisse und heißt Stichprobenraum, Mengen $A \in \mathcal{A}$ [d.h. $A \subset \Omega$] heißen Ereignisse. Das Tripel $(\Omega, \mathcal{A}, \mathbf{P})$ heißt Wahrscheinlichkeitsraum.

Bem.: (i) σ-Algebren werden in Kapitel 6 definiert. Bis dahin nehmen wir immer an, dass $\mathcal{A}$ die Potenzmenge von Ω ist, d.h. $\mathcal{A} := \mathcal{P}(\Omega) := \{A \,|\, A \subset \Omega\}$. [für abzählbare Ω kann man immer $\mathcal{A} = \mathcal{P}(\Omega)$ verwenden; für überabzählbare Ω muss man die Menge der möglichen Ereignisse leider einschränken - deshalb verwendet man σ-Algebren]

(ii) Verbal: $P(A)$ ist die W't, dass das Ereignis A eintritt, $P(A \cap B)$ ist z.B. die W't, dass das Ereignis A und das Ereignis B eintreten, $P(A \cup B)$ ist die W't, dass das Ereignis A oder das Ereignis B eintritt, $P(A^c)$ ist die W't, dass das Ereignis A nicht eintritt.

(iii) Man schreibt auch $\sum_{i=1}^{\infty} A_i$ anstelle von $\bigcup_{i=1}^{\infty} A_i$ falls $A_i \cap A_j = \emptyset \quad \forall i \neq j$.

Beispiel 1.2 (Laplace-Verteilung)

$$\Omega \text{ endlich} \quad , \quad \mathcal{A} = \mathcal{P}(\Omega) \quad , \quad \mathbf{P}(A) := \frac{|A|}{|\Omega|} \quad \Big(= \frac{\#\text{ günstige Fälle}}{\#\text{ mögliche Fälle}} \Big). \tag{1}$$

Bei einer Laplace-Verteilung ordnet man jedem möglichen Ergebnis die gleiche Wahrscheinlichkeit $1/|\Omega|$ zu.

Konkrete Beispiele: (i) Einmaliges Würfeln mit einem Würfel:

$$\Omega = \{1, \ldots, 6\}, \quad A = \{6\} \ \text{[Ereignis eine "6" zu würfeln]}, \quad \mathbf{P}(A) = \tfrac{|A|}{|\Omega|} = \tfrac{1}{6}$$

(ii) Zweimaliges Würfeln. W't eine "10" als Summe zu erhalten?

$$\Omega = \{(1,1), (1,2), \ldots, (6,6)\}, \quad |\Omega| = 6^2 = 36;$$

$$A = \{(4,6), (5,5), (6,4)\}, \quad |A| = 3, \quad \mathbf{P}(A) = \tfrac{3}{36}$$

Weitere Ereignisse:

"Pasch": $B = \{(1,1), (2,2), \ldots, (6,6)\}, \quad \mathbf{P}(B) = \tfrac{6}{36}$;

"Summe ungerade": $C = \{(1,2), (1,4), \ldots\}, \quad \mathbf{P}(C) = \tfrac{18}{36}$.

$\mathbf{P}(A \cap B) = \mathbf{P}(\{(5,5)\}) = \tfrac{1}{36}$.

□

Beispiel 1.3 Werfen einer Reißzwecke. Sei o (bzw. u) das Ergebnis, dass sie mit der Spitze nach oben (unten) fällt, d.h.

$$\Omega = \{o, u\} \quad p = \mathbf{P}(\{o\})$$

p ist idR. unbekannt mit $p \neq \frac{1}{2} = \frac{1}{|\Omega|}$, d.h. $\mathbf{P}$ ist keine Laplace-Verteilung.
[Unterscheidung: zufällig-unbekannt. $\mathbf{P}(p = \frac{1}{2})$ ist unsinnig.] □

Satz 1.4 (Eigenschaften einer Wahrscheinlichkeitsverteilung)
Für $A, B, A_i \in \mathcal{A}$ gilt

(i) $\mathbf{P}(A^c) = 1 - \mathbf{P}(A)$ *insbesondere* $\mathbf{P}(\emptyset) = 0$*;*
(ii) $A \subset B \quad \Rightarrow \quad \mathbf{P}(A) \leq \mathbf{P}(B)$*;*
(iii) $\mathbf{P}(\bigcup_{i=1}^n A_i) \leq \sum_{i=1}^n \mathbf{P}(A_i)$*;*
(iv) $\mathbf{P}(A \cup B) = \mathbf{P}(A) + \mathbf{P}(B) - \mathbf{P}(A \cap B)$.

Beweis. trivial [Aussagen an Mengenbildern verdeutlichen!]
z.B. (iv) $A \cup B = A \backslash B + B \backslash A + A \cap B$ ["+" ist disjunkte Vereinigung]

$$\begin{aligned} \Rightarrow \quad \mathbf{P}(A \cup B) &= \mathbf{P}(A \backslash B) + \mathbf{P}(B \backslash A) + \mathbf{P}(A \cap B) \\ &= \mathbf{P}(A) + \mathbf{P}(B) - \mathbf{P}(A \cap B) \quad (\text{da } A = A \backslash B + A \cap B) \end{aligned}$$

□

Beispiel 1.5 (Laplace-Verteilung + Motivation für Kombinatorik)
Wette: "Unter den Hörern der Vorlesung sind mindestens zwei, die am gleichen Tag Geburtstag haben." (*) Ich wäre bereit, 1 EURO gegen 1 EURO zu setzen. Ist die Wette für Sie interessant? [Abstimmung]

Idee: Sei $A = (*)$ Wette lohnend $\Leftrightarrow \mathbf{P}(A) > \mathbf{P}(A^c)$
$\Leftrightarrow \mathbf{P}(A) > \frac{1}{2}$.

Problem: Liegt überhaupt ein Experiment vor. Der Ausgang ist nicht zufällig, sondern nur unbekannt! [Diskussion]

Modell: Jeder Hörer hat zufällig seinen Geburtstag ausgewählt aus 365 Tagen.

Annahmen (zur Vereinfachung): Es gibt kein Schaltjahr;
Alle Geburtstage sind gleich-wahrscheinlich.

Wahrscheinlichkeitsraum: $\Omega = \{(\omega_1, \ldots, \omega_r) \,|\, \omega_i \in \{1, \ldots, 365\}\}$ ($r = \#$ Hörer);
$\mathbf{P}$ Laplace-Verteilung;
$A = \{(\omega_1, \ldots, \omega_r) \in \Omega \,|\, \exists\, i \neq j : \omega_i = \omega_j\}$;
$A^c = \{(\omega_1, \ldots, \omega_r) \in \Omega \,|\, \omega_i \neq \omega_j \ \forall i \neq j\}$.

Damit gilt wegen $e^x \approx 1 + x$ für $|x|$ klein

$$\begin{aligned}\mathbf{P}(A) &= 1 - \mathbf{P}(A^c) = 1 - \frac{|A^c|}{|\Omega|} \\ &= 1 - \frac{365 \cdot 364 \cdots (365 - r + 1)}{365^r} \\ &= 1 - [1 \cdot (1 - \frac{1}{365}) \cdots (1 - \frac{r-1}{365})] \\ &\approx 1 - \Big[1 \cdot \underbrace{\exp\Big\{ \sum_{k=1}^{r-1} \big(- \frac{k}{365} \big) \Big\}}_{e^{-r(r-1)/730}} \Big] \approx 1 - e^{-r^2/730}.\end{aligned}$$

[da $\exp\{r/730\} \approx 1$]. Für $r = 30\,;\, 40\,;\, 50$ folgt $\mathbf{P}(A) = 0,71\,;\, 0,89\,;\, 0,97$.

Wir haben folgende Formeln aus der Kombinatorik verwendet:

$365 \cdot 364 \cdots (365 - r + 1)$	Ziehen (in Reihenfolge) ohne Zurücklegen;
365^r	Ziehen (in Reihenfolge) mit Zurücklegen.

□

Satz 1.6 (Kombinatorik) *Beim Ziehen einer Stichprobe vom Umfang r aus n Elementen gibt es folgende Anzahl von Möglichkeiten:*

	mit Zurücklegen	*ohne Zurücklegen*
in Reihenfolge	(i) n^r	(ii) $n(n-1)\cdots(n-r+1)$
ohne Reihenfolge	(iv) $\binom{n+r-1}{r}$	(iii) $\binom{n}{r}$

Beweis.

(i) Klar. Formal verwendet man

$$\Omega_1 := \Big\{(\omega_1, \ldots, \omega_r) \,\Big|\, \omega_r \in \{1, \ldots, n\}\Big\} \quad \text{mit} \quad |\Omega_1| = n^r.$$

(ii) Sei

$$\Omega_2 := \Big\{(\omega_1, \ldots, \omega_r) \,\Big|\, \omega_r \in \{1, \ldots, n\}, \omega_i \neq \omega_j \; \forall i \neq j\Big\}.$$

Offensichtlich gilt $|\Omega_2| = n(n-1)\cdots(n-r+1)$.

(iii) Sei

$$\Omega_3 := \Big\{\{\omega_1, \ldots, \omega_r\} \,\Big|\, \omega_r \in \{1, \ldots, n\}, \omega_i \neq \omega_j \; \forall i \neq j\Big\}.$$

Es gilt $|\Omega_3| = \frac{|\Omega_2|}{r!} = \frac{n\cdots(n-r+1)}{r!} = \binom{n}{r}$. [$r!$ = # Permutationen von r Elementen]
Für den Beweis von (iv) halten wir fest, dass $|\Omega_3| = |\Omega_3'(n)|$ mit

$$\Omega_3'(n) := \Big\{(\omega_1, \ldots, \omega_r) \,\Big|\, 1 \leq \omega_1 < \ldots < \omega_r \leq n\Big\}.$$

(iv) Sei

$$\Omega_4(n) := \Big\{(\omega_1, \ldots, \omega_r) \,\Big|\, 1 \leq \omega_1 \leq \ldots \leq \omega_r \leq n\Big\}.$$

Offensichtlich ist $h : \Omega_4(n) \to \Omega_3'(n+r-1)$ mit

$$(\omega_1, \ldots, \omega_r) \longmapsto (\omega_1, \omega_2 + 1, \ldots, \omega_r + r - 1)$$

eine Bijektion. Damit gilt

$$|\Omega_4(n)| = |\Omega_3'(n+r-1)| = \binom{n+r-1}{r}.$$

□

Beispiel 1.7 (Lotto) Es gibt $\frac{49\cdot 48\cdots 44}{6!} = \binom{49}{6} = 13.983.816$ Möglichkeiten aus 49 Zahlen 6 auszuwählen.

Beispiel 1.8 (Schätzung eines Wildbestandes) In einem See seien N Fische. Um N zu schätzen, werden M (z.B. $M = 10$) Fische gefangen, markiert und wieder ausgesetzt. Nach einiger Zeit werden n ($n = 20$) Fische gefangen, darunter befinden sich k ($k = 2$) markierte. Wie viele Fische sind im See? Was "glauben" Sie?

Diskussion:

(i) Mindestens 28 Fische! Ist das auch eine plausible Schätzung?
(ii) Höchstens? Unbeschränkt viele.
(iii) Gilt exakt $\frac{N}{M} = \frac{n}{k}$, so folgt $N = 10 \cdot \frac{20}{2} = 100$. Wäre 99 oder 101 weniger plausibel?

Wir wollen zunächst den Vorgang wahrscheinlichkeitstheoretisch beschreiben. Nehmen wir an, es wären genau $N = 100$ Fische im See. Wie groß ist dann die Wahrscheinlichkeit, dass von den $n = 20$ gefangenen Fischen genau $k = 2$ markiert sind?

Kenngrößen:

N	unbekannter (oder bekannter) Parameter	nicht zufällig
$M = 10$	bekannter Parameter	nicht zufällig
$n = 20$	bekannter Parameter	nicht zufällig
k	(bekanntes) Ergebnis des Experiments	zufällig
	(Experiment = Fangen der Fische)	

[Konsequenz: $\mathbf{P}(N = 100)$ ist unsinnig!!]

Modell: Sei F_i der i-te Fisch, d.h. wir haben

$$\underbrace{F_1, \ldots, F_M}_{M \text{ markierte}}, \quad \underbrace{F_{M+1}, \ldots, F_N}_{N-M \text{ unmarkierte}}.$$

- Sei Ω der Stichprobenraum, d.h. die Menge aller möglichen Ergebnisse unseres Fangexperiments von n ($n = 20$) Fischen:

$$\Omega = \Big\{ A \subset \{F_1, \ldots, F_N\} \,\Big|\, |A| = n \Big\}.$$

- Sei $\mathbf{P}$ die Laplace-Verteilung auf Ω, d.h. $\mathbf{P}(A) := \frac{|A|}{|\Omega|}$ mit $|\Omega| = \binom{N}{n}$. [Ziehen ohne Reihenfolge ohne Zurücklegen]
  Wir setzen damit insbesondere voraus, dass alle Ergebnisse unseres Fangexperiments gleich wahrscheinlich sind. [Diskussion dieser Annahme]

In diesem Modell sei nun E_k das Ereignis, dass von den n gefangenen Fischen genau k markiert sind:

$$E_k = \Big\{ A \subset \{F_1, \ldots, F_N\} \,\Big|\, |A \cap \{F_1, \ldots, F_M\}| = k, \quad |A \cap \{F_{M+1}, \ldots, F_N\}| = n - k \Big\}.$$

Es gilt

$$|E_k| = \binom{M}{k} \binom{N-M}{n-k}$$

und damit

$$\mathbf{P}(E_k) = \mathbf{P}_{N,M,n}(E_k) = \frac{|E_k|}{|\Omega|} = \frac{\binom{M}{k}\binom{N-M}{n-k}}{\binom{N}{n}}.$$

Zahlenbeispiel:

$$N = 100 \quad M = 10$$
$$n = 20 \qquad k = 2$$
$$\mathbf{P}(E_2) = \frac{\binom{10}{2}\binom{90}{18}}{\binom{100}{20}} \approx 0,32.$$

Statistische Problemstellung: Schätze N ! [nochmal: N ist unbekannt, aber nicht zufällig!] Eine Möglichkeit ist, N durch den Wert zu schätzen, für den die Wahrscheinlichkeit $\mathbf{P}_{N,M,n}(E_k)$ (mit festen M, n, k) maximal wird, d.h. maximiere

$$L_N = \frac{\binom{M}{k}\binom{N-M}{n-k}}{\binom{N}{n}}$$

bzgl. N (Maximum Likelihood Schätzer). Es gilt [nachrechnen!]

$$\frac{L_N}{L_{N-1}} = \ldots = \frac{(N-M)(N-n)}{(N-M-n+k)N}$$

d.h.

$$\begin{aligned}
\frac{L_N}{L_{N-1}} > 1 &\Leftrightarrow (N-M)(N-n) > (N-M-n+k)N \\
&\Leftrightarrow N^2 - MN - Nn + Mn > N^2 - MN - nN + kN \\
&\Leftrightarrow Mn > kN \\
&\Leftrightarrow \frac{nM}{k} > N \qquad \text{d.h. Maximum } N_{\max} \approx \frac{nM}{k} \ (= 100).
\end{aligned}$$

[Der Maximum Likelihood Schätzers wird ausführlich in Kapitel 15 behandelt].

Bemerkung: Es macht offensichtlich wenig Sinn, für N eine einzelne Zahl anzugeben (z.B. $N = 100$ - es kämen genauso gut 99 oder 101 in Betracht). Gesucht ist sinnvollerweise ein Intervall, das den unbekannten Parameter N mit hoher Wahrscheinlichkeit enthält. Das führt zu sog. 'Konfidenzintervallen' (s. Kapitel 9). □

2 Bedingte Wahrscheinlichkeit und Stochastische Unabhängigkeit

In diesem Kapitel werden die bedingte Wahrscheinlichkeit und die stochastische Unabhängigkeit für Ereignisse definiert. Als wichtige Eigenschaften werden die Formel von Bayes und der Satz von der totalen Wahrscheinlichkeit bewiesen. In einem Beispiel wird angedeutet, wie die Formel von Bayes in Expertensystemen verwendet wird.

Beispiel 2.1 (Kongestive Herzinsuffizienz)
Kongestive Herzinsuffizienz = krankhafte Blutvolumenzunahme innerhalb der Herzkammern [Kongestion=Blutüberfüllung].
Man wendet u.a. eine Digitalis Therapie an. Mögliche Nebenwirkung: Digitalis Vergiftung.

Beller u.a. (1971) haben Daten über 135 Digitalis-Patienten ausgewertet:

T_+: Test auf hohe Digitalis-Konzentration im Blut positiv;
T_-: Test auf hohe Digitalis-Konzentration im Blut negativ;
V_+: Digitalis Vergiftung (Vergiftungs-Symptome vorhanden);
V_-: keine Digitalis Vergiftung (keine Symptome vorhanden).

Man erhielt die Häufigkeitstafel

	V_+	V_-	Summe
T_+	25	14	39
T_-	18	78	96
Summe	43	92	135

Die relative Häufigkeit, dass der Test positiv ist und keine Vergiftung vorliegt, beträgt $H(T_+ \cap V_-) = \frac{14}{135} = 0.104$. Analog gibt es eine unbekannte Wahrscheinlichkeit $\mathbf{P}(T_+ \cap V_-)$ für dieses Ereignis. [Wie man von der relativen Häufigkeit bei einer Stichprobe auf die

Wahrscheinlichkeit schließt, soll später behandelt werden.] Wir haben aber die (berechtigte) Vorstellung, dass die einzelnen Wahrscheinlichkeiten nah bei den relativen Häufigkeiten liegen. Man erhält die folgende Tafel der relativen Häufigkeiten

	V_+	V_-	$\sum$
T_+	0.19	0.10	0.29
T_-	0.13	0.58	0.71
$\sum$	0.32	0.68	1.00

Frage: *Angenommen ein Arzt weiß, dass der Test ein positives Ergebnis gezeigt hat. Wie groß ist dann die Wahrscheinlichkeit für eine Vergiftung?* Wir schreiben hierfür $P(V_+ \,|\, T_+)$. Für die relative Häufigkeit gilt

$$H(V_+ \,|\, T_+) = \frac{25}{39} = \frac{25/135}{39/135} = \frac{0.19}{0.29} = \frac{H(V_+ \cap T_+)}{H(T_+)} = 0.64$$

während $H(V_+) = 0.32$ und (analog berechnet) $H(V_+ \,|\, T_-) = 0.19$. Falls der Test negativ ist, so beträgt die (bedingte) Häufigkeit keine Vergiftung zu haben

$$H(V_- \,|\, T_-) = \frac{78}{96} = \frac{78/135}{96/135} = \frac{0.58}{0.71} = \frac{H(V_- \cap T_-)}{H(T_-)} = 0.81$$

während $H(V_-) = 0.68$ und (analog berechnet) $H(V_- \,|\, T_+) = 0.36$.

Analog definieren wir nun die bedingte Wahrscheinlichkeit.

Definition 2.2 Seien A und B zwei Ereignisse mit $\mathbf{P}(B) > 0$. Dann ist die bedingte Wahrscheinlichkeit von A gegeben B definiert durch

$$\mathbf{P}(A \,|\, B) := \frac{\mathbf{P}(A \cap B)}{\mathbf{P}(B)}\,.$$

Satz 2.3 *Ist* $\mathbf{P}$ *eine Wahrscheinlichkeitsverteilung auf* $\mathcal{A}$ *und* $B \in \mathcal{A}$ *mit* $\mathbf{P}(B) > 0$*, so ist auch* $\mathbf{P}(\,\cdot \,|\, B)$ *eine W'verteilung auf* $\mathcal{A}$*.* [$\mathbf{P}(A \,|\, \cdot\,)$ ist jedoch keine W'verteilung !!!]

Beweis. Ü-Aufgabe □

Satz 2.4 (Satz von der totalen Wahrscheinlichkeit) *Seien* $B_1, \ldots, B_n \in \mathcal{A}$ *mit* $\cup_{i=1}^n B_i = \Omega$, $B_i \cap B_j = \emptyset \; \forall\, i \neq j$ *und* $\mathbf{P}(B_i) > 0 \; \forall i$. *Dann gilt für alle* $A \in \mathcal{A}$

$$\mathbf{P}(A) = \sum_{i=1}^{n} \mathbf{P}(A \,|\, B_i)\, \mathbf{P}(B_i).$$

Beweis. [Situation $\cup_{i=1}^n B_i = \Omega$ mit Mengenbild skizzieren!]

$$\begin{aligned}
\mathbf{P}(A) &= \mathbf{P}(A \cap \Omega) = \mathbf{P}(A \cap (\cup_{i=1}^n B_i)) = \mathbf{P}(\cup_{i=1}^n (A \cap B_i)) \\
&= \sum_{i=1}^{n} \mathbf{P}(A \cap B_i) \qquad \text{da} \quad (A \cap B_i) \cap (A \cap B_j) = \emptyset \qquad \forall\, i \neq j \\
&= \sum_{i=1}^{n} \mathbf{P}(A \,|\, B_i)\, \mathbf{P}(B_i).
\end{aligned}$$

□

Beispiel 2.5 Ziehen von 2 Karten aus einem Skatblatt. Wahrscheinlichkeit, beim zweiten Zug ein Ass zu ziehen (Ereignis A)?

$$B_1 \cong \text{ Ass im ersten Zug;}$$

$$B_2 := B_1^c \cong \text{ kein Ass im ersten Zug.}$$

Es gilt

$$\begin{aligned}
\mathbf{P}(A) &= \mathbf{P}(A|B_1)\, \mathbf{P}(B_1) + \mathbf{P}(A|B_2)\, \mathbf{P}(B_2) \\
&= \frac{3}{31} \cdot \frac{4}{32} + \frac{4}{31} \cdot \frac{28}{32} \\
&= \frac{1}{31 \cdot 32}\,(12 + 112) = \frac{124}{31 \cdot 32} = \frac{1}{8}.
\end{aligned}$$

Wie groß ist $\mathbf{P}(B_1|A)$?

$$\begin{aligned}\mathbf{P}(B_1|A) &= \frac{\mathbf{P}(B_1 \cap A)}{\mathbf{P}(A)} = \frac{\mathbf{P}(A|B_1)\,\mathbf{P}(B_1)}{\mathbf{P}(A|B_1)\,\mathbf{P}(B_1) + \mathbf{P}(A|B_2)\,\mathbf{P}(B_2)} \\ &= \frac{\frac{3}{31} \cdot \frac{1}{8}}{\frac{1}{8}} = \frac{3}{31}.\end{aligned}$$

□

Satz 2.6 (Formel von Bayes) *Seien $A, B_1, \ldots, B_n \in \mathcal{A}$ mit $\cup_{i=1}^n B_i = \Omega$, $B_i \cap B_j = \emptyset$ $\forall i \neq j$, $\mathbf{P}(A) > 0$, $\mathbf{P}(B_j) > 0$ $\forall j$. Dann gilt*

$$\mathbf{P}(B_j \,|\, A) = \frac{\mathbf{P}(A \,|\, B_j)\,\mathbf{P}(B_j)}{\sum_{i=1}^n \mathbf{P}(A \,|\, B_i)\,\mathbf{P}(B_i)}\ .$$

Beweis. Analog zu oben:

$$\mathbf{P}(B_j \,|\, A) = \frac{\mathbf{P}(B_j \cap A)}{\mathbf{P}(A)} = \frac{\mathbf{P}(A \,|\, B_j)\,\mathbf{P}(B_j)}{\sum_{i} \mathbf{P}(A \,|\, B_i)\,\mathbf{P}(B_i)}\ .$$

□

Beispiel 2.7 (Diagnose von Angina Pectoris)
Die Krankheit „Angina Pectoris“ (Brustenge) bedeutet eine Unterversorgung des Herzmuskels mit Sauerstoff. Die Symptome sind Herzschmerzen und körperliche Untätigkeit. Zur Diagnose wird eine Herz-Fluoroskopie durchgeführt, bei der die Anzahl der verkalkten Koronararterien bestimmt wird.

Es werde zufällig ein Patient ausgewählt: [Experiment erklären; zufällig-unbekannt].

A_+:	Der Patient hat Angina Pectoris;
A_-:	Der Patient hat kein Angina Pectoris;
T_0, T_1, T_2, T_3:	Der Patient hat 0, 1, 2, 3 verkalkte Arterien.

Die folgenden Wahrscheinlichkeiten setzen wir als bekannt voraus: [z.B. ”geschätzt” durch die relativen Häufigkeiten bei einer sehr großen Stichprobe]

i	$\mathbf{P}(T_i\|A_+)$	$\mathbf{P}(T_i\|A_-)$
0	0,42	0,96
1	0,24	0,02
2	0,20	0,02
3	0,14	0,00

Es gilt

$$\mathbf{P}(A_+ \,|\, T_i) = \frac{\mathbf{P}(T_i \,|\, A_+)\, \mathbf{P}(A_+)}{\mathbf{P}(T_i \,|\, A_+)\, \mathbf{P}(A_+) \;+\; \mathbf{P}(T_i \,|\, A_-)\, \mathbf{P}(A_-)} \,.$$

Falls man nun auch $\mathbf{P}(A+)$ kennt (und damit auch $\mathbf{P}(A-) = 1 - \mathbf{P}(A+)$), kann man die bedingten Wahrscheinlichkeiten $\mathbf{P}(A_+ \,|\, T_i)$ bestimmen. Beispiele:

1) $\mathbf{P}(A_+) = 0,05$ (für einen Mann, Alter 30-39, der wegen Brustschmerzen eine Praxis aufsucht)

   Die Formel von Bayes ergibt dann: $\mathbf{P}(A_+ \,|\, T_0) = 0,02$ und $\mathbf{P}(A_+ \,|\, T_1) = 0,39$.

2) $\mathbf{P}(A_+) = 0,92$ (für einen Mann, Alter 50-59, der wegen Brustschmerzen eine Praxis aufsucht)

   Die Formel von Bayes ergibt dann: $\mathbf{P}(A_+ \,|\, T_0) = 0,83$ und $\mathbf{P}(A_+ \,|\, T_1) = 0,99$.

Die Formel von Bayes wird auf diese Art auch in größeren „Expertensystemen“ verwendet.

□

Stochastische Unabhängigkeit:

Intuition: Zwei Ereignisse sollten unabhängig sein, wenn

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

und

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

Definition 2.8 (Unabhängigkeit)

(i) Zwei Ereignisse A und B heißen stochastisch unabhängig falls

$$\mathbf{P}(A \cap B) = \mathbf{P}(A)\,\mathbf{P}(B).$$

(ii) n Ereignisse $A_1, \ldots, A_n$ heißen gemeinsam stochastisch unabhängig falls

$$\mathbf{P}(A_{i_1} \cap \ldots \cap A_{i_m}) = \mathbf{P}(A_{i_1}) \cdots \mathbf{P}(A_{i_m}) \tag{2}$$

für alle $\{i_1, \ldots i_m\} \subset \{1, \ldots, n\}$.

Bemerkung 2.9 (i) (ii) garantiert gerade, dass die Information über das Eintreten einer Teilmenge der Ereignisse A_i keine Information über das Eintreten eines anderen Ereignisses A_{i_0} enthält, also z.B.

$$\mathbf{P}(A_1 \,|\, A_2 \cap \ldots \cap A_k) = \frac{\mathbf{P}(A_1 \cap \ldots \cap A_k)}{\mathbf{P}(A_2 \cap \ldots \cap A_k)} = \frac{\prod_{i=1}^{k} \mathbf{P}(A_i)}{\prod_{i=2}^{k} \mathbf{P}(A_i)} = \mathbf{P}(A_1).$$

(ii) Stochastische Unabhängigkeit bedeutet nicht, dass zwei Mengen "nichts miteinander gemeinsam haben", also disjunkt sind!

Beispiel 2.10 Ziehen einer Karte aus einem Kartenspiel

$$\Omega = \Big\{ \underset{\substack{\uparrow \\ \text{(Farbe, Karte)}}}{(i,j)} \;\Big|\; i \in \{1,\ldots,4\}\,,\, j \in \{1,\ldots,13\} \Big\}, \quad \mathbf{P} \text{ Laplace-Verteilung}$$

$$\begin{aligned} A &= \big\{(i,1) \,|\, i \in \{1,\ldots,4\}\big\} && \text{(es wird ein Ass gezogen)} \\ H &= \big\{(1,j) \,|\, j \in \{1,\ldots,13\}\big\} && \text{(es wird Herz gezogen)} \\ A \cap H &= \{(1,1)\} \end{aligned}$$

Es gilt

$$\begin{aligned} \mathbf{P}(A) &= \frac{|A|}{|\Omega|} = \frac{4}{52} = \frac{1}{13}\,, \quad \mathbf{P}(H) = \frac{|H|}{|\Omega|} = \frac{13}{52} = \frac{1}{4}, \\ \mathbf{P}(A \cap H) &= \frac{|A \cap H|}{|\Omega|} = \frac{1}{52} = \mathbf{P}(A)\,\mathbf{P}(H), \end{aligned}$$

d.h. A und H sind stochastisch unabhängig. □

[Nachfolgendes Beispiel evtl. weglassen. Das Beispiel ist aber insofern wichtig, als es die möglichen Konsequenzen einer Unabhängigkeitsannahme bei einer angewandten Modellierung demonstriert.]

Beispiel 2.11 (Chemische Anlage, 4 Kühlsysteme) $A_i \cong$ Ereignis, dass das Kühlsystem i während eines Produktionsvorgangs ausfällt. Sei $\mathbf{P}(A_i) = 0,01$.

$A_1, \ldots, A_4$ stoch. unabh. $\quad\Rightarrow\quad \mathbf{P}(A_1 \cap \ldots \cap A_4) = \prod \mathbf{P}(A_i) = 0.01^4 = 10^{-8}$.
[diskutieren: Extremfall der Abhängigkeit z.B. bei einem Stromkreislauf, Sicherheitsstudien in Kernkraftwerken] □

Beispiel 2.12 (Werfen einer fairen Münze)

$\Omega = \{(K,K),(K,Z),(Z,K),(Z,Z)\}$, $\mathbf{P}$ Laplace-Verteilung;

$A \cong$ beim 1. Wurf "Kopf",
$A = \{(K,K),(K,Z)\}$, $\mathbf{P}(A) = 1/2$;

$B \cong$ beim 2. Wurf "Kopf",
$B = \{(K,K,)(Z,K)\}$, $\mathbf{P}(B) = 1/2$;

$\mathbf{P}(A \cap B) = 1/4$;

$C \cong$ genau einmal Kopf,
$C = \{(Z,K),(K,Z)\}$, $\mathbf{P}(C) = 1/2$;

$\mathbf{P}(A \cap C) = 1/4\,, \mathbf{P}(B \cap C) = 1/4\,.$

aber: $\mathbf{P}(A \cap B \cap C) = \mathbf{P}(\emptyset) = 0 \neq \mathbf{P}(A)\,\mathbf{P}(B)\,\mathbf{P}(C)$.

<u>d.h.</u> paarweise unabh. $\not\Rightarrow$ gemeinsam unabh. □

Satz 2.13 *Seien die Ereignisse $A_1, \ldots, A_n$ gemeinsam stochastisch unabhängig. Dann sind auch*
$B_1, \ldots, B_n$ mit $B_i \in \{A_i, A_i^c\}$ gemeinsam stochastisch unabhängig.

Beweis. Induktion über n.

$$
\begin{aligned}
&(A_1 \cap A_2) + (A_1 \cap A_2^c) = A_1 \\
&\quad \Rightarrow\ \mathbf{P}(A_1 \cap A_2) + \mathbf{P}(A_1 \cap A_2^c) = \mathbf{P}(A_1) \\
&\qquad\quad \underset{\mathbf{P}(A_1)\,\mathbf{P}(A_2)}{\|} \\
&\quad \Rightarrow\ \mathbf{P}(A_1 \cap A_2^c) = \mathbf{P}(A_1)\,[1 - \mathbf{P}(A_2)] = \mathbf{P}(A_1)\,\mathbf{P}(A_2^c) \\
&\Rightarrow\ \mathbf{P}(A_1^c \cap A_2) = \mathbf{P}(A_1^c)\,\mathbf{P}(A_2) \quad \text{(durch Vertauschung der Indices)} \\
&\Rightarrow\ \mathbf{P}(A_1^c \cap A_2^c) = \mathbf{P}(A_1^c)\,\mathbf{P}(A_2^c) \quad \text{(iterativ)}
\end{aligned}
$$

Sei die Beh. für $n-1$ richtig und seien $A_1, \ldots, A_n$ stoch. unabh.

$\Rightarrow B_{i_1}, \ldots, B_{i_m}$ stoch. unabh. $\forall\ \{i_1, \ldots, i_m\} \subset \{1, \ldots, n\}$ mit $m \leq n-1$.

Bleibt zu zeigen: $\mathbf{P}(B_1 \cap \ldots \cap B_n) = \mathbf{P}(B_1) \cdots \mathbf{P}(B_n)$.
[bei Zeitmangel direkt zu Fall $k = 1$ gehen]

Seien o.E. die Komplementmengen A_i^c die ersten B_i, d.h. z.z.

$$\mathbf{P}(A_1^c \cap \ldots \cap A_k^c \cap A_{k+1} \cap \ldots \cap A_n) = \mathbf{P}(A_1^c) \cdots \mathbf{P}(A_k^c)\,\mathbf{P}(A_{k+1}) \cdots \mathbf{P}(A_n).$$

$k = 1$:

$$
\begin{aligned}
\mathbf{P}(A_1^c \cap A_2 \cap \ldots \cap A_n) + \mathbf{P}(A_1 \cap \ldots \cap A_n) &= \mathbf{P}(A_2 \cap \ldots \cap A_n) = \mathbf{P}(A_2) \cdots \mathbf{P}(A_n) \\
&\qquad\quad \underset{\mathbf{P}(A_1) \cdots \mathbf{P}(A_n)}{\|} \\
\Rightarrow\ \mathbf{P}(A_1^c \cap A_2 \cap \ldots \cap A_n) &= \underbrace{[1 - \mathbf{P}(A_1)]}_{=\mathbf{P}(A_1^c)}\,\mathbf{P}(A_2) \cdots \mathbf{P}(A_n).
\end{aligned}
$$

$k-1 \to k$: [analog - evtl. weglassen]

$$
\begin{aligned}
&\mathbf{P}(A_1^c \cap \ldots \cap A_k^c \cap A_{k+1} \cap \ldots \cap A_n) + \mathbf{P}(A_1^c \cap \ldots \cap A_{k-1}^c \cap A_k \cap \ldots \cap A_n) \\
&\qquad = \mathbf{P}(A_1^c \cap \ldots \cap A_{k-1}^c \cap A_{k+1} \cap \ldots \cap A_n) \\
&\qquad = \mathbf{P}(A_1^c) \cdots \mathbf{P}(A_{k-1}^c)\,\mathbf{P}(A_{k+1}) \cdots \mathbf{P}(A_n)
\end{aligned}
$$

$$\overset{\text{analog}}{\Rightarrow}\ \mathbf{P}(A_1^c \cap \ldots \cap A_k^c \cap A_{k+1} \cap \ldots \cap A_n) = \mathbf{P}(A_1^c) \cdots \mathbf{P}(A_k^c)\,\mathbf{P}(A_{k+1}) \cdots \mathbf{P}(A_n).$$

□

3 Diskrete Verteilungen

In diesem Kapitel werden die wichtigsten diskreten Verteilungen, d.h. Verteilungen mit abzählbarem Träger definiert.

Definition 3.1 Eine Wahrscheinlichkeitsverteilung heißt diskrete Verteilung falls Ω höchstens abzählbar ist. $p(w) := \mathbf{P}(\{w\})$ heißt Zähldichte der Verteilung.

Bemerkung: Oft gilt $\Omega = \{0, \ldots, n\}$.

Proposition 3.2 *Ist Ω abzählbar, $p : \Omega \to [0,1]$ eine Abbildung mit $\sum_{\omega\in\Omega} p(\omega) = 1$, so ist durch $\mathbf{P}(A) := \Sigma_{\omega\in A}\, p(\omega)\ \forall\ A \in \mathcal{P}(\Omega)$ eindeutig eine Wahrscheinlichkeitsverteilung auf $\mathcal{P}(\Omega)$ definiert.*

Beweis. Es gilt $\mathbf{P}(\Omega) = 1$ und $\mathbf{P}(A) \geq 0$. Die σ-Additivität ist offensichtlich. □

Beispiel 3.3 (Die Bernoulli-Verteilung)

$$\Omega = \{0,1\}\,, \quad p(1) = p \text{ mit } p \in [0,1]\,, \quad p(0) = 1 - p.$$

Beispiel: Werfen einer Reißzwecke. p ist dabei unbekannt. □

Beispiel 3.4 (Die Binomial-Verteilung $\mathcal{B}(n,p)$)

$$\Omega = \{0, \ldots, n\}\,, \quad p(k) = \binom{n}{k} p^k (1-p)^{n-k}.$$

$$\left[\text{ Es gilt } \sum_{k=0}^{n} p(k) = \sum_{k=0}^{n} \binom{n}{k} p^k (1-p)^{n-k} = [p + (1-p)]^n = 1\,.\right]$$

Anwendung: Ein Experiment habe die möglichen Ergebnisse $0, 1$ (Misserfolg / Erfolg) mit Erfolgswahrscheinlichkeit p (z.B. $p = 1/6$ für die "6" beim Würfeln). Wir zeigen, dass die Anzahl der Erfolge bei n unabhängigen Wiederholungen des Experiments $\mathcal{B}(n,p)$-verteilt ist:

Sei $A_j \cong$ Erfolg im j-ten Experiment, $\mathbf{P}(A_j) = p$, $\mathbf{P}(A_j^c) = 1 - p$. Gesucht wird die $W't$ für genau k Erfolge. "Unabhängig" bedeutet hier, dass die $A_1, \ldots, A_n$ stochastisch unabhängig sind. Es gilt (E ist im folgenden die Indexmenge der Erfolge):

$$\begin{aligned}
\mathbf{P}(\,k \text{ Erfolge}\,) &= \mathbf{P}\Bigg(\bigcup_{\substack{E\subset\{1,\ldots,n\}\\ |E|=k}} \Big[\Big\{ \bigcap_{j\in E} A_j \Big\} \cap \Big\{ \bigcap_{j\in E^c} A_j^c \Big\} \Big] \Bigg) \\
&= \sum_{\substack{E\subset\{1,\ldots,n\}\\ |E|=k}} \mathbf{P}\left(\Big\{ \bigcap_{j\in E} A_j \Big\} \cap \Big\{ \bigcap_{j\in E^c} A_j^c \Big\} \right) \\
&= \sum_{E} \Big\{ \prod_{j\in E} \mathbf{P}(A_j) \Big\} \Big\{ \prod_{j\in E^c} \mathbf{P}(A_j^c) \Big\} \\
&= \sum_{E} p^{|E|}\,(1-p)^{|E^c|} \\
&= \binom{n}{k} p^k\,(1-p)^{n-k}\,.
\end{aligned}$$

[Bemerkung: Leser, die bereits Zufallsvariable kennen, sollten zusätzlich Beispiel 5.6 anschauen. Dort wird deutlich, dass obiges $\mathbf{P}$ auf dem ursprünglichen Wahrscheinlichkeitsraum mit $\Omega = \{(\omega_1, \ldots, \omega_n) \mid \omega_i \in \{0,1\}\}$ lebt und die $\mathcal{B}(n,p)$-Verteilung die induzierte Verteilung auf dem Wahrscheinlichkeitsraum $\Omega = \{0, \ldots, n\}$ ist.]

Beispiele:

(i) Es wird 10-mal gewürfelt. Die Wahrscheinlichkeit, dass genau k-mal eine "6" auftritt, beträgt

$$p(k) = \binom{10}{k} \left(\frac{1}{6}\right)^k \left(\frac{5}{6}\right)^{10-k}.$$

(ii) Beide Elternteile sind gesunde Träger einer rezessiven Erbkrankheit. Wie groß ist bei 4 Kindern die Wahrscheinlichkeit, dass k Kinder krank sind

$$p(k) = \binom{4}{k} \left(\frac{1}{4}\right)^k \left(\frac{3}{4}\right)^{4-k}.$$

□

Beispiel 3.5 (Die geometrische Verteilung $\mathcal{G}(p)$)

$$\Omega = \mathbb{N}\,, \quad p(k) = (1-p)^{k-1}p\,.$$

$$\left[\sum_{k=1}^{\infty} p(k) = \sum_{k=1}^{\infty}(1-p)^{k-1}p = \frac{p}{1-(1-p)} = 1\,.\right]$$

Anwendung: Man wiederholt ein Experiment solange (unabhängig), bis zum ersten mal "Erfolg" eintritt. $p(k)$ gibt die Wahrscheinlichkeit für genau k Wiederholungen bis zum Erfolgsfall an.

Beweis: [wie oben - evtl. weglassen] Seien A_j wie oben. Dann gilt

$$\begin{aligned}
\mathbf{P}(k \text{ Wiederholungen}) &= \mathbf{P}\Big(\Big\{\bigcap_{j=1}^{k-1} A_j^c\Big\} \cap A_k\Big) \\
&= \Big\{\prod_{j=1}^{k-1} \mathbf{P}(A_j^c)\Big\}\, \mathbf{P}(A_k) \\
&= (1-p)^{k-1}\, p\,.
\end{aligned}$$

Beispiel: Würfeln, bis das erste Mal eine "6" auftritt. □

Beispiel 3.6 (Die hypergeometrische Verteilung $\mathcal{H}(N, M, n)$)

$$\Omega = \{0, \ldots, n\}\,, \quad p(k) = \frac{\binom{M}{k}\binom{N-M}{n-k}}{\binom{N}{n}}\,.$$

Anwendung: Schätzung Wildbestand/Qualitätskontrolle. □

Beispiel 3.7 (Die Poisson-Verteilung $\mathcal{P}(\lambda)$)

$$\Omega = \mathbb{N}_0\,, \quad p(k) = \frac{\lambda^k}{k!}\, e^{-\lambda} \quad \lambda > 0\,.$$

$$\left[\sum_{k=0}^{\infty} p(k) = \sum_{k=0}^{\infty} \frac{\lambda^k}{k!}\, e^{-\lambda} = e^{-\lambda}\, e^{\lambda} = 1\,.\right]$$

□

Für große n sind die Wahrscheinlichkeiten der $\mathcal{B}(n,p)$-Verteilung schwer zu berechnen. Man verwendet deshalb häufig eine Approximation:

Proposition 3.8 (Approximation der Binomialverteilung)
Sei $p_n = \frac{\lambda}{n}$ mit $\lambda > 0$ (λ fest). Dann gilt für $k \in \mathbb{N}_0$

$$\lim_{n\to\infty} \binom{n}{k} p_n^k (1-p_n)^{n-k} = e^{-\lambda}\frac{\lambda^k}{k!}.$$

Beweis.

$$\begin{aligned}\binom{n}{k}\, p_n^k\, (1-p_n)^{n-k} &= \frac{n!}{k!\,(n-k)!}\,\frac{\lambda^k}{n^k}\,(1-\frac{\lambda}{n})^n\,(1-\frac{\lambda}{n})^{-k}\\ &= \frac{\lambda^k}{k!}\,\frac{n!}{(n-k)!\,n^k}\,(1-\frac{\lambda}{n})^n\,(1-\frac{\lambda}{n})^{-k}\\ &\to \frac{\lambda^k}{k!}\cdot 1\cdot e^{-\lambda}\cdot 1\,.\end{aligned}$$

□

Bemerkung: Damit kann man für hinreichend großes n und np "mittelgroß" die $\mathcal{B}(n,p)$-Verteilung durch eine $\mathcal{P}(\lambda)$-Verteilung approximieren [zur Güte der Approximation vgl. z.B. Krengel]. Die Abhängigkeit von $p = p_n$ von n wird nur für die mathematische Beschreibung verwendet. Sie ist nicht so gemeint, dass p wirklich von n abhängt. Für p "mittelgroß" und np "groß" verwendet man übrigens anstelle der $\mathcal{P}(\lambda)$-Verteilung eine Normalverteilung (siehe W'theorie I).

Beispiel 3.9 In einem Land beträgt die Geburtenrate 730 Geburten pro 100.000 Einwohner pro Jahr. Wie ist die Anzahl der Geburten pro Tag in einer Stadt S mit 280.000 Einwohnern verteilt? Sei p Wahrscheinlichkeit, dass ein einzelner Bewohner an einem festen Tag ein Kind bekommt. (alle gleiche Wahrscheinlichkeit - auch für Männer; bessere Modellierung z.B. über Anzahl der Frauen zwischen 20 und 40) $p = \frac{730}{365}/100.000 = \frac{2}{100.000}$. Wir haben $n = 280.000$ unabhängige "Wiederholungen" des Experiments d.h. n groß, p klein, $np = 5.6$ mittel $\to$ Anzahl der Geburten pro Tag in S ist $\mathcal{P}(5.6)$-verteilt. □

4 Testen von Parametern / Neyman-Pearson Lemma

Statistische Tests gehören zu den zentralen Methoden der Statistik. In diesem Kapitel werden als einfaches Beispiel Tests von einelementigen Hypothesen bei diskreten Verteilungen eingeführt. Der optimale Test wird durch das Neyman-Pearson-Lemma gegeben.

Beispiel 4.1 (Einführung statistische Tests)
Das Medikament NERVOSTOP heilt in 60% aller Fälle ein Magengeschwür (bekannt).

Herstellerfirma PHARMAFIX: zusätzlicher Wirkstoff → NERVOSTOP PLUS
Beh.: Heilungsrate 70% → Antrag auf Zulassung

Sachberater SKEPTISCH:
Beh.: zusätzlicher Wirkstoff ist unwirksam (nur teurer); die Heilungsrate ist nach wie vor 60 %

Es wird ein Experiment durchgeführt: von 100 Testpersonen werden 66 geheilt. Soll das Medikament NERVOSTOP PLUS zugelassen werden?

Modell: $\Omega = \underset{\substack{\uparrow \\ \text{(Anzahl möglicher Erfolge)}}}{\{0, \ldots, n\}}, \quad n = 100;$

[Modellbildung kurz diskutieren, vor allem die Unabhängigkeit]

Frage: liegt $\mathbf{P} = \mathcal{B}(n, p), \quad p = 0,6 \quad$ (Hypothese H_0)
oder $\mathbf{Q} = \mathcal{B}(n, q), \quad q = 0,7 \quad$ (Alternativhypothese H_A) vor?

Bemerkung (trivial): Es können sowohl unter $\mathbf{P}$ wie unter $\mathbf{Q}$ 66 Personen geheilt worden sein, nur mit den unterschiedlichen Wahrscheinlichkeiten

$$\mathbf{P}(\{66\}) = \binom{100}{66}(0,6)^{66}(0,4)^{34} = \ldots \approx 0,04,$$

und

$$\mathbf{Q}(\{66\}) = \binom{100}{66}(0,7)^{66}(0,3)^{34} = \ldots \approx 0,06.$$

Aufgabe: Entscheidung zwischen $\mathbf{P}$ und $\mathbf{Q}$ anhand des Ergebnisses $k = 66$. Gesucht ist eine Entscheidungsfunktion (auch Test genannt)

$$\phi : \Omega \to \{\mathbf{P}, \mathbf{Q}\},$$

die möglichst "optimal" ist.

Problem: Man kann sich falsch entscheiden. Es gibt 2 Typen von Fehlern:

(i) Entscheidung für $\mathbf{Q}$ obwohl $\mathbf{P}$ richtig ist.
Wahrscheinlichkeit: $\mathbf{P}\big(\phi = \mathbf{Q}\big) = \mathbf{P}(\{k \,|\, \phi(k) = \mathbf{Q}\})$ Fehler 1. Art;

(ii) Entscheidung für $\mathbf{P}$ obwohl $\mathbf{Q}$ richtig ist
Wahrscheinlichkeit: $\mathbf{Q}\big(\phi = \mathbf{P}\big) = \mathbf{Q}(\{k \,|\, \phi(k) = \mathbf{P}\})$ Fehler 2. Art.

Gesucht ist nun der "beste" Test ϕ^* (d.h. die beste Entscheidungsfunktion). Was heißt hier "bester" Test? Das Problem bei der Definition ist, dass man nicht beide Fehler gleichzeitig klein machen kann. Beispiel: Setze $\phi \equiv \mathbf{P}$. Dann gilt:

$$\begin{aligned} \mathbf{P}\big(\phi = \mathbf{Q}\big) &= 0 \quad \text{klein!} \\ \mathbf{Q}\big(\phi = \mathbf{P}\big) &= 1 \quad \text{groß!} \end{aligned}$$

Definition bester Test: Die beiden Fehler haben idR unterschiedliche Konsequenzen, deshalb

- gibt man eine Obergrenze für den Fehler 1. Art fest vor (z.B. ≤ 0.01);
- und sucht dann denjenigen Test, der den Fehler 2. Art unter dieser Nebenbedingung minimiert

Mathematisch: Minimiere $\mathbf{Q}\big(\phi = \mathbf{P}\big)$ bzgl. ϕ unter der Nebenbedingung $\mathbf{P}\big(\phi = \mathbf{Q}\big) \leq \alpha$. Falls eine Lösung ϕ^* existiert, so heißt diese bester Test zum Niveau α.

Heuristische Idee für ein gutes ϕ:

$$\phi(k) = \begin{cases} \mathbf{P}, & k < c \\ \mathbf{Q}, & k \geq c \end{cases}$$

wobei c geeignet bestimmt werden muss.

Mit dem folgenden Lemma zum Testen von zwei einfachen Hypothesen werden wir zeigen, dass dieses wirklich der optimale Test ϕ^* ist. ['einfach' bedeutet in diesem Zusammenhang, dass H_0 und H_A jeweils einelementig sind.]

Satz 4.2 (Neyman-Pearson Lemma) *Sei Ω abzählbar und seien $\mathbf{P}$ und $\mathbf{Q}$ zwei Wahrscheinlichkeitsverteilungen auf Ω mit Zähldichten $p(\omega)$ und $q(\omega)$.*
$L(\omega) := \frac{q(\omega)}{p(\omega)} \leq \infty$ heißt Likelihood-Quotient von $\mathbf{Q}$ bzgl. $\mathbf{P}$. Sei ferner

$$\phi^* : \Omega \to \{\mathbf{P}, \mathbf{Q}\}$$

$$\phi^*(\omega) := \begin{cases} \mathbf{P}, & L(\omega) < c^* \\ \mathbf{Q}, & L(\omega) \geq c^* \end{cases}$$

mit $\mathbf{P}\big(\phi^* = \mathbf{Q}\big) = \mathbf{P}(L(\omega) \geq c^*) = \alpha$ (*). Dann gilt für jede andere Entscheidungsfunktion $\phi : \Omega \to \{\mathbf{P}, \mathbf{Q}\}$ mit $\mathbf{P}\big(\phi = \mathbf{Q}\big) \leq \alpha$

$$\mathbf{Q}\big(\phi = \mathbf{P}\big) \geq \mathbf{Q}\big(\phi^* = \mathbf{P}\big),$$

d.h. ϕ^* minimiert $\mathbf{Q}\big(\phi = \mathbf{P}\big)$ unter der Nebenbedingung $\mathbf{P}\big(\phi = \mathbf{Q}\big) \leq \alpha$.

[Die Existenz eines Tests ϕ^* mit (*) wird hier angenommen. Diese Bedingung ist allerdings bei diskreten Verteilungen für viele α nicht erfüllt.]

Beweis.
Seien $A^* := \{\omega \,|\, \phi^*(\omega) = \mathbf{Q}\}$ und $A := \{\omega \,|\, \phi(\omega) = \mathbf{Q}\}$ (Ablehnbereiche)

Es gilt:

$$\begin{aligned}
& \omega \in A^* \Leftrightarrow q(\omega) - c^* p(\omega) \geq 0 \\
\Rightarrow \quad & \sum_{\omega \in A^*} \big[q(\omega) - c^* p(\omega)\big] \geq \sum_{\omega \in A} \big[q(\omega) - c^* p(\omega)\big] \\
\Rightarrow \quad & \mathbf{Q}\big(\underbrace{\phi^* = \mathbf{Q}}_{=A^*}\big) - c^* \, \mathbf{P}\left(\phi^* = \mathbf{Q}\right) \geq \mathbf{Q}\big(\underbrace{\phi = \mathbf{Q}}_{=A}\big) - c^* \, \mathbf{P}\big(\phi = \mathbf{Q}\big) \\
\Rightarrow \quad & \mathbf{Q}\big(\phi^* = \mathbf{Q}\big) - \mathbf{Q}\big(\phi = \mathbf{Q}\big) \geq c^* \left\{ \underbrace{\mathbf{P}\big(\phi^* = \mathbf{Q}\big)}_{=\alpha} - \underbrace{\mathbf{P}\big(\phi = \mathbf{Q}\big)}_{\leq \alpha} \right\} \geq 0 \\
\Rightarrow \quad & \mathbf{Q}\big(\phi = \mathbf{P}\big) = 1 - \mathbf{Q}\big(\phi = \mathbf{Q}\big) \geq 1 - \mathbf{Q}\big(\phi^* = \mathbf{Q}\big) = \mathbf{Q}\big(\phi^* = \mathbf{P}\big).
\end{aligned}$$

□

Beispiel 4.3 ($\mathcal{B}(n,p)$-Verteilungen/Fortsetzung von Beispiel 4.1) Wir betrachten ϕ^* für den Medikamententest

$$p(k) = \binom{n}{k} p^k (1-p)^{n-k}, \quad p = 0,6;$$

$$q(k) = \binom{n}{k} q^k (1-q)^{n-k}, \quad q = 0,7;$$

$$\Rightarrow L(k) = \frac{q(k)}{p(k)} = \left(\frac{q}{p}\right)^k \left(\frac{1-q}{1-p}\right)^{n-k} = \left(\frac{q}{1-q} \Big/ \frac{p}{1-p}\right)^k \left(\frac{1-q}{1-p}\right)^n \geq c^*.$$

Wegen $q > p$ folgt $\frac{q}{1-q} \Big/ \frac{p}{1-p} > 1$ und damit

$$L(k) \geq c^* \quad \Leftrightarrow \quad k \geq k^*.$$

Wähle nun k^* so, dass

$$\mathbf{P}\big(\phi^* = \mathbf{Q}\big) = \mathbf{P}\big(L(k) \geq c^*\big) = \mathbf{P}(k \geq k^*) = \alpha.$$

Zahlenbeispiel: $n = 100,\ p = 0.6,\ \alpha \approx 0,01$ (nicht jedes α ist wählbar!). Bestimmung von k^*:

$$\mathbf{P}(k \geq k^*) = \mathbf{P}\big(\{k^*, \ldots, n\}\big) = \sum_{k=k^*}^{n} p(k) = \sum_{k=k^*}^{n} \binom{n}{k} p^k (1-p)^{n-k} \stackrel{!}{=} \alpha.$$

Mit einer Approximation der Binomialverteilung (s. Anhang 4.1) erhalten wir $k^* = 72$ bei $\alpha = 0,0095$.

[Diskussion! - insbesondere der Tatsache, dass $k^* > 70 = nq$.]

[Die Grenze k^* ist unabhängig von q, d.h. wir erhalten das gleiche k^* für alle q. Daher heißt ϕ^* gleichmäßig bester Test von $H_0 = \{\mathcal{B}(n,p)\}$ gegen $H_A = \{\mathcal{B}(n,q) \,|\, q > p\}$.]

Fehler 1. Art:

$$\mathbf{P}\big(\phi^* = \mathbf{Q}\big) = \mathbf{P}(k \geq k^*) = 0.0095.$$

Fehler 2. Art: Den Fehler 2. Art kann man ebenfalls näherungsweise mit einer Approximation der Binomialverteilung herleiten (s. Anhang 4.1). Man erhält

$$\mathbf{Q}(\phi^* = \mathbf{P}) = \mathbf{Q}(k < k^*) = \sum_{k=0}^{k^*-1} \binom{n}{k} q^k (1-q)^{n-k} \approx 0.63.$$

Damit ist der Fehler 2. Art sehr groß. Da der Test optimal ist, kann man diesen Fehler aber nicht verkleinern. Ausweg: vergrößere n. Man erhält analog

n = 200:

Herleitung analog: $\mathbf{P}\big(\phi^* = \mathbf{Q}\big) = \mathbf{P}(k \geq k^*) = \sum_{k=k^*}^{200} \binom{200}{k} p^k (1-p)^{200-k} \approx 0.01$

$\Rightarrow \quad k^* = 137 \quad (\frac{137}{200} = 0.685\,!)$

$\Rightarrow \quad \mathbf{Q}\big(\phi^* = \mathbf{P}\big) = \mathbf{Q}(k < k^*) \approx 0.29$ (s. Anhang 4.1).

n = 300:

analog: $\quad k^* = 201 \quad (\frac{201}{300} = 0.67\,!) \quad \Rightarrow \quad \mathbf{Q}\big(\phi^* = \mathbf{P}\big) \approx 0.12.$

"Design of Experiments": Man gibt auch den Fehler 2. Art vor $\big($z.B. $\mathbf{Q}(\phi^* = \mathbf{P}) \leq 0.01\big)$ und bestimmt das kleinste n, für das der optimale Test diese Fehlergrenzen einhält.

Gesucht sind also k^* und n mit

$$\sum_{k=k^*}^{n} \binom{n}{k} p^k (1-p)^{n-k} \approx 0.01, \quad p = 0.6;$$

$$\sum_{k=0}^{k^*-1} \binom{n}{k} q^k (1-q)^{n-k} \approx 0.01, \quad q = 0.7.$$

Man erhält (s. Anhang 4.1) $n = 482$ und $k^* = ?$. □

Bemerkung 4.4 (i) Anstelle von Testfunktionen

$$\phi : \Omega \to \{\mathbf{P}, \mathbf{Q}\}$$

betrachtet man normalerweise Testfunktionen

$$\phi : \Omega \to \{0, 1\},$$

wobei $\phi = 0$ eine Entscheidung für H_0 und $\phi = 1$ eine Entscheidung für H_A bedeutet (in obigem Fall war $H_0 = \{\mathbf{P}\}$ und $H_A = \{\mathbf{Q}\}$).

(ii) Man beachte, dass im obigen Satz das Niveau α nicht frei wählbar ist. Die Aussage des Satzes gilt nur für solche α, für die es ein c^* gibt mit $\mathbf{P}(L(\omega) \geq c^*) = \alpha$.

[Für sog. stetige Verteilungen gilt ein ähnliches Resultat (Satz 6.18), wo dieses Problem nicht auftaucht. Um dieses Problem auch bei diskreten Verteilungen zu überwinden, werden sog. "randomisierte" Tests verwendet (die dann in einer höheren Statistik-Vorlesung behandelt werden).]

Zusammenfassung 4.5 (Optimaler Test)

1. Wir möchten die einfachen Hypothesen $H_0 = \{\mathbf{P}\}$ gegen $H_A = \{\mathbf{Q}\}$ aufgrund einer Beobachtung (bzw. einer Statistik - hier: Anzahl der Erfolge) testen.

2. Das Niveau α für den Fehler 1. Art $\mathbf{P}(\phi = \mathbf{Q})$ wird vorgeben, z.B. $\alpha = 0,01$.

3. Der optimale Test $\big($d.h. $\mathbf{Q}\big(\phi = \mathbf{P}\big)$ minimal$\big)$ ergibt sich aus dem Neyman-Pearson-Lemma:
$$\phi^*(\omega) = \begin{cases} \mathbf{P}, & L(\omega) := \frac{q(\omega)}{p(\omega)} < c^* \\ \mathbf{Q}, & L(\omega) \geq c^*. \end{cases}$$

4. c^* wird aus der Bedingung $\mathbf{P}\big(\phi^* = \mathbf{Q}\big) = \mathbf{P}\big(L(\omega) \geq c^*\big) = \alpha$ bestimmt (Fehler 1. Art).

5. Evtl. wird das Niveau α nicht exakt angenommen. Man erhält dann nur einen optimalen Test zu einem Niveau $\alpha' \approx \alpha$.

6. Ist $L(\omega)$ monoton in ω, so kann man den Test vereinfachen:

$$\phi^*(\omega) = \begin{cases} \mathbf{P} & \omega < \omega^* \\ \mathbf{Q}, & \omega \geq \omega^* \end{cases},$$

   wobei ω^* bestimmt wird durch $\mathbf{P}(\omega \geq \omega^*) = \alpha$. Da ω^* dann von $\mathbf{Q}$ nicht abhängt, erhält man idR einen gleichmäßig besten Test gegen eine größere Gegenhypothese, z.B.

$$H_0 = \{\mathcal{B}(n,p) \,|\, p = p_0\} \quad \text{gegen} \quad H_A = \{\mathcal{B}(n,p) \,|\, p > p_0\}.$$

7. Man beachte die Unsymmetrie zwischen den Fehlern 1. und 2. Art.

Beispiel 4.6 ($\mathcal{H}(N,M,n)$ - Verteilung)
Teste $H_0 = \{N = N_0\}$ gegen $H_A = \{N = N_A\}$ mit $N_0 < N_A$ zum Niveau α (vorgegeben). Sei

$$h(N,M,n,k) := p(k) = \frac{\binom{M}{k}\binom{N-M}{n-k}}{\binom{N}{n}}.$$

[Beispiel: Fische im See wie in Beispiel 1.8: M, n sind bekannt, k (Ergebnis des Experiments) ist ebenfalls bekannt, N ist unbekannt]
Satz 4.2 liefert die optimale Entscheidungsfunktion:

$$\phi^*(k) = \begin{cases} N_0, & L(k) < c^* \\ N_A, & L(k) \geq c^* \end{cases} \quad \text{mit} \quad L(k) = \frac{h(N_A,M,n,k)}{h(N_0,M,n,k)},$$

wobei c^* bestimmt wird durch $\mathbf{P}(\phi^* = N_A) = \mathbf{P}(L(k) \geq c^*) = \alpha$.

Behauptung: $L(k)$ ist monoton in k.
Beweis:

$$\frac{h(N+1,M,n,k)}{h(N,M,n,k)} = \frac{\binom{M}{k}\binom{N+1-M}{n-k}}{\binom{N+1}{n}} \Big/ \frac{\binom{M}{k}\binom{N-M}{n-k}}{\binom{N}{n}}$$
$$= \frac{(N+1-M)(N-n+1)}{(N-M-n+k+1)(N+1)}$$

$$\Rightarrow \quad L(k) = \prod_{j=0}^{N_A-N_0-1} \frac{h(N_0+j+1,M,n,k)}{h(N_0+j,M,n,k)} \quad \text{monoton fallend in } k.$$

Fazit:
(i) Der optimale Test ist

$$\phi^*(k) = \begin{cases} N_0, & k > k^* \\ N_A, & k \le k^* \end{cases}$$

mit $\mathbf{P}(k \le k^*) \approx \alpha$;
(ii) ϕ^* ist auch gleichmäßig bester Test für $H_0 = \{N_0\}$ gegen $H_A = \{N \mid N > N_0\}$. $\square$

4.1 Anhang: Fehler 1. und 2. Art beim Binomialtest

Zum Verständnis dieses Anhangs sind Kenntnisse über Zufallsvariable und den zentralen Grenzwertsatz notwendig, d.h. insbesondere aus Kapitel 5, 6, 8 und 14.

Wir wollen nun die konkreten Werte aus Beispiel 4.3 (insbesondere die Fehler 1. und 2. Art sowie k^*) mit einer Approximation der Binomialverteilung herleiten. Da p nicht klein ist verwenden wir anstelle der Poissonapproximation die in Kapitel 14 behandelte Approximation durch die Normalverteilung basierend auf dem zentralen Grenzwertsatz. Um die Notation zu vereinfachen ersetzen wir die bisherigen Größen $p\,(=0.6)$ und $q\,(=0.7)$ durch p_0 bzw. p_A. Die Fehler 1. und 2. Art sind

$$\mathbf{P}\big(\phi^* = \mathbf{Q}\big) = \mathbf{P}(k \ge k^*) = \sum_{k=k^*}^{n} \binom{n}{k} p_0^k (1-p_0)^{n-k}$$

bzw.

$$\mathbf{Q}(\phi^* = \mathbf{P}) = \mathbf{Q}(k < k^*) = \sum_{k=0}^{k^*-1} \binom{n}{k} p_A^k (1-p_A)^{n-k}.$$

Sei

$$X_i := \begin{cases} 1, & \text{Person i geheilt,} \\ 0, & \text{Person i nicht geheilt.} \end{cases}$$

Dann gilt für die Anzahl der Heilerfolge $X = \sum_{i=1}^n X_i$ und es folgt mit dem zentralen Grenzwertsatz in Satz 14.4

$$\frac{X - np}{\sqrt{np(1-p)}} \;=\; \sqrt{n}\,\frac{\frac{1}{n}\sum_{i=1}^n X_i - p}{\sqrt{p(1-p)}} \;\xrightarrow{\mathcal{D}}\; Z$$

mit $Z \sim \mathcal{N}(0,1)$. Die Verteilungen $\mathbf{P}$ und $\mathbf{Q}$ sind nun die Verteilungen $\mathbf{P}_{p_0}^X$ mit $p_0 = 0.6$ bzw. $\mathbf{P}_{p_A}^X$ mit $p_A = 0.7$. Damit erhält man als Approximation für den Fehler 1. Art

$$\begin{aligned}
\mathbf{P}\big(\phi^* = \mathbf{Q}\big) = \mathbf{P}(k \geq k^*) \;&=\; \mathbf{P}_{p_0}\Big(X > k^* - \frac{1}{2}\Big) \\
&=\; \mathbf{P}_{p_0}\left(\frac{X - np}{\sqrt{np(1-p)}} > \frac{k^* - \frac{1}{2} - np}{\sqrt{np(1-p)}}\right) \quad \text{mit } p = p_0 \\
&\approx\; \mathbf{P}\left(Z > \frac{k^* - \frac{1}{2} - np}{\sqrt{np(1-p)}}\right) \quad \text{mit } p = p_0 \\
&=\; 1 - \Phi\left(\frac{k^* - 1/2 - np_0}{\sqrt{np_0(1-p_0)}}\right) \\
&=\; 1 - \Phi\left(\frac{k^* - 60.5}{4.90}\right) \quad \text{für } p_0 = 0.6,\ n = 100
\end{aligned}$$

sowie analog für den Fehler 2. Art

$$\mathbf{Q}(\phi^* = \mathbf{P}) \;=\; \mathbf{Q}(k < k^*) \;=\; \mathbf{P}_{p_A}\big(X \leq k^* - \frac{1}{2}\big)$$

$$\approx\; \Phi\left(\frac{k^* - 1/2 - np_A}{\sqrt{np_A(1-p_A)}}\right).$$

$$=\; \Phi\left(\frac{k^* - 70.5}{4.58}\right) \quad \text{für } p_A = 0.7,\ n = 100.$$

Mit diesen Ausdrücken und einer Tafel für Φ lassen sich die konkreten Werte aus Beispiel 4.3 berechnen:

Bestimmung von k^* und Fehler 1. Art:

$$\mathbf{P}(k \geq k^*) \approx 1 - \Phi\left(\frac{k^* - 60.5}{4.90}\right) \quad \stackrel{!}{=} \quad \alpha = 0.01$$

$$\Leftrightarrow\; k^* \;=\; 60.5 + 4.9 * \underbrace{\Phi^{-1}\big(1 - 0.01\big)}_{=2.33} \;=\; 71.9$$

Wir setzen $k^* = 72$ und erhalten für den Fehler 1. Art exakt (abgesehen von der Normal-Approximation)

$$\mathbf{P}(k \geq k^*) = 1 - \Phi\left(\frac{72 - 60.5}{4.90}\right) \;=\; 1 - \underbrace{\Phi(0.918)}_{0.9905} \;=\; 0.0095 \;<\; 0.01$$

Der zugehörige Test ist damit optimal zum Niveau $\alpha = 0.0095$.

Fehler 2. Art: Mit $k^* = 72$ erhalten wir

$$\mathbf{Q}(k < k^*) \;\approx\; \Phi\left(\frac{72 - 70.5}{4.58}\right) \;=\; \Phi(0.33) \;=\; 0.63.$$

Werte für $n = 200$: Die Bestimmung von k^* ergibt

$$\mathbf{P}(k \geq k^*) \approx 1 - \Phi\left(\frac{k^* - 0.5 - 120}{\sqrt{48}}\right) \quad \stackrel{!}{=} \quad 0.01$$

$$\Leftrightarrow\; k^* \;=\; 120.5 + 6.93 * \underbrace{\Phi^{-1}\big(1 - 0.01\big)}_{=2.33} \;=\; 136.6$$

d.h. wir setzen $k^* = 137$. Fehler 2. Art:

$$\mathbf{Q}(k < k^*) \;\approx\; \Phi\left(\frac{136.5 - 140}{\sqrt{42}}\right) \;=\; \Phi(-0.54) \;=\; 0.29.$$

Werte für $n = 300$: Die Bestimmung von k^* ergibt

$$\mathbf{P}(k \geq k^*) \approx 1 - \Phi\left(\frac{k^* - 0.5 - 180}{\sqrt{72}}\right) \quad \stackrel{!}{=} \quad 0.01$$

$$\Leftrightarrow \; k^* \;=\; 180.5 + 8.49 * \underbrace{\Phi^{-1}\big(1 - 0.01\big)}_{=2.33} \;=\; 200.3$$

d.h. wir setzen $k^* = 201$ (wir runden k^* auf, damit das Niveau eingehalten wird!). Fehler 2. Art:

$$\mathbf{Q}(k < k^*) \;\approx\; \Phi\left(\frac{200.5 - 210}{\sqrt{63}}\right) \;=\; \Phi(-1.20) \;=\; 0.12.$$

Design of Experiments:

Schreibt man die oben verwendeten Ausdrücke für die Berechnung von k^* in Abhängigkeit von n, so erhält man

$$k^* \;=\; 0.6n + 0.5 + 2.33\sqrt{0.24n}.$$

Der Fehler 2. Art beträgt damit

$$\mathbf{Q}(k < k^*) \;\approx\; \Phi\left(\frac{0.6n + 2.33\sqrt{0.24n} - 0.7n}{\sqrt{0.21n}}\right) \;=\; \Phi\big(-0.22\sqrt{n} + 2.5\big).$$

Damit gilt

$$\mathbf{Q}(k < k^*) \quad \stackrel{!}{=} \quad 0.01$$

$$\Leftrightarrow \; -0.22\sqrt{n} + 2.5 \;=\; \Phi^{-1}\big(0.01\big) = -2.33$$

$$\Leftrightarrow \; n = 482$$

Bestimmung der Einzelwerte $\mathbf{P}(\{66\})$ und $\mathbf{Q}(\{66\})$:

Mit der obigen Normalapproximation erhält man

$$\begin{aligned}\mathbf{P}(\{66\}) &= \mathbf{P}_{p_0}\big(X \le 66.5\big) - \mathbf{P}_{p_0}\big(X \le 65.5\big) \\ &\approx \Phi\left(\frac{66.5-60}{\sqrt{24}}\right) - \Phi\left(\frac{65.5-60}{\sqrt{24}}\right) = 0.91 - 0.87 = 0.04\end{aligned}$$

und

$$\begin{aligned}\mathbf{Q}(\{66\}) &= \mathbf{P}_{p_A}\big(X \le 66.5\big) - \mathbf{P}_{p_A}\big(X \le 65.5\big) \\ &\approx \Phi\left(\frac{66.5-70}{\sqrt{21}}\right) - \Phi\left(\frac{65.5-70}{\sqrt{21}}\right) = 0.22 - 0.16 = 0.06.\end{aligned}$$

5 Diskrete Zufallsvariable

In diesem Kapitel werden Zufallsvariable als Abbildungen von Ω in einen Bildraum definiert. Es wird gezeigt, wie Zufallsvariable eine Wahrscheinlichkeitsverteilung auf dem Bildraum induzieren. In einem Beispiel wird die Prognose von Zufallsvariablen behandelt.

[diskret: Ω abzählbar, genauer: $\Omega^X =$ Bild(X) abzählbar]

Beispiel 5.1 Dreimaliges Werfen einer "fairen" Münze

$$\Omega = \{(\omega_1, \omega_2, \omega_3) \mid \omega_j \in \{ \underset{\substack{\uparrow \\ \text{Kopf}}}{0} , \underset{\substack{\uparrow \\ \text{Zahl}}}{1} \}\},$$

$$\mathbf{P} \text{ Laplace Verteilung auf } \Omega \text{ d.h. } \mathbf{P}(A) = \frac{|A|}{|\Omega|} = \frac{|A|}{8}.$$

Frage: Wie groß ist die Wahrscheinlichkeit, dass k-mal "Zahl" erscheint?
2 Möglichkeiten:

(i) $A_k = \{(\omega_1, \omega_2, \omega_3) \mid \Sigma w_j = k\}, \quad \mathbf{P}(A_k) = \frac{|A_k|}{|\Omega|}$.

(ii) Kapitel 3: Anzahl "Zahl" ist $\mathcal{B}(3, \frac{1}{2})$-verteilt. Wir haben hier ein anderes Ω und ein anderes $\mathbf{P}$, nämlich

$$\Omega^X = \{0, 1, 2, 3\} \quad \text{und} \quad \mathbf{P}^X(\{k\}) = \binom{3}{k} (\frac{1}{2})^k (\frac{1}{2})^{3-k}.$$

Wir können $(\Omega^X, \mathbf{P}^X)$ aus $(\Omega, \mathbf{P})$ erhalten mittels der Abbildung

$$\begin{aligned} X: \quad & \Omega \to \mathbb{R} \\ & \omega = (\omega_1, \omega_2, \omega_3) \mapsto \sum_{j=1}^{3} \omega_j \ , \end{aligned}$$

$$\Omega^X = \text{Bild } X = \{x \in \mathbb{R} \mid \exists\, \omega \in \Omega :\ x = X(\omega)\},$$
$$\mathbf{P}^X = \text{W'verteilung auf } \Omega^X \text{ mit}$$
$$\mathbf{P}^X(B) = \mathbf{P}(\underbrace{X^{-1}(B)}_{\substack{= \text{Urbild von B} \\ = \{\omega \in \Omega \mid X(\omega) \in B\}}}), \qquad B \in \mathcal{P}(\Omega^X),$$

z.B.

$$\begin{aligned}\mathbf{P}^X(\{1\}) &= \mathbf{P}(\{\omega \in \Omega | X(\omega) = 1\}) \\ &= \mathbf{P}(\{(1,0,0),(0,1,0),(0,0,1)\}) = \mathbf{P}(A_1) = \frac{3}{8}.\end{aligned}$$

Es gilt $\mathbf{P}^X = \mathcal{B}(3, \frac{1}{2})$.

X heißt Zufallsvariable, P^X die von X induzierte Verteilung . □

Definition 5.2 Sei Ω abzählbar mit Wahrscheinlichkeitsverteilung $\mathbf{P}$. Eine Abbildung $X : \Omega \to \mathbb{R}$ heißt diskrete Zufallsvariable (ZVA). Die durch X induzierte Verteilung $\mathbf{P}^X(A) := \mathbf{P}(X^{-1}(A))$ ist eine Wahrscheinlichkeitsverteilung auf $\Omega^X = \text{Bild } X$ und $\mathcal{P}(\Omega^X)$. Die Funktion $F_X(x) = \mathbf{P}(\{\omega \mid X(\omega) \leq x\}) = \mathbf{P}(X \leq x)$ heißt Verteilungsfunktion von X. $\mathcal{T}(X) = \{x \in \mathbb{R} \mid \mathbf{P}^X(\{x\}) > 0\}$ heißt Träger von $\mathbf{P}^X$.

[Bem.: X ist keine Variable, sondern eine Funktion.]

Bemerkung 5.3

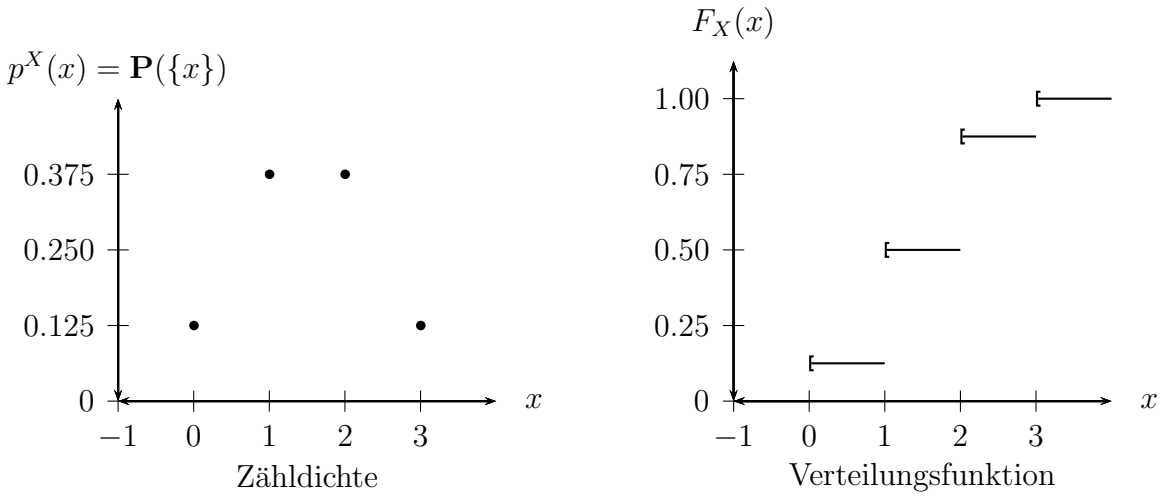
(i) Man rechnet leicht nach, dass $\mathbf{P}^X$ wirklich eine Wahrscheinlichkeitsverteilung ist.

(ii) Man schreibt $X \sim \mathbf{P}^X$ (insbesondere bei konkreten Verteilungen), z.B. $X \sim \mathcal{B}(n,p)$.

(iii) X diskrete ZVA bedeutet genauer, dass Ω^X abzählbar ist [Ω darf überabzählbar sein. Man braucht dann aber den zusätzlichen Begriff der Messbarkeit von $X \to$ Statistik I].

(iv) Man rechnet leicht nach, dass F_X monoton wachsend und rechtsseitig stetig mit $\lim_{x \to -\infty} F_X(x) = 0$ und $\lim_{x \to \infty} F_X(x) = 1$ ist. Man kann die Verteilung $\mathbf{P}^X$ von X

auf zwei Arten graphisch veranschaulichen - mit der bereits bekannten Zähldichte p^X oder der Verteilungsfunktion $F_X(x) = \mathbf{P}(X \leq x)$. Für obiges Beispiel erhält man



Beispiel 5.4 Zweimaliges Würfeln. Wie groß ist die Wahrscheinlichkeit, eine "8" als Augensumme zu erhalten?

$$\Omega = \{(i,j) \,|\, i,j \in \{1,\ldots,6\}\}, \quad \mathbf{P} \text{ Laplace-Verteilung}$$

$$\begin{aligned} X : \Omega &\to \mathbb{R} \\ (i,j) &\mapsto i+j \quad \text{(Augensumme)}. \end{aligned}$$

Dann ist $\Omega^X = \{2,\ldots,12\}$. Für $\mathbf{P}^X$ gilt z.B.

$$\begin{aligned} \mathbf{P}^X\big(\{8\}\big) &= \mathbf{P}\big(X^{-1}(\{8\})\big) \\ &= \mathbf{P}\big(\{\omega \in \Omega \,|\, X(\omega) = 8\}\big) \\ &= \mathbf{P}\big(\{(2,6),(3,5),(4,4),(5,3),(6,2)\}\big) \\ &= \frac{5}{|\Omega|} = \frac{5}{36}. \end{aligned}$$

□

Bemerkung 5.5 Man beachte, dass für die Beantwortung der Frage nach der "W't, eine "8" zu erhalten", nur der Raum Ω^X und die Verteilung $\mathbf{P}^X$ von Interesse sind. Man verzichtet deswegen häufig auf die Angabe von Ω und $\mathbf{P}$, z.B.

$$
\begin{aligned}
&X = \text{Anzahl der Kunden in einem Geschäft innerhalb der nächsten Stunde,}\\
&\mathbf{P}^X = \mathcal{P}(\lambda), \quad \text{(z.B. mit } \lambda = 20)\\
&\Omega^X = \mathbb{N}_0.
\end{aligned}
$$

Hier werden Ω und $\mathbf{P}$ nicht angegeben.

Beispiel 5.6 Es werde n mal gewürfelt. X sei die Anzahl der Würfe mit Augenzahl 6.

$$
\begin{aligned}
&\Omega = \{(\omega_1, \ldots, \omega_n) \mid \omega_i \in \{1, \ldots 6\}\},\\
&\mathbf{P}(A) = \frac{|A|}{|\Omega|} \quad \text{Laplaceverteilung,}\\
&X : \Omega \to \mathbb{R} \quad \text{mit} \quad X(\omega) = \sum_{i=1}^{n} \mathrm{I}_{\{\omega_i = 6\}}(\omega),\\
&\mathbf{P}^X\big(\{k\}\big) = \mathbf{P}\big(X = k\big) = \mathbf{P}\big(\{\omega \mid \sum_{i=1}^{n} \mathrm{I}_{\{\omega_i = 6\}}(\omega) = k\}\big) = \binom{n}{k} \big(\frac{1}{6}\big)^k \big(\frac{5}{6}\big)^{n-k},\\
&\Omega^X = \{0, \ldots, n\},\\
&\mathbf{P}^X \text{ ist die } \mathcal{B}(n, \frac{1}{6})\text{-Verteilung auf } \Omega^X, \text{ formal } X \sim \mathcal{B}(n, \frac{1}{6}).
\end{aligned}
$$

Wiederum ist die Angabe von Ω und $\mathbf{P}$ nicht erforderlich (um die Wahrscheinlichkeit zu berechnen, k mal eine "6" zu würfeln). □

Beispiel 5.7 (Vorhersage einer Zufallsvariable) Seien $X_1, \ldots, X_n, X_{n+1}$ identisch verteilte Zufallsvariable mit

$$X_i \in \{0, 1\}, \quad \mathbf{P}(X_i = 1) = p \quad \forall i \in \{1, \ldots, n+1\}.$$

[man braucht außerdem eine Unabhängigkeitsannahme - s. unten]
Was ist eine gute Vorhersage (Prädiktion) von X_{n+1}, wenn man die Werte von $X_1, \dots X_n$ kennt? Gütekriterium:

$$\boxed{\mathbf{P}_p\big(V(X_1,\dots,X_n) = X_{n+1}\big) = \max} \quad (*)$$

$\beta_V(p) := \mathbf{P}_p(V(X_1,\dots,X_n) = X_{n+1})$ heißt Gütefunktion.

[$\mathbf{P}$ hängt von p ab; V ist als Funktion von $X_1,\dots,X_n$ auch eine ZVA]

Beispiel

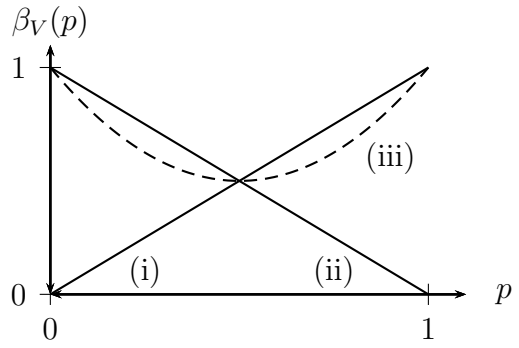
(i) $V \equiv 1 \quad \beta_V(p) = \mathbf{P}(X_{n+1} = 1) = p$
(ii) $V \equiv 0 \quad \beta_V(p) = \mathbf{P}(X_{n+1} = 0) = 1 - p$
(iii) $V = X_n$

$$\begin{aligned} \beta_V(p) &= \mathbf{P}(X_n = X_{n+1}) = \mathbf{P}(X_n = 0, X_{n+1} = 0) + \mathbf{P}(X_n = 1, X_{n+1} = 1) \\ &\overset{\text{unabh.}}{=} (1-p)^2 + p^2 = 2p^2 - 2p + 1 = 2\left(p - \frac{1}{2}\right)^2 + \frac{1}{2} \end{aligned}$$

(iv) Idee: Wähle

$$V = \begin{cases} 1 & \Sigma X_i > n/2 \\ ? & \Sigma X_i = n/2 \\ 0 & \Sigma X_i < n/2 \end{cases}$$

Zur Vereinfachung sei n ungerade.



Intuitiv ist es klar, dass man eine Zusatzbedingung braucht, um (*) glm. zu maximieren.

Zusatzannahme: V ist symmetrisch, d.h.

$$V(1-X_1,\ldots,1-X_n) = 1 - V(X_1,\ldots,X_n) \quad \text{plausibel!}$$

(a) Beh.: $\beta_V(p) = p + (1-2p)\,\mathbf{P}_p(V=0)$

Bew.:

$$\begin{aligned}
\beta_V(p) &= \mathbf{P}_p(V=0, X_{n+1}=0) + \mathbf{P}_p(V=1, X_{n+1}=1) \\
&\overset{\text{unabh.}}{=} \mathbf{P}_p(V=0)\,\mathbf{P}_p(X_{n+1}=0) + \mathbf{P}_p(V=1)\,\mathbf{P}_p(X_{n+1}=1) \\
&= (1-p)\,\mathbf{P}_p(V=0) + p\,[1-\mathbf{P}_p(V=0)] \\
&= p + (1-2p)\,\mathbf{P}_p(V=0)
\end{aligned}$$

(b) Bem.: (a) $\Rightarrow$ $\beta_V(\frac{1}{2}) = \frac{1}{2} \quad \forall\, V$ plausibel wegen Unabhängigkeit!

(c) Beh.: V symmetrisch $\Rightarrow$ $\mathbf{P}_{\frac{1}{2}}(V=0) = \frac{1}{2}$ plausibel!

Bew.: Es gilt

$$\mathbf{P}_p\big(X_i=1\big) = p = \mathbf{P}_{1-p}\big(1-X_i=1\big),$$

d.h. die $1-X_i$ haben bei Vorliegen des Parameters $1-p$ die gleiche Verteilung wie die X_i bei Vorliegen des Parameters p. [Genauer braucht man dieses Argument für

den Zufallsvektor $(1-X_1,\ldots,1-X_n)'$, wo es aber ebenfalls richtig ist]. Daraus folgt

$$\begin{aligned}\mathbf{P}_p\big(V(X_1,\ldots,X_n)=0\big) &= \mathbf{P}_{1-p}\big(V(1-X_1,\ldots,1-X_n)=0\big)\\ &= \mathbf{P}_{1-p}\big(V(X_1,\ldots X_n\big)=1)\\ &= 1-\mathbf{P}_{1-p}\big(V(X_1,\ldots,X_n\big)=0)\end{aligned}$$

$$\Rightarrow\ \mathbf{P}_{\frac{1}{2}}(V=0)=1-\mathbf{P}_{\frac{1}{2}}(V=0)$$

$$\Rightarrow\ \text{Beh.}$$

(d) $\boxed{p<\frac{1}{2}}$ Aus (a) folgt: $\beta_V(p)$ maximal $\Leftrightarrow$ $\mathbf{P}_p(V=0)$ maximal.
Zusätzliche Nebenbedingung: $\mathbf{P}_{\frac{1}{2}}(V=0)=\frac{1}{2}$ (gilt, falls V symmetrisch).
Sei $\Omega=\big\{(\omega_1,\ldots,\omega_n)\mid\omega_i\in\{0,1\}\big\}$ und $A\subset\Omega$ mit $A:=\{\omega\in\Omega\mid V(\omega)=0\}$.
Also maximiere $\mathbf{P}_p(A)$ unter $\mathbf{P}_{\frac{1}{2}}(A)=\frac{1}{2}$

$$\begin{aligned}\mathbf{P}_p(A) &= \sum_{\omega\in A}\mathbf{P}_p(\{w\})\\ &\overset{\text{unabh.}}{=} \sum_{\omega\in A}p^{\Sigma\omega_j}\,(1-p)^{n-\Sigma w_j}\\ &= \sum_{\omega\in A}\Big(\frac{p}{1-p}\Big)^{\Sigma\omega_j}\,(1-p)^n \quad (*)\end{aligned}$$

$\mathbf{P}_{\frac{1}{2}}(A)=\frac{1}{2}\ \Rightarrow\ \sum_{\omega\in A}(\frac{1}{2})^n=\frac{1}{2}\ \Rightarrow\ |A|=2^{n-1}$.

$p<\frac{1}{2}\ \Rightarrow\ \left(\frac{p}{1-p}\right)<1$, d.h. (*) wird maximiert, wenn A diejenigen 2^{n-1} ω enthält, für die Σw_j am kleinsten ist, d.h. diejenigen 2^{n-1} ω, für die $\Sigma w_j<\frac{n}{2}$ gilt. Also:

$$V_{opt}=0\ \Leftrightarrow\ \sum_{j=1}^{n}X_j<\frac{n}{2}.$$

(e) $\boxed{p>\frac{1}{2}}$ analog:
Minimiere $\mathbf{P}_p(V=0)\overset{\text{s.o.}}{=}1-\mathbf{P}_{1-p}(V=0)$ unter $\mathbf{P}_{\frac{1}{2}}(V=0)=\frac{1}{2}$. Dies liefert wie unter (d):

$$V_{opt}=0\ \Leftrightarrow\ \sum_{j=1}^{n}X_j<\frac{n}{2}.$$

Ergebnis: optimale Prognose

$$V_{opt}(X_1,\ldots,X_n) = \begin{cases} 0 & \text{falls} \quad \sum_{j=1}^n X_j < \frac{n}{2} \\ 1 & \text{falls} \quad \sum_{j=1}^n X_j > \frac{n}{2} \end{cases} .$$

[n gerade: randomisieren]

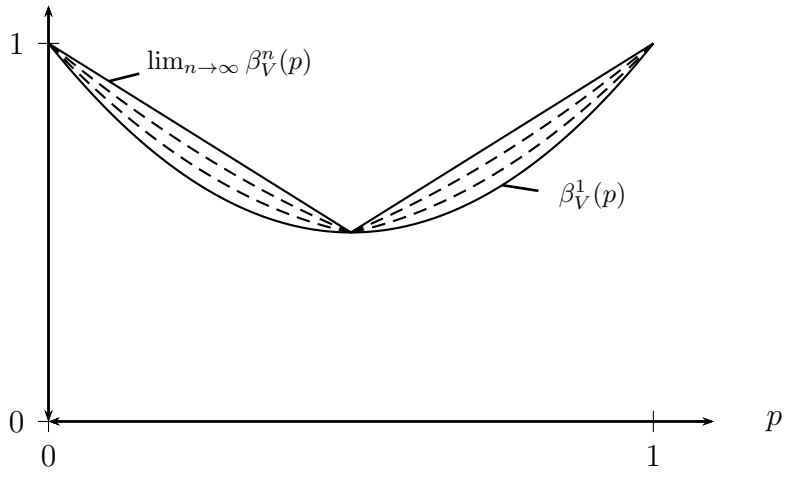
(f) Gütefunktion:

$$\begin{aligned} \beta_V(p) &= p + (1-2p)\,\mathbf{P}_p(V=0) \\ &= p + (1-2p) \sum_{k=0}^{[n/2]} \binom{n}{k} p^k (1-p)^{n-k} \\ &=: \beta_V^n(p) \end{aligned}$$

Wir zeigen unten, dass

$$\lim_{n\to\infty} \beta_V^n(p) = |p - \frac{1}{2}| + \frac{1}{2}.$$

Grafisch dargestellt haben wir damit folgende Situation:



[es ist klar, dass $\beta_V^1(p)$ gleich $\beta_V(p)$ aus Bsp.(iii) ist]

Bemerkungen:

(i) Der Limes ist die Gütefunktion der besten Vorhersage bei bekanntem p. Falls man die relative Häufigkeit $\widehat{p} = \frac{1}{n}\sum_{j=1}^{n} X_j$ als Schätzwert für den unbekannten Parameter p verwendet, so erhält man

$$V_{opt}(X_1, \ldots, X_n) = \begin{cases} 0 & \text{falls} \quad \widehat{p} < \frac{1}{2} \\ 1 & \text{falls} \quad \widehat{p} > \frac{1}{2} \end{cases} .$$

Wir werden später sehen, dass $\widehat{p}$ für $n \to \infty$ gegen p konvergiert (in einem noch zu definierenden Sinne). Damit ist heuristisch klar, warum die Gütefunktionen gegeneinander konvergieren.

(ii) An 2 Stellen haben wir die stochastische Unabhängigkeit von Ereignissen vorausgesetzt. Genauer müssen wir die stochastische Unabhängigkeit der ZVAs annehmen (Definition s. Kapitel 8). Es ist klar, dass wir eine solche Unabhängigkeitsannahme brauchen - sonst würde man auch eine andere Prognose verwenden.

Herleitung des Limes:

$p > \frac{1}{2}$:

$$\begin{aligned} & \sum_{k=0}^{[n/2]} \binom{n}{k} p^k \, (1-p)^{n-k} \\ = \quad & (1-p)^n \sum_{k=0}^{[n/2]} \binom{n}{k} \Big(\underbrace{\frac{p}{1-p}}_{>1} \Big)^k \\ \leq \quad & (1-p)^n \left(\frac{p}{1-p} \right)^{n/2} 2^n \\ = \quad & [4\,p\,(1-p)]^{n/2} \quad \overset{n\to\infty}{\longrightarrow} \; 0. \qquad [\text{wegen } 4\,p\,(1-p) < 1] \end{aligned}$$

$p = \frac{1}{2}$:

$$\sum_{k=0}^{[n/2]} \binom{n}{k} \Big(\frac{1}{2}\Big)^n = \Big(\frac{1}{2}\Big)^n \frac{1}{2}\, 2^n = \frac{1}{2}$$

$\underline{p < \frac{1}{2}}$:

$$\begin{aligned}
\sum_{k=0}^{[n/2]} \binom{n}{k} p^k (1-p)^{n-k} &= 1 - \sum_{k=[\frac{n}{2}]+1}^{n} \binom{n}{k} p^k (1-p)^{n-k} \\
&= 1 - \sum_{l=0}^{[n/2]} \underbrace{\binom{n}{n-l}}_{\binom{n}{l}} p^{n-l} (1-p)^l \\
&\xrightarrow{n\to\infty} \quad 1 - 0 = 1.
\end{aligned}$$

Also folgt:

$$\begin{aligned}
\lim_{n\to\infty} \beta_V^n(p) &= p + (1-2p) \begin{cases} 1 & , \quad p < 1/2 \\ 1/2 & , \quad p = 1/2 \\ 0 & , \quad p > 1/2 \end{cases} \\
&= |p - \frac{1}{2}| + \frac{1}{2}.
\end{aligned}$$

□

6 Stetige Verteilungen und stetige Zufallsvariable

Zur sauberen Definition stetiger Verteilungen und stetiger Zufallsvariable benötigt man Kenntnisse der Maßtheorie, die die Studierenden idR beim Besuch dieser Grundvorlesung noch nicht haben. In diesem Skriptum werden die wirklich notwendigen Begriffe und Resultate der Maßtheorie kurz dargestellt, wobei auf Beweise verzichtet wird. Danach werden stetige Verteilungen und stetige Zufallsvariable definiert und die wichtigsten Beispiele für stetige Verteilungen präsentiert - darunter auch die Normalverteilung. Ziel ist es, alle Definitionen und Ergebnisse inhaltlich und von der Notation her genauso darzustellen wie sie auch später (bei Kenntnis der maßtheoretischen Grundlagen) dargestellt werden. Gleichzeitig wird auf die verbleibenden Lücken in den Beweisen hingewiesen. Zum Schluss wird das Testen einfacher Hypothesen und das Neyman-Pearson-Lemma für stetige Verteilungen behandelt.

Bisher: diskrete Verteilungen mit abzählbarem Ω bzw. Ω^X [bei induzierten Verteilungen];
Jetzt: $\Omega \subset \mathbb{R}$ (d.h. in der Regel überabzählbar), bzw. X ZVA mit $\Omega^X \subset \mathbb{R}$;
Beispiele: Länge von Schulkindern, Lebenszeit einer Glühbirne.

Beispiel 6.1 Ein Student erwartet seine Freundin. Sie hat versprochen, um 16^{00} Uhr zu kommen. Er hält hingegen alle Zeitpunkte zwischen 16^{00} Uhr und 17^{00} Uhr für "gleichwahrscheinlich".

[Diskussion: Was bedeutet hier "gleichwahrscheinlich"?]

Approximation: Teile das Intervall $[a, b]$ in n gleiche Teile mit Mittelpunkten $x_1, \dots x_n$ und Wahrscheinlichkeiten $\frac{1}{n}$. Seien $a \le c < d \le b$ und X die Ankunftszeit.

$$\mathbf{P}(c \le X \le d) \approx \sum_{c \le x_i \le d} \frac{1}{n} = \frac{b-a}{n} \sum_{c \le x_i \le d} \frac{1}{b-a} \longrightarrow \int_c^d \frac{1}{b-a}\, dx.$$

Ist z.B. die Wahrscheinlichkeit in der Mitte des Zeitintervalls größer, so werden wir analog

$$\mathbf{P}(c \le X \le d) = \int_c^d f(x)\, dx$$

erhalten, wobei $f(x) \geq 0$ und $\int_a^b f(x)\,dx = 1$ gilt. $f(x)$ heißt Wahrscheinlichkeitsdichte. Wir werden zeigen, dass dadurch allgemein eine Wahrscheinlichkeitsverteilung definiert wird (stetige Verteilungen). □

Für die exakte Definition von stetigen Verteilungen braucht man Ergebnisse aus der Maßtheorie. Wir werden diese kurz ohne Beweis zitieren (mit **MT** kenntlich gemacht).

Satz 6.2 (MT) *Es gibt keine Wahrscheinlichkeitsverteilung auf* $\big([a,b], \mathcal{P}([a,b])\big)$ *mit*

$$\mathbf{P}\big([c,d]\big) = \int_c^d \frac{1}{b-a}\,dx \qquad \text{für alle } a \leq c < d \leq b.$$

Konsequenz: Man definiert $\mathbf{P}$ auf einem kleineren Mengensystem $\mathcal{A}$, einer sogenannten σ-Algebra:

Definition 6.3 (MT) *Sei* $\Omega \neq \emptyset$. $\mathcal{A} \subset \mathcal{P}(\Omega)$ *heißt* σ*-Algebra über* Ω*, falls gilt:*

(i) $\Omega \in \mathcal{A}$,
(ii) $A \in \mathcal{A} \Rightarrow A^c \in \mathcal{A}$,
(iii) $A_i \in \mathcal{A}\ \forall i \in \mathbb{N} \Rightarrow \bigcup_{i=1}^{\infty} A_i \in \mathcal{A}$.

Bemerkung: Wegen der de-Morgan'schen Regel liegen damit auch alle abzählbaren Durchschnitte wieder in $\mathcal{A}$.

Beispiel 6.4 (i) $\mathcal{P}(\Omega)$ ist eine σ-Algebra.
(ii) $\mathcal{A} = \{\emptyset, A, A^c, \Omega\}$ ist eine σ-Algebra,

Proposition/Definition 6.5 (MT) Sei $\mathcal{E}$ eine Menge von Teilmengen von Ω (in unserem Beispiel $\Omega = [a,b]$, $\mathcal{E} = \{[c,d] \mid a \leq c < d \leq b\}$). Dann gibt es eine eindeutig bestimmte kleinste σ-Algebra $A(\mathcal{E})$ über Ω, die $\mathcal{E}$ enthält. Es gilt

$$A(\mathcal{E}) = \bigcap_{\mathcal{A} \in\, \mathcal{S}(\mathcal{E})} \mathcal{A}\,,$$

wobei $\mathcal{S}(\mathcal{E}) = \{\mathcal{A} \mid \mathcal{A}\ \sigma\text{-Algebra mit } \mathcal{E} \subset \mathcal{A}\}$. $A(\mathcal{E})$ heißt die von $\mathcal{E}$ erzeugte σ-Algebra.

Beweis. Es gilt $\mathcal{P}(\Omega) \in S(\mathcal{E})$, d.h. $S(\mathcal{E}) \neq \emptyset$. Man muss nur noch zeigen, dass der Durchschnitt von beliebig vielen σ-Algebren wieder eine σ-Algebra ist (Ü-Aufgabe). □

Definition 6.6 (MT) *Die von* $\mathcal{E} = \{[c,d] \mid c,d \in \mathbb{R}\}$ *erzeugte* σ*-Algebra* $\mathcal{B} = \mathcal{B}_{\mathbb{R}} = A(\mathcal{E})$ *heißt Borelsche* σ*-Algebra auf* $\mathbb{R}$*. Analog* $\mathcal{B}_{[a,b]}$*.*

Proposition 6.7 (MT) $\mathcal{B}_{\mathbb{R}}$ *wird auch von folgenden Mengensystemen erzeugt:*

$$\mathcal{E}_1 = \{(a,b] \mid a < b\}, \quad \mathcal{E}_2 = \{(-\infty, a] \mid a \in \mathbb{R}\},$$
$$\mathcal{E}_3 = \{(a,b) \mid a < b\}, \quad \mathcal{E}_4 = \{O \subset \mathbb{R} \mid O \text{ offen}\}.$$

Beweis. teilweise Ü-Aufgabe. □

Bemerkung: Damit enthält $\mathcal{B}_{\mathbb{R}}$ insbesondere alle Intervalle, offenen und abgeschlossenen Mengen sowie deren (unendliche) Vereinigungen und Durchschnitte. Man kann aber zeigen, dass $\mathcal{B}_{\mathbb{R}} \neq \mathcal{P}(\mathbb{R})$.

Satz 6.8 (MT) *Sei* $F : \mathbb{R} \to \mathbb{R}$ *eine rechtsseitig stetige, monoton nicht fallende Funktion mit* $\lim_{x\to-\infty} F(x) = 0$ *und* $\lim_{x\to\infty} F(x) = 1$*. Dann gibt es eine eindeutig bestimmte Wahrscheinlichkeitsverteilung* $\mathbf{P}$ *auf* $(\mathbb{R}, \mathcal{B})$ *mit*

$$\mathbf{P}((a,b]) = F(b) - F(a)$$

Beweis. Maßtheorie (Spezialfall des Maßerweiterungssatzes). □

[mündliche Bemerkung: Verlangt man nur, dass $F : \mathbb{R} \to \mathbb{R}$ eine rechtsseitig stetige, monoton nicht fallende Funktion ist, dann gilt der Satz ebenfalls. Allerdings ist dann $\mathbf{P}$ kein W'verteilung mehr, sondern nur noch Maß. Für $F(x) = x$ erhält man insbesondere das Lebesgue-Maß.]

Wir verwenden jetzt den obigen Satz, um stetige Verteilungen zu definieren:

Proposition/Definition 6.9 (Stetige Verteilungen) Sei $f : \mathbb{R} \to [0, K]$ auf allen Intervallen $[a, b]$ Riemann-integrierbar mit $\int_{-\infty}^{\infty} f(y)\, dy = \lim_{a,b\to\infty} \int_{-a}^{b} f(y)\, dy = 1$. Dann erfüllt $F(x) = \int_{-\infty}^{x} f(y)\, dy$ die obigen Bedingungen und es gilt

$$\mathbf{P}\big((a, b]\big) = \int_a^b f(y)\, dy\,.$$

Eine solche Verteilung heißt stetige Verteilung. $\mathbf{P}$ ist eine Verteilung auf $(\mathbb{R}, \mathcal{B})$.

Bemerkung: Für stetige Verteilungen gilt $\mathbf{P}(\{a\}) = 0$ (Ü-Aufgabe).

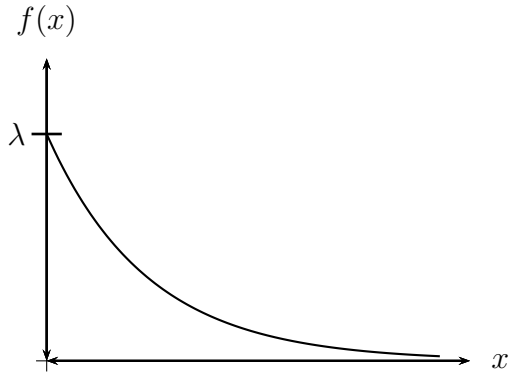
Beispiel 6.10 (Die Gleichverteilung $\mathcal{R}[a, b]$)
[auch Rechteck- oder Uniform-Verteilung genannt]

$$f(x) = \frac{1}{b-a}\, \mathrm{I}_{[a,b]}(x)\,, \qquad F(x) = \begin{cases} 0 & x \le a \\ \frac{x-a}{b-a} & a \le x \le b \\ 1 & x \ge b \end{cases}\,.$$

[$f(x)$ ist normiert!]

Beispiel 6.11 (Die Exponentialverteilung $\mathcal{E}(\lambda)$)

$$f(x) = \begin{cases} \lambda e^{-\lambda x} & , \quad x \ge 0 \\ 0 & , \quad x < 0 \end{cases}$$



Es gilt

$$\int_0^\infty \lambda e^{-\lambda x} dx = -e^{-\lambda x}\Big|_0^\infty = 1$$

und

$$F(x) = \int_0^x \lambda e^{-\lambda y} dy = \begin{cases} 1 - e^{-\lambda x} & , \quad x \geq 0 \\ 0 & , \quad x < 0 \end{cases} .$$

Anwendung: Die Lebensdauer einer Glühbirne ist näherungsweise $\mathcal{E}(\lambda)$-verteilt. □

Beispiel 6.12 (Die Normalverteilung $\mathcal{N}(\mu, \sigma^2)$)

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

Es gilt

$$\int_{-\infty}^{\infty} f_{\mu,\sigma}(x)\, dx \overset{y=\frac{x-\mu}{\sigma}}{=} \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}}\, e^{-\frac{1}{2}y^2}\, dy = (*)$$

und

$$\Big(\int_{-\infty}^{\infty} e^{-\frac{1}{2}y^2} dy\Big)^2 = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} e^{-\frac{1}{2}(x^2+y^2)} dx\, dy = \int_0^{2\pi}\int_0^{\infty} r\, e^{-\frac{1}{2}r^2} dr\, d\varphi$$
$$= 2\pi \left(- e^{-\frac{1}{2}r^2}\right)\Big|_0^{\infty} = 2\pi ,$$

d.h. $(*) = 1$.

Bezeichnung: $\varphi(x) := f_{0,1}(x) , \quad \Phi(x) := \int_{-\infty}^{x} f_{0,1}(y)\, dy .$
(Dichte und Verteilungsfunktion der Standard-Normalverteilung).

Anwendung: Fehlermodell bei Messungen; viele Messwerte sind näherungsweise $\mathcal{N}(\mu, \sigma^2)$ verteilt. □

Wie bei diskreten Zufallsvariablen in Kapitel 5 definieren wir nun stetige Zufallsvariable als Zufallsvariable, deren induzierte Verteilung eine stetige Verteilung ist. Konkret betrachten wir hier $X : \Omega \to \mathbb{R}$ mit einem Ausgangsraum $(\Omega, \mathcal{A}, \mathbf{P})$ und dem Bildraum $(\mathbb{R}, \mathcal{B}, \mathbf{P}^X)$, d.h. $\Omega^X \subset \mathbb{R}$ und $\mathbf{P}^X(A) := \mathbf{P}(X^{-1}(A))$.

Definition 6.13 (Stetige Zufallsvariable) Ist X eine messbare **(MT)** Zufallsvariable mit Bildmaß $\mathbf{P}^X\big((a,b]\big) := \mathbf{P}\big(X^{-1}((a,b])\big) = \int_a^b f(y)\, dy$, so heißt X stetige (oder stetig verteilte) Zufallsvariable mit Wahrscheinlichkeitsdichte $f(y)$ und Verteilungsfunktion $F(x) = \int_{-\infty}^{x} f(y)\, dy$.

Bemerkung 6.14

(i) Man braucht für obige Definition, dass $X^{-1}((a,b])$ in der σ - Algebra $\mathcal{A}$ liegt - sonst ist $\mathbf{P}\big(X^{-1}((a,b])\big)$ nicht definiert. Das besagt aber gerade die Annahme der Messbarkeit **(MT)**, die wir hier nicht vertiefen wollen. Wir bemerken lediglich, dass diese Annahme idR erfüllt ist, z.B. für alle stetigen oder monotonen Funktionen.

(ii) Wie bei diskreten Verteilungen schreibt man $X \sim \mathbf{P}^X$, z.B. $X \sim \mathcal{N}(\mu, \sigma^2)$.

(iii) Wie schon bei diskreten Zufallsvariablen spielt der Ausgangsraum $(\Omega, \mathcal{A}, \mathbf{P})$ häufig keine Rolle und wird dann auch nicht angegeben.

(iv) Bezeichnungskollision: Wenn X eine stetige Zufallsvariable ist, dann nimmt $\mathbf{P}^X$ die Rolle von $\mathbf{P}$ in den bisherigen Ausführungen dieses Kapitels (z.B. in Satz 6.8 und Definition 6.9) ein während das $\mathbf{P}$ aus Definition 6.13 ein anderes ist (nämlich die W-Verteilung auf dem Ausgangsraum).

(v) Wir ergänzen (ohne Beweis) ein paar vertiefende Überlegungen aus der Maßtheorie:
1) Definiert man wie in Kapitel 5 die Verteilungsfunktion von X durch $F_X(x) = \mathbf{P}(X \le x) = \mathbf{P}^X\big((-\infty, x]\big)$, so kann man zeigen, dass F_X die Voraussetzungen von Satz 6.8 erfüllt, d.h. es folgt aus Satz 6.8, dass $\mathbf{P}^X$ durch F_X bereits eindeutig bestimmt ist (gilt für beliebige ZVAs).
2) Andererseits kann man (unter Verwendung der Messbarkeit von X) nachrechnen, dass $\mathbf{P}^X(A) := \mathbf{P}(X^{-1}(A))$ eine Wahrscheinlichkeitsverteilung auf $(\mathbb{R}, \mathcal{B})$ ist. Wegen der Eindeutigkeit in Satz 6.8 handelt es sich damit um dieselbe Verteilung.

Bemerkung 6.15 (Lineare Transformation von Zufallsvariablen)

X Ergebnis der Temperaturmessung einer Flüssigkeit in $°C$, Y Ergebnis in $°F$.
Es gilt $Y = \frac{9}{5}X + 32$, allgemein: $Y = aX + b$.
Sei X stetig verteilt mit Dichte f_X und Vf F_X (z.B. $X \sim \mathcal{N}(\mu, \sigma^2)$). Wie ist Y verteilt, d.h. wie sehen f_Y und F_Y aus?

$$F_Y(y) = \mathbf{P}^Y((-\infty, y]) = \mathbf{P}(Y \le y) = \mathbf{P}(aX + b \le y) = \mathbf{P}(X \le \frac{y-b}{a}) = F_X(\frac{y-b}{a})$$

$$\Rightarrow \qquad f_Y(y) = \frac{d}{dy} F_X(\frac{y-b}{a}) = \frac{1}{a} \cdot f_X(\frac{y-b}{a}) .$$

Spezialfall: $X \sim \mathcal{N}(\mu, \sigma^2)$

$$\Rightarrow f_X(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{1}{2\sigma^2}(x-\mu)^2\right]$$
$$\Rightarrow f_Y(y) = \frac{1}{\sqrt{2\pi}} \frac{1}{\sigma a} \exp\left[-\frac{1}{2\sigma^2 a^2}(y-b-a\mu)^2\right]$$
$$\Rightarrow Y = aX + b \sim \mathcal{N}(a\mu + b, a^2\sigma^2)\,.$$

Satz 6.16 (Transformation von Zufallsvariablen)
Sei X eine stetig verteilte Zufallsvariable mit Dichte $f_X(x)$ und sei $Y = g(X)$ mit g streng monoton und differenzierbar. Dann ist auch Y stetig verteilt mit Dichte $f_Y(y) = f_X\big(g^{-1}(y)\big) \cdot |\frac{d}{dy} g^{-1}(y)|$ für $y = g(x)$ mit $f_X(x) > 0$ und $f_Y(y) = 0$ sonst.

Beweis. Sei g monoton wachsend. Dann gilt:

$$\begin{aligned}\mathbf{P}^Y\big([a,b]\big) &= \mathbf{P}\big(g(X) \in [a,b]\big) = \mathbf{P}\big(X \in [g^{-1}(a), g^{-1}(b)]\big) \\ &= \int_{g^{-1}(a)}^{g^{-1}(b)} f_X(x)\,dx = \int_a^b \underbrace{f_X\big(g^{-1}(y)\big)\,\frac{d}{dy} g^{-1}(y)}_{=f_Y(y)}\,dy.\end{aligned}$$

Der Fall g monoton fallend verläuft analog. □

Beispiel 6.17 (Die χ^2-Verteilung)
Sei $X \sim \mathcal{N}(0,1)$ und $Y = X^2$. Wie sieht $f_Y(y)$ aus?
$f_X(x) = \frac{1}{\sqrt{2\pi}}\, e^{-x^2/2}$, $g(x) = x^2$, $g^{-1}(y) = \sqrt{y}$, $\frac{d}{dy} g^{-1}(y) = \frac{1}{2\sqrt{y}}$.
Problem: g ist nicht monoton. Die Herleitung verläuft aber analog:
Es gilt für $0 \le a < b < \infty$:

$$\begin{aligned}\mathbf{P}\big(Y \in [a,b]\big) &= \mathbf{P}\big(X \in [\sqrt{a}, \sqrt{b}]\big) + \mathbf{P}\big(X \in [-\sqrt{b}, -\sqrt{a}\,]\big) \\ &= 2\,\mathbf{P}\big(X \in [\sqrt{a}, \sqrt{b}]\big) \\ &= 2\int_a^b \frac{1}{\sqrt{2\pi}}\, e^{-y/2}\, \frac{1}{2} y^{-1/2}\,dy\end{aligned}$$

$$\Rightarrow \quad f_Y(y) = \frac{1}{\sqrt{2\pi}}\, y^{-1/2} e^{-y/2}\,.$$

Y heißt χ^2-verteilt (mit einem Freiheitsgrad). □

Satz 6.18 (Neyman-Pearson Lemma)
Seien $\mathbf{P}^X$ und $\mathbf{P}^Y$ zwei stetige Wahrscheinlichkeitsverteilungen auf $(\mathbb{R}, \mathcal{B})$ mit Wahrscheinlichkeitsdichten f_X und f_Y. $L(z) := \frac{f_Y(z)}{f_X(z)} \leq \infty$ heißt der Likelihood-Quotient von $\mathbf{P}^Y$ bzgl. $\mathbf{P}^X$. Für das Testproblem $H_0 = \{\mathbf{P}^X\}$ gegen $H_A = \{\mathbf{P}^Y\}$ zum Niveau α ist

$$\phi^* : \mathbb{R} \to \{\overset{H_0}{\downarrow} 0\,,\ \overset{H_A}{\downarrow} 1\}$$
$$z \mapsto \begin{cases} 1 & , \quad L(z) \geq c^* \\ 0 & , \quad L(z) < c^* \end{cases}$$

mit $\mathbf{P}^X(\phi^* = 1) = \mathbf{P}^X\big(\{z \mid L(z) \geq c^*\}\big) = \alpha$ ein bester Test.

Beweis. Analog zu Satz 4.2 (**MT**) [Man benötigt, dass der Annahmebereich in $\mathcal{B}$ liegt und das Intergral darüber definiert ist.] □

Beispiel 6.19 Testen von $H_0 = \{\mathcal{N}(\mu_0, \sigma^2)\}$ gegen $H_A = \{\mathcal{N}(\mu_A, \sigma^2)\}$ mit $\mu_0 < \mu_A$ bei bekanntem σ^2. [Man schreibt oft auch $H_0 : \mu = \mu_0$ gegen $H_A : \mu = \mu_A$].

$$\begin{aligned} L(z) = \frac{f_Y(z)}{f_X(z)} &= \frac{e^{-\frac{1}{2\sigma^2}(z-\mu_A)^2}}{e^{-\frac{1}{2\sigma^2}(z-\mu_0)^2}} \\ &= e^{\frac{1}{2\sigma^2}[(z-\mu_0)^2-(z-\mu_A)^2]} \\ &= e^{\frac{1}{2\sigma^2}[z^2-2\mu_0 z+\mu_0^2-z^2+2\mu_A z-\mu_A^2]} \\ &= e^{\frac{2}{2\sigma^2}(\mu_A-\mu_0)z}\, e^{\frac{\mu_0^2-\mu_A^2}{2\sigma^2}} \\ &\geq c^* \end{aligned}$$

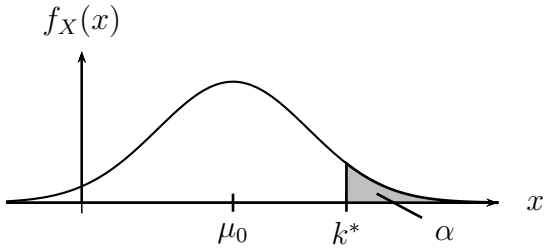
$$\iff \quad z \geq k^* \qquad \text{(monotoner Dichte-Quotient).}$$

Also:

$$\phi^*(z) = \begin{cases} 1 & , \quad z \geq k^* \\ 0 & , \quad z < k^* \end{cases} \qquad \begin{array}{l} \text{mit } \mathbf{P}^X(\phi^* = 1) = \mathbf{P}^X\big([k^*, \infty)\big) = \int_{k^*}^{\infty} f_X(z)\, dz = \alpha, \\ f_X \text{ Dichte der } \mathcal{N}(\mu_0, \sigma^2)\text{-Verteilung.} \end{array}$$

ist glm. bester Test für $H_0 = \{\mathcal{N}(\mu_0, \sigma^2)\}$ gegen $H_A = \{\mathcal{N}(\mu, \sigma^2) \mid \mu > \mu_0\}$.

Berechnung von k^*:



$$\begin{aligned} \mathbf{P}^X\big([k^*, \infty)\big) = \mathbf{P}(X \geq k^*) &= \mathbf{P}\big(\underbrace{\frac{X - \mu_0}{\sigma}}_{\sim \mathcal{N}(0,1)} \geq \frac{k^* - \mu_0}{\sigma}\big) \\ &= 1 - \Phi\left(\frac{k^* - \mu_0}{\sigma}\right) \stackrel{!}{=} \alpha. \end{aligned}$$

Man bestimmt die Lösung, indem man aus einer Tabelle (z.B. in Krengel oder Rice im Anhang) das α-Quantil $u_{1-\alpha} := \Phi^{-1}(1-\alpha)$ abliest und dann die Gleichung $u_{1-\alpha} = \frac{k^* - \mu_0}{\sigma}$ nach k^* auflöst, d.h. $k^* = \mu_0 + \sigma u_{1-\alpha}$.

Bemerkung: Es ist unrealistisch, solch einen Test auf der Grundlage nur einer einzigen Beobachtung durchzuführen. Wenn man n unabhängige identisch verteilte Beobachtungen hat, so erhält man den entsprechenden Test, wenn man die "multivariaten" Verteilungen anschaut (s. Satz 8.20). Das Neyman-Pearson Lemma gilt für diese Situation analog. Außerdem wird in Satz 8.20 die realistischere Situation $H_0 = \{\mathcal{N}(\mu, \sigma^2) \mid \mu \leq \mu_0\}$ gegen $H_A = \{\mathcal{N}(\mu, \sigma^2) \mid \mu > \mu_0\}$ betrachtet. $\square$

7 Erwartungswert und Varianz von Zufallsvariablen

Der Erwartungswert einer Zufallsvariable wird allgemein als Maßintegral von $g(x) = x$ bzgl. der von X induzierten Verteilung definiert. Da wir den Begriff des Maßintegrals aber nicht zur Verfügung haben, verwenden wir zwei getrennte Definitionen für diskrete und stetige ZVAs. Als weitere Kenngröße wird die Varianz definiert. Die Beziehung zwischen Varianz und der Streuung einer ZVA um den Erwartungswert ergibt sich aus der Tschebyscheff-Ungleichung.

Definition 7.1 (Erwartungswert)

(i) Sei X eine diskrete ZVA mit Zähldichte $p(x)$ und Träger $\{x_1, x_2, \ldots\}$. Dann heißt

$$\mathbf{E}X = \sum_i x_i \, p(x_i)$$

Erwartungswert von X, falls $\sum |x_i| \, p(x_i) < \infty$.

(ii) Sei X stetige ZVA mit Wahrscheinlichkeitsdichte $f(x)$. Dann heißt

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

Erwartungswert von X, falls $\int |x| f(x) \, dx < \infty$.

Bemerkung: **(MT)** Man beachte, dass es sich bei dem Erwartungswert im Grunde um den Erwartungswert der von X induzierten Verteilung $\mathbf{P}^X$ handelt (mit Zähldichte $p(x)$ bzw. W'dichte $f(x)$). Im Rahmen der Maßtheorie verwendet man als einheitliche Definition das Maßintegral $\mathbf{E}X := \int x \, d\mathbf{P}^X(x) = \int X \, d\mathbf{P}$. Obige Definitionen sind dann Spezialfälle.

Beispiel 7.2 (Erwarteter Gewinn beim Roulette)
$\Omega = \{0, \ldots, 36\}$, $\mathbf{P}$ Laplace-Verteilung, 1 € Einsatz auf "ungerade", X sei der Gewinn.

$$X(\omega) = \begin{cases} 1 & , \text{ falls } \omega \text{ ungerade} \\ -1 & , \text{ falls } \omega \text{ gerade} \end{cases};$$

$$\Omega^X = \{-1, 1\};$$

$$p(1) = \mathbf{P}^X\big(\{1\}\big) = \frac{18}{37},$$
$$p(-1) = \mathbf{P}^X\big(\{-1\}\big) = \frac{19}{37},$$

$$\mathbf{E}X = 1 \cdot \frac{18}{37} + (-1) \cdot \frac{19}{37} = -\frac{1}{37} \approx -0,03.$$

Man verliert also im Mittel 3 Ct pro Spiel. □

Beispiel 7.3

$$X \sim \mathcal{P}(\lambda), \qquad p(k) = \frac{\lambda^k}{k!} e^{-\lambda},$$

$$\mathbf{E}X = \sum_{k=0}^{\infty} k\, p(k) = \sum_{k=0}^{\infty} k\, \frac{\lambda^k}{k!}\, e^{-\lambda} = \lambda \cdot \sum_{k=1}^{\infty} \frac{\lambda^{k-1}}{(k-1)!}\, e^{-\lambda} = \lambda.$$

□

Beispiel 7.4

$$X \sim \mathcal{N}(\mu, \sigma^2), \qquad f(x) = \frac{1}{\sqrt{2\pi}\sigma}\, \exp\{-\frac{1}{2\sigma^2}\,(x-\mu)^2\},$$

$$\begin{aligned}
\mathbf{E}X &= \int_{-\infty}^{\infty} x\, f(x)\, dx \\
&\overset{y=x-\mu}{=} \int_{-\infty}^{\infty} y\, \frac{1}{\sqrt{2\pi}\sigma}\, \exp\{-\frac{1}{2\sigma^2} y^2\}\, dy + \int_{-\infty}^{\infty} \mu\, \frac{1}{\sqrt{2\pi}\sigma}\, \exp\{-\frac{1}{2\sigma^2} y^2\}\, dy \\
&= 0 + \mu = \mu.
\end{aligned}$$

□

Satz 7.5 *Für $g : \mathbb{R} \to \mathbb{R}$ und*

(i) X diskrete ZVA mit Zähldichte $p(x)$ und Träger $\{x_i \mid i \in \mathbb{N}\}$ gilt:

$$\mathbf{E}\, g(X) = \sum_{i=1}^{\infty} g(x_i)\, p(x_i)\ , \qquad \textit{falls } \sum |g(x_i)|\, p(x_i) < \infty.$$

(ii) X stetige ZVA mit Wahrscheinlichkeitsdichte $f(x)$ [sowie g messbar (MT)] gilt:

$$\mathbf{E}\, g(X) = \int_{-\infty}^{\infty} g(x)\, f(x)\, dx\ , \qquad \textit{falls } \int |g(x)|\, f(x)\, dx < \infty.$$

Beweis. [in der VL nur Teil (i) vortragen]

(i) Sei $Y := g(X)$, $\mathcal{Y} = \{y_1, y_2, \ldots\}$ und $A_i = g^{-1}\big(\{y_i\}\big)$. Dann gilt:

$$\begin{aligned}\mathbf{E}\, g(X) = \mathbf{E}\, Y &= \sum_i y_i\, \mathbf{P}^Y\big(\{y_i\}\big) = \sum_i y_i \sum_{x_j \in A_i} p(x_j) \\ &= \sum_i \sum_{x_j \in A_i} y_i\, p(x_j) = \sum_i \sum_{x_j \in A_i} g(x_j)\, p(x_j) \\ &= \sum_{j=1}^{\infty} g(x_j)\, p(x_j).\end{aligned}$$

(ii) Der Beweis im stetigen Fall ist komplizierter **($\to$ MT)**. Ist g zusätzlich streng monoton und differenzierbar, so folgt die Aussage mit $Y := g(X)$ aber unmittelbar aus Satz 6.16:

$$\mathbf{E}(g(X)) = \int_{-\infty}^{\infty} y\, f_Y(y)\, dy = \int_{-\infty}^{\infty} y\, f_X\big(g^{-1}(y)\big) |\frac{d}{dy} g^{-1}(y)|\, dy = \int_{-\infty}^{\infty} g(x) f_X(x) dx$$

[Man kann immer von $-\infty$ bis ∞ integrieren, da ggf. $f_Y(y) = 0$ bzw. $f_X(x) = 0$.]

[(MT) Allgemein folgt das Resultat aus der Transformationsformel für Maßintegrale: $\mathbf{E}(g(X)) = \int y\, d\mathbf{P}^{g(X)}(y) = \int g(x)\, d\mathbf{P}^X(x)$.] □

Bemerkung: Abgesehen von Spezialfällen gilt $\mathbf{E}\, g(X) \neq g(\mathbf{E}X)$.

Korollar 7.6 *Es gilt für alle* $a \leq b$

$$\mathbf{E}\,\mathrm{I}_{[a,b]}(X) = \mathbf{P}^X\big([a,b]\big).$$

Bemerkung: Allgemeiner gilt für alle $A \in \mathcal{B}$ $\mathbf{E}\,\mathrm{I}_A(X) = \mathbf{P}^X(A)$. Allerdings ist die Funktion $\mathrm{I}_A(x)$ nicht für alle $A \in \mathcal{B}$ Riemann-integrierbar. In diesen Fällen braucht man für die Definition von $\mathbf{E}\,\mathrm{I}_A(X)$ den Begriff des Maßintegrals **(MT)**.

Korollar 7.7 *Sei* X *Zufallsvariable und* $a, b \in \mathbb{R}$. *Dann gilt:*

$$\mathbf{E}(aX + b) = a\,\mathbf{E}X + b.$$

Satz 7.8 *Seien* X *und* Y *Zufallsvariable. Dann gilt:*

$$\mathbf{E}(X + Y) = \mathbf{E}X + \mathbf{E}Y.$$

Beweis. (für X, Y diskrete ZVA)
Sei

$$\begin{aligned} &\{x_1, x_2, \ldots\} \text{ Träger von } X, && A_i := X^{-1}\big(\{x_i\}\big), \\ &\{y_1, y_2, \ldots\} \text{ Träger von } Y, && B_j := Y^{-1}\big(\{y_j\}\big). \end{aligned}$$

Dann gilt

$$\begin{aligned} \mathbf{E}(X+Y) &= \sum_{i,j} (x_i + y_j)\,\mathbf{P}(A_i \cap B_j) \\ &= \sum_i x_i \sum_j \mathbf{P}(A_i \cap B_j) + \sum_j y_j \sum_i \mathbf{P}(A_i \cap B_j) \\ &= \sum_i x_i\,\mathbf{P}(A_i \cap \Omega) + \sum_j y_j\,\mathbf{P}(\Omega \cap B_j) \\ &= \mathbf{E}X + \mathbf{E}Y. \end{aligned}$$

□

Beispiel 7.9 Es sollen n Blutproben untersucht werden. p sei die Wahrscheinlichkeit, dass ein Bluttest negativ ausfällt.

1. Methode: Untersuche alle n Proben getrennt.

2. Methode: Teile die Proben in m Gruppen a k Präparate ein ($n = mk$) und mische die Proben innerhalb einer Gruppe.

$$\text{Test} \left\langle \begin{array}{lcl} \text{negativ} & \to & \text{alle } k \text{ Proben negativ} \\ \text{positiv} & \to & \text{untersuche alle } k \text{ Proben getrennt} \end{array} \right.$$

Sei X_i die Anzahl der Untersuchungen in der i-ten Gruppe.
Gesucht:

$$\begin{aligned} \mathbf{E}\,(\sum_{i=1}^{m} X_i) = \sum_{i=1}^{m} \mathbf{E}X_i &= \sum_{i=1}^{m} \left(1 \cdot p^k + (k+1)\,(1-p^k)\right) \\ &= m(k+1-kp^k) \\ &= n \underbrace{(1 + \frac{1}{k} - p^k)}_{=:A(k,p)} \overset{?}{\leqslant} n. \end{aligned}$$

Man kann $A(k,p)$ nach k minimieren:
z.B. $p = 0.99, \quad k \approx 10, \quad 1 + \frac{1}{k} - p^k \approx 0.2.$

□

Definition 7.10 *Sei X eine ZVA mit $\mathbf{E}X^2 < \infty$. Dann heißt*

$$\mathrm{Var}(X) = \mathbf{E}\big[\big(X - \mathbf{E}(X)\big)^2\big]$$

die Varianz von X. $\sigma = \sqrt{\mathrm{Var}(X)}$ heißt Standardabweichung von X. Die Varianz ist ein Maß für die "Streuung" der Verteilung um den Erwartungswert.

Satz 7.11 *[Sei X ZVA mit $\mathbf{E}X^2 < \infty$ und $a, b \in \mathbb{R}$.] Es gilt*

(i) $\mathrm{Var}(X) = \mathbf{E}(X^2) - (\mathbf{E}X)^2$*;*
(ii) $\mathrm{Var}(aX+b) = a^2\,\mathrm{Var}X$*.*

Beweis.
(i)

$$\begin{aligned}\mathrm{Var}X &= \mathbf{E}(X-\mathbf{E}X)^2\\ &= \mathbf{E}\big(X^2 - 2X\,\mathbf{E}X + (\mathbf{E}X)^2\big)\\ &= \mathbf{E}X^2 - 2\,(\mathbf{E}X)^2 + (\mathbf{E}X)^2 = \mathbf{E}X^2 - (\mathbf{E}X)^2.\end{aligned}$$

(ii)

$$\begin{aligned}\mathrm{Var}(aX+b) &= \mathbf{E}\Big\{\big[aX+b-\mathbf{E}(aX+b)\big]^2\Big\}\\ &= \mathbf{E}\big\{a^2[X-\mathbf{E}X]^2\big\}\\ &= a^2\,\mathrm{Var}(X).\end{aligned}$$

□

Beispiel 7.12 $X \sim \mathcal{N}(0,1) \quad [\Rightarrow \mathbf{E}X = \mu = 0]$.

$$\begin{aligned}\mathrm{Var}X &= \mathbf{E}\big[(X - \underbrace{\mathbf{E}X}_{=0})^2\big] = \int_{-\infty}^{\infty} x^2\varphi(x)\,dx\\ &= \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} x^2\,\exp(-\frac{1}{2}x^2)\,dx\\ &\overset{\text{partielle Integration}}{=} \frac{1}{\sqrt{2\pi}}\Big[-x\exp(-\frac{1}{2}x^2)\Big]_{-\infty}^{\infty} + \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty}\exp(-\frac{1}{2}x^2)\,dx\\ &= 0+1 = 1.\end{aligned}$$

Sei nun $Y = aX + b$. Dann gilt

$$\begin{aligned}&(6.16) \Rightarrow Y \sim \mathcal{N}(b, a^2); \qquad \text{[lineare Transf. von ZVAs]}\\ &(7.11) \Rightarrow \mathrm{Var}Y = a^2\,\mathrm{Var}X = a^2 \quad \text{und} \quad \mathbf{E}Y = a\,\mathbf{E}X + b = b;\end{aligned}$$

d.h.

$$\boxed{\begin{array}{c} Z \sim \mathcal{N}(\mu, \sigma^2) \\ \Rightarrow \mathbf{E}Z = \mu, \quad \mathrm{Var}Z = \sigma^2 \end{array}}$$

□

Beispiel 7.13 $X \sim \mathcal{P}(\lambda) \quad [\Rightarrow \mathbf{E}X = \lambda\,]$

$$\Rightarrow \quad \mathbf{E}\big[X(X-1)\big] = \sum_{k=0}^{\infty} k(k-1)\frac{\lambda^k}{k!}\,e^{-\lambda}$$
$$\uparrow \text{ [Trick]}$$
$$= \lambda^2 \sum_{k=2}^{\infty} \frac{\lambda^{k-2}}{(k-2)!}\,e^{-\lambda} = \lambda^2$$

$$\Rightarrow \quad \mathbf{E}X^2 = \mathbf{E}\big[X(X-1)\big] + \mathbf{E}X = \lambda^2 + \lambda$$
$$\Rightarrow \quad \mathrm{Var}X = \mathbf{E}X^2 - (\mathbf{E}X)^2 = \lambda^2 + \lambda - \lambda^2 = \lambda.$$

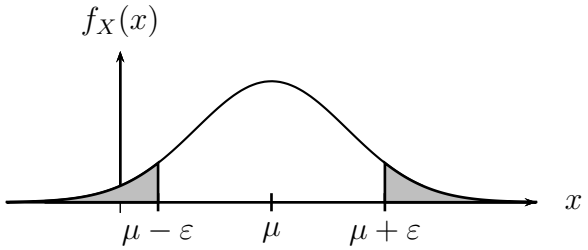
□

Bemerkung: $\mathbf{E}X^2 \neq (\mathbf{E}X)^2 \quad$ [Ausnahme: $X \equiv c$].

Bemerkung 7.14 $\mathbf{E}X$ und $\mathrm{Var}X$ sind im Grunde Kenngrößen der von X induzierten Verteilung $\mathbf{P}^X$, d.h. sie hängen nicht direkt von X sondern nur von $\mathbf{P}^X$ ab. Für diejenigen, die mit dem Maßtheorie vertraut sind, schreiben wir noch einmal die Definitionen als Maßintegral auf (die obigen Definitionen sind Spezialfälle davon):

$$\mu = \mathbf{E}X := \int x\, d\mathbf{P}^X(x) \quad \Big(= \int X\, d\mathbf{P} \Big),$$
$$\sigma^2 = \mathrm{Var}X := \int (x-\mu)^2\, d\mathbf{P}^X(x) \quad \Big(= \int (X-\mu)^2\, d\mathbf{P} \Big).$$

Wir wollen nun ausführlich die Frage diskutieren, weshalb $\mathrm{Var}X$ ein Maß für die "Streuung" von X ist. Dafür untersuchen wir den Ausdruck $\mathbf{P}(|X - \mathbf{E}X| \geq \varepsilon)$.



Satz 7.15 (Tschebyscheff-Ungleichung) *Sei* $\mathbf{E}X^2 < \infty$. *Dann gilt für alle* $a \in \mathbb{R}$ *und* $\varepsilon > 0$

$$\mathbf{P}(|X - a| \geq \varepsilon) \leq \frac{\mathbf{E}(X-a)^2}{\varepsilon^2}.$$

Für $a = \mathbf{E}X$ *erhält man* $\mathbf{P}(|X - \mathbf{E}X| \geq \varepsilon) \leq \frac{\mathrm{Var}(X)}{\varepsilon^2}$.

Beweis. (für X diskrete ZVA)
Sei $\mathcal{T}(X) = \{x_1, x_2, \ldots\}$. Dann gilt

$$\begin{aligned}
\mathbf{P}(|X-a| \geq \varepsilon) &= \sum_{\substack{i=1 \\ |x_i - a| \geq \varepsilon}}^{\infty} \mathbf{P}(X = x_i) \\
&\leq \sum_{\substack{i=1 \\ |x_i - a| \geq \varepsilon}}^{\infty} \frac{(x_i - a)^2}{\varepsilon^2} \mathbf{P}(X = x_i) \\
&\leq \frac{1}{\varepsilon^2} \sum_{i=1}^{\infty} (x_i - a)^2 \mathbf{P}(X = x_i) = \frac{1}{\varepsilon^2} \mathbf{E}(X-a)^2.
\end{aligned}$$

□

Korollar 7.16

(i) Sei $\mathbf{E}X^2 < \infty$ *und* $\sigma = \sqrt{\mathrm{Var}X} > 0$. *Dann gilt*

$$\mathbf{P}(|X - \mathbf{E}X| \geq k\sigma) \leq \frac{1}{k^2};$$

also z.B.

$$\mathbf{P}(|X - \mathbf{E}X| \geq 2\sigma) \leq \frac{1}{4} = 0.25;$$
$$\mathbf{P}(|X - \mathbf{E}X| \geq 3\sigma) \leq \frac{1}{9} \approx 0.11.$$

(ii) Gilt $\mathrm{Var}X = 0$*, so folgt* $\mathbf{P}(X = \mathbf{E}X) = 1$.

Bemerkung 7.17 Die Tschebyscheff-Ungleichung ist nur eine grobe Abschätzung. Sie gilt für alle ZVAs. Kennt man die Verteilung von X, so kann man die Wahrscheinlichkeit genau ausrechnen. Sei z.B. $X \sim \mathcal{N}(\mu, \sigma^2)$. Dann gilt

$$\begin{aligned}
\mathbf{P}\big(|X - \mu| < k\sigma\big) &= \mathbf{P}\big(\big|\frac{X-\mu}{\sigma}\big| < k\big) \\
&= \mathbf{P}\big(-k < \underbrace{\frac{X-\mu}{\sigma}}_{\sim\mathcal{N}(0,1)} \leq k\big) - \underbrace{\mathbf{P}\big(\frac{X-\mu}{\sigma} = k\big)}_{=0} \\
&= \Phi(k) - \Phi(-k) \\
&\overset{(*)}{=} 2\,\Phi(k) - 1.
\end{aligned}$$

Damit folgt

$$\mathbf{P}\big(|X - \mu| \geq k\sigma\big) = 1 - \big(2\,\Phi(k) - 1\big) = 2 - 2\,\Phi(k);$$

z.B.

$$\mathbf{P}\big(|X - \mu| \geq \sigma\big) \approx 2 - 2 \cdot 0,84 = 0,32;$$
$$\mathbf{P}\big(|X - \mu| \geq 2\sigma\big) \approx 2 - 2 \cdot 0,98 = 0,04;$$
$$\mathbf{P}\big(|X - \mu| \geq 3\sigma\big) \approx 2 - 2 \cdot 0,9987 = 0,0026.$$

Zu $(*)$: Die Dichte φ der $\mathcal{N}(0,1)$ -Verteilung ist symmetrisch, d.h. $\varphi(x) = \varphi(-x)$. Damit folgt für die Verteilungsfunktion Φ:

$$\begin{aligned}
\Phi(x) = \int_{-\infty}^{x} \varphi(y)\,dy &= 1 - \int_{x}^{\infty} \varphi(y)\,dy \\
&= 1 - \int_{-\infty}^{-x} \varphi(y)\,dy \\
&= 1 - \Phi(-x).
\end{aligned}$$

□

8 Mehrdimensionale Verteilungen und Stochastische Unabhängigkeit von Zufallsvariablen

In diesem Kapitel werden mehrdimensionale Verteilungen definiert und gemeinsam verteilte Zufallsvariable betrachtet. Das wichtigste Beispiel dafür ist die multivariate Normalverteilung, die aber erst in Kapitel 12 behandelt wird. Wir werden sehen, dass die gemeinsame Verteilung von mehreren Zufallsvariablen bei gleichen eindimensionalen Verteilungen verschieden sein kann. Außerdem definieren wir die Stochastische Unabhängigkeit von Zufallsvariablen. Als Anwendung leiten wir einen gleichmäßig besten Test für den Erwartungswert bei Normalverteilungen her - diesmal basierend auf n unabhängigen Beobachtungen.

Beispiel 8.1 Gegeben seien zwei Zufallsvariable, z.B.

$$X = \text{Größe eines Fisches},$$
$$Y = \text{Alter eines Fisches}.$$

Kann man von den Werten von X auf die Werte von Y schließen oder sind X und Y "unabhängig"? Offensichtlich gibt es eine Abhängigkeit zwischen X und Y. Aus Gründen der Klarheit studieren wir zwei einfachere ZVAs (dreimaliges Werfen einer Münze):

$$\Omega = \{(\omega_1, \omega_2, \omega_3) \mid \omega_i \in \{0, 1\}\};$$

$$\begin{array}{ll} X(\omega) = \omega_1, & \Omega^X = \{0, 1\}; \\ Y(\omega) = \sum_{i=1}^{3} \omega_i, & \Omega^Y = \{0, 1, 2, 3\}. \end{array}$$

Analog zum eindimensionalen Fall definiert $\mathbf{P}^{X,Y}(C) := \mathbf{P}\big((X, Y)^{-1}(C)\big)$ eine Verteilung auf $\big(\Omega^X \times \Omega^Y, \mathcal{P}(\Omega^X \times \Omega^Y)\big)$ wobei $\Omega^X \times \Omega^Y := \big\{(x, y) \mid x \in \Omega^X, y \in \Omega^Y\big\}$. $\mathbf{P}^{X,Y}$ heißt die

gemeinsame Verteilung von (X, Y).

Beispiel: $C = \{(0,1),(1,2)\}$ $\quad (X,Y)^{-1}(C) = \big\{(0,1,0),(0,0,1),(1,1,0),(1,0,1)\big\}$

$$\Rightarrow \mathbf{P}^{X,Y}(C) = \frac{|(X,Y)^{-1}(C)|}{|\Omega|} = \frac{4}{8} = \frac{1}{2}.$$

$p_{X,Y}(x_i, y_j) := \mathbf{P}^{X,Y}\big(\{(x_i, y_j)\}\big)$ heißt gemeinsame Zähldichte von X und Y. Man erhält folgende Zähldichte:

x_i \ y_i	0	1	2	3
0	1/8	2/8	1/8	0
1	0	1/8	2/8	1/8

Man kann aus $p_{X,Y}(x_i, y_j)$ die Zähldichte $p_X(x_i)$ von X berechnen, indem man bzgl. der anderen Komponente aufsummiert:

$$\begin{aligned} p_X(x_i) &= \mathbf{P}(X = x_i) = \mathbf{P}(X = x_i, Y \text{ beliebig}) \\ &= \sum_{j=1}^{\infty} \mathbf{P}(X = x_i, Y = y_j) = \sum_{j=1}^{\infty} p_{X,Y}(x_i, y_j). \end{aligned}$$

Analoges gilt auch für stetige Verteilungen. □

Für einen Zufallsvektor $(X_1, \ldots, X_n)$ wollen wir nun wie bei eindimensionalen Verteilungen die induzierte Verteilung durch

$$\mathbf{P}^{X_1,\ldots,X_n}(A) = \mathbf{P}\big((X_1, \ldots, X_n)^{-1}(A)\big) = \mathbf{P}\big(\{\omega \,|\, (X_1(\omega), \ldots, X_n(\omega)) \in A\}\big)$$

für "geeignete" $A \in \mathbb{R}^n$ definieren. Die Probleme dabei sind völlig analog zu denen in Kapitel 6 (vgl. insbesondere Definition 6.6, Satz 6.8 und Bemerkung 6.14(v)).

Definition 8.2 (MT) *Die von dem Mengensystem*

$$\mathcal{E} = \big\{(a_1, b_1] \times \ldots \times (a_n, b_n] \,\big|\, a_i < b_i, a_i, b_i \in \mathbb{R}\big\}$$

erzeugte σ-Algebra $\mathcal{B}^n$ heißt Borelsche σ - Algebra auf $\mathbb{R}^n$.

Bemerkung: Unter anderem liegen alle n-dimensionalen Rechtecke, Kugeln, Zylinder, etc. in $\mathcal{B}^n$.

Seien nun $X_1, \ldots, X_n$ messbare **(MT)** Zufallsvariable auf einem gemeinsamen Wahrscheinlichkeitsraum $(\Omega, \mathcal{A}, \mathbf{P})$ und

$$F(x_1, \ldots, x_n) := \mathbf{P}(X_1 \leq x_1, \ldots, X_n \leq x_n)$$

die gemeinsame Verteilungsfunktion von $(X_1, \ldots, X_n)$. Wie im Fall $n = 1$ gilt:

Proposition 8.3 (MT) *F definiert eine eindeutig bestimmte Wahrscheinlichkeitsverteilung $\mathbf{P}^{X_1,\ldots,X_n}$ auf $(\mathbb{R}^n, \mathcal{B}^n)$ mit $\mathbf{P}^{X_1,\ldots,X_n}(A) = \mathbf{P}\big((X_1, \ldots, X_n)^{-1}(A)\big)$ für alle $A \in \mathcal{B}^n$.*

Beweis. MT - wie für n=1 Spezialfall des Maßerweiterungssatzes. □

Bemerkungen: (i) Im diskreten Fall gilt $\mathcal{P}(\Omega^{X_1} \times \ldots \times \Omega^{X_n}) \subset \mathcal{B}^n$, d.h. F definiert dann auch eine eindeutig bestimmte Wahrscheinlichkeitsverteilung auf
$\big(\Omega^{X_1} \times \ldots \times \Omega^{X_n}, \mathcal{P}(\Omega^{X_1} \times \ldots \times \Omega^{X_n})\big)$.
(ii) Wie im Kapitel 6 benötigt man zur Definition einer Verteilung auf $(\mathbb{R}^n, \mathcal{B}^n)$ keine ZVAs, sondern lediglich eine Funktion F mit bestimmten Eigenschaften.

Definition 8.4

(i) *Diskrete Zufallsvariable: Sind alle Zufallsvariable diskret, so heißt $p(x_1, \ldots, x_n) := \mathbf{P}(X_1 = x_1, \ldots, X_n = x_n)$ die gemeinsame Zähldichte. Für $I \subset \{1, \ldots, n\}$ gilt*

$$p_{(X_i; i \in I)}(x_i; i \in I) = \mathbf{P}(X_i = x_i; i \in I) = \sum_{\substack{x_j \\ j \in I^C}} p(x_1, \ldots, x_n).$$

(ii) *Stetige Zufallsvariable: Gilt $F(x_1, \ldots, x_n) = \int_{-\infty}^{x_1} \ldots \int_{-\infty}^{x_n} f(y_1, \ldots, y_n)\, dy_n \cdots dy_1$, so heißt $f(x_1, \ldots, x_n)$ die gemeinsame Wahrscheinlichkeitsdichte von $X_1, \ldots, X_n$. Für $I \subset \{1, \ldots, n\}$ ist*

$$f_{(X_i; i \in I)}(x_i; i \in I) = \int \cdots \int f(x_1, \ldots, x_n) \prod_{j \in I^C} (dx_j)$$

die Wahrscheinlichkeitsdichte von $(X_i, i \in I)$ (Marginaldichte).

Beispiel/Bemerkung 8.5 Sei (X_1, X_2) stetig verteilt mit Wahrscheinlichkeitsdichte $f(x_1, x_2)$. Dann gilt

$$\begin{aligned}
\mathbf{P}\big(X_1 \in (a,b], X_2 \in (c,d]\big) &= \mathbf{P}(X_1 \le b, X_2 \le d) - \mathbf{P}(X_1 \le a, X_2 \le d) \\
&\quad - \mathbf{P}(X_1 \le b, X_2 \le c) + \mathbf{P}(X_1 \le a, X_2 \le c) \\
&= F(b,d) - F(a,d) - F(b,c) + F(a,c) \\
&= \int_{-\infty}^{b}\int_{-\infty}^{d} f(x_1,x_2)\,dx_2dx_1 - \int_{-\infty}^{a}\int_{-\infty}^{d} f(x_1,x_2)\,dx_2dx_1 \\
&\quad - \int_{-\infty}^{b}\int_{-\infty}^{c} f(x_1,x_2)\,dx_2dx_1 + \int_{-\infty}^{a}\int_{-\infty}^{c} f(x_1,x_2)\,dx_2dx_1 \\
&= \int_{a}^{b}\int_{c}^{d} f(x_1,x_2)\,dx_2dx_1.
\end{aligned}$$

Allgemeiner gilt für alle Mengen $A \in \mathcal{B}^2$

$$\mathbf{P}\big((X_1,X_2) \in A\big) = \iint_A f(x_1,x_2)\,dx_1dx_2$$

(z.B. für Kreise, Dreiecke, etc.), wobei man für allgemeine $A \in \mathcal{B}^2$ die rechte Seite zunächst als Lebesgue-Intergral auffassen muss **(MT)**. Falls das obige Integral aber als Riemann-Integral existiert, so sind die beiden Integrale gleich **(MT)**. □

Allgemein gilt:

Satz 8.6 *Sei $(X_1, \ldots, X_n)$ stetig verteilt mit Wahrscheinlichkeitsdichte $f(x_1, \ldots, x_n)$. Dann gilt*

$$\mathbf{P}\big(X_1 \in (a_1,b_1], \ldots, X_n \in (a_n,b_n]\big) = \int_{a_1}^{b_1} \ldots \int_{a_n}^{b_n} f(x_1,\ldots,x_n)\,dx_n \cdots dx_1.$$

und

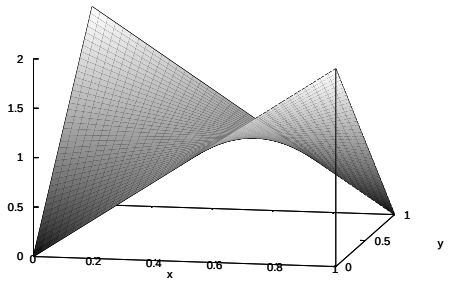
$$\mathbf{P}\big((X_1,\ldots,X_n) \in A\big) = \int_A \cdots \int f(x_1,\ldots,x_n)\,dx_n \cdots dx_1$$

für solche $A \in \mathcal{B}^n$, für die das Riemann-Integral existiert.

Beweis. Erster Teil analog zu obigem Beispiel. Teil 2 $\rightarrow$ **MT**. [Eindeutigkeit der Maßfortsetzung + Gleichheit von Riemann- und Lebesgue-Integral.] $\square$

Beispiel 8.7 Seien X, Y gemeinsam verteilt mit folgenden Dichten:

(i)
$$f(x,y) = \begin{cases} 2x + 2y - 4\,xy & , \quad x, y \in [0,1] \\ 0 & , \quad \text{sonst} \end{cases}.$$

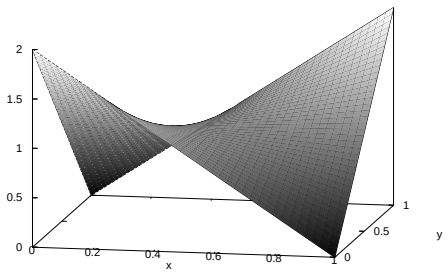


Es gilt

$$f(0,0) = 0 \text{ und}$$
$$f_X(x) = \int_0^1 \big(2x + 2y - 4xy\big)\, dy = 2x + 1 - 2x = 1, \quad x \in [0,1],$$
$$f_Y(y) = \int_0^1 \big(2x + 2y - 4xy\big)\, dx = 1, \quad y \in [0,1],$$

d.h. $X \sim \mathcal{R}[0,1]$ und $Y \sim \mathcal{R}[0,1]$.

(ii)
$$f(x,y) = \begin{cases} 2 - 2x - 2y + 4xy & , \quad x, y \in [0,1] \\ 0 & , \quad \text{sonst} \end{cases}.$$



Es gilt

$$f(0,0) = 2 \text{ und}$$
$$f_X(x) = \int_0^1 \big(2 - 2x - 2y + 4xy\big)\, dy = 2 - 2x - 1 + 2x = 1, \quad x \in [0,1],$$
$$f_Y(y) = 1, \quad y \in [0,1],$$

d.h. $X \sim \mathcal{R}[0,1]$ und $Y \sim \mathcal{R}[0,1]$.

(iii)

$$f(x,y) = \begin{cases} 1 & , \quad x,y \in [0,1] \\ 0 & , \quad \text{sonst} \end{cases} .$$

Es gilt

$$f_X(x) = 1, \quad x \in [0,1],$$
$$f_Y(y) = 1, \quad y \in [0,1],$$

d.h. $X \sim \mathcal{R}[0,1]$ und $Y \sim \mathcal{R}[0,1]$.

Obwohl in allen 3 Fällen X und Y $\mathcal{R}[0,1]$ - verteilt sind, ist die gemeinsame Verteilung in allen Fällen unterschiedlich. □

Bemerkung: Das wichtigste Beispiel für eine mehrdimensionale Verteilung ist die multivariate Normalverteilung, die ausführlich in Kapitel 12 behandelt wird. Man beachte schon

einmal die Plots der 2-dimensionalen Dichten in Kapitel 12.

Beispiel 8.8 (Motivation zur Unabhängigkeit von Zufallsvariablen)
Zur Erinnerung:

Zwei Ereignisse A und B sind unabhängig.

$\Leftrightarrow \mathbf{P}(A \cap B) = \mathbf{P}(A)\,\mathbf{P}(B)$

$\Leftrightarrow \mathbf{P}(A|B) = \mathbf{P}(A) \qquad (\text{ falls } \mathbf{P}(B) > 0\ )$

[$\Leftrightarrow$ Das Eintreten des Ereignisses B liefert keine Information über die Wahrscheinlichkeit des Eintretens von A.]

Seien nun X, Y Zufallsvariable. Wann sind X und Y stochastisch unabhängig?

Zweimaliges Würfeln:

$$\Omega = \big\{\omega = (\omega_1, \omega_2) \mid \omega_i \in \{1, \ldots, 6\}\big\}, \quad |\Omega| = 36;$$
$$X(\omega) = \omega_1, \qquad Y(\omega) = \omega_2;$$

$$\mathbf{P}(X = 3, Y = 2) = \frac{1}{36} = \frac{1}{6} \cdot \frac{1}{6} = \mathbf{P}(X = 3)\,\mathbf{P}(Y = 2),$$

d.h. die Ereignisse $\{X = 3\}$ und $\{Y = 2\}$ sind stochastisch unabhängig. Für die Unabhängigkeit von X und Y wollen wir verlangen, dass alle Ereignisse $\{X = a\}$ und $\{Y = b\}$ für $a, b \in \{1, \ldots, 6\}$ unabhängig sind. Allgemeiner sollen $\{X \in A\}$ und $\{Y \in B\}$ für $A, B \subset \{1, \ldots, 6\}$ unabhängig sein.

Sei $Z(\omega) := \omega_1 + \omega_2$. Sind X und Z unabhängig?

$$\{Z = 7\} = \{(1,6), (2,5), (3,4), (4,3), (5,2), (6,1)\},$$
$$\big|\{X = a\} \cap \{Z = 7\}\big| = \big|\{(a, 7-a)\}\big| = 1 \qquad \forall a \in \{1, \ldots, 6\}$$
$$\Rightarrow \mathbf{P}\big(X = a, Z = 7\big) = \frac{1}{36} = \mathbf{P}\big(X = a\big)\,\mathbf{P}\big(Z = 7\big)$$
$$\Rightarrow \{X = a\} \text{ und } \{Z = 7\} \text{ sind unabhängig } \forall a$$

aber

$$\{Z = 12\} \cap \{X = 1\} = \emptyset$$
$$\Rightarrow X \text{ und } Z \text{ sind } \underline{\text{nicht}} \text{ stochastisch unabhängig.}$$

□

Definition 8.9 (Stochastische Unabhängigkeit)
$X_1, \ldots, X_n : \Omega \to \mathbb{R}$ *heißen* *stochastisch unabhängig, falls*

$$\mathbf{P}(X_1 = x_1, \ldots, X_n = x_n) = \mathbf{P}(X_1 = x_1) \cdots \mathbf{P}(X_n = x_n) \quad \forall x_i \in \mathbb{R} \quad [x_i \in \mathcal{T}(X_i) \text{ reicht}]$$

(diskrete ZVA), bzw.

$$\mathbf{P}\big(X_1 \in (a_1, b_1], \ldots, X_n \in (a_n, b_n]\big) = \prod_{i=1}^{n} \mathbf{P}\big(X_i \in (a_i, b_i]\big) \qquad \forall a_i, b_i \in \mathbb{R} \tag{3}$$

(stetige ZVA) gilt.

Proposition 8.10 (MT) $X_1, \ldots, X_n$ *sind genau dann stochastisch unabhängig sind, wenn*

$$\mathbf{P}(X_1 \in B_1, \ldots, X_n \in B_n) = \prod_{i=1}^{n} \mathbf{P}(X_i \in B_i) \quad \forall B_i \in \mathcal{B}.$$

[gilt für stetige und diskrete ZVAs]. Insbesondere gilt

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

Beweis. "⇐" ist trivial. "⇒" folgt im diskreten Fall durch aufsummieren. Im allgemeinen Fall benötigt man wiederum den Maßerweiterungssatz **(MT)**. □

Satz 8.11 *Seien X_i stetig verteilte ZVAs. Dann gilt*

$$X_i \text{ stochastisch unabhängig} \quad \Leftrightarrow \quad f_{X_1 \ldots X_n}(x_1, \ldots, x_n) = \prod_{i=1}^{n} f_{X_i}(x_i). \tag{4}$$

Beweis. "⇐":

$$\begin{aligned}
&\mathbf{P}\big(X_1 \in (a_1, b_1], \ldots, X_n \in (a_n, b_n]\big) \\
&= \int_{a_1}^{b_1} \ldots \int_{a_n}^{b_n} f(x_1, \ldots, x_n)\, dx_n \cdots dx_1 = \prod_{i=1}^{n} \int_{a_i}^{b_i} f_{X_i}(x_i)\, dx_i \\
&= \prod_{i=1}^{n} \mathbf{P}\big(X_i \in (a_i, b_i]\big).
\end{aligned}$$

"⇒": Bilde partielle Ableitungen von (3) bezüglich b_i. □

Beispiel 8.12 (Fortsetzung von Beispiel 8.7)
Von den drei Dichten aus Beispiel 8.7 ist nur im Fall (iii) die Bedingung (4) erfüllt, d.h. nur im Fall (iii) sind die ZVAs X und Y stochastisch unabhängig.

Beispiel 8.13 (Faltung von unabhängigen Zufallsvariablen)
Seien $X \sim \mathcal{B}(n,p)$, $Y \sim \mathcal{B}(m,p)$ und X, Y stochastisch unabhängig. $X + Y \sim\ ?$

$$\begin{aligned}
\mathbf{P}(X+Y=k) &= \sum_{l=0}^{k} \mathbf{P}(X+Y=k, X=l) \\
&\overset{X,Y\text{unabh.}}{=} \sum_{l=0}^{k} \mathbf{P}(X=l)\, \mathbf{P}(Y=k-l) \\
&= \sum_{l=0}^{k} \binom{n}{l} p^l\, q^{n-l} \binom{m}{k-l} p^{k-l}\, q^{m-(k-l)} \\
&= \sum_{l=0}^{k} \binom{n}{l}\binom{m}{k-l} p^k q^{n+m-k} \\
&= \binom{n+m}{k} p^k q^{n+m-k} \\
&\quad\nwarrow \\
&\qquad \text{[z.B. Normierung der Hypergeometrischen Verteilung]}
\end{aligned}$$

$$\Rightarrow X + Y \sim \mathcal{B}(n+m, p).$$

Induktiv: $X_1, \ldots, X_n \sim \mathcal{B}(1,p) \quad \Rightarrow \quad \sum_{i=1}^{n} X_i \sim \mathcal{B}(n,p)$. □

Bemerkung 8.14 Faltungseigenschaften anderer Verteilungen (X, Y jeweils stochastisch unabhängig)

$$\begin{aligned} X \sim \mathcal{P}(\lambda),\, Y \sim \mathcal{P}(\mu) \quad &\Rightarrow \quad X + Y \sim \mathcal{P}(\lambda + \mu)\,, \\ X \sim \mathcal{N}(\mu_1, \sigma_1^2),\, Y \sim \mathcal{N}(\mu_2, \sigma_2^2) \quad &\Rightarrow \quad X + Y \sim \mathcal{N}(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2)\,. \end{aligned}$$

Satz 8.15 *Seien $X_1, \ldots, X_n$ stochastisch unabhängig, $h_1, \ldots, h_n$ (messbare (**MT**)) Funktionen. Dann sind auch $h_1(X_1), \ldots, h_n(X_n)$ stochastisch unabhängig.*

Beweis. Für alle $B_i \in \mathcal{B}$ gilt

$$\begin{aligned} \mathbf{P}\big(h_1(X_1) \in B_1, \ldots, h_n(X_n) \in B_n\big) &= \mathbf{P}\big(X_1 \in h_1^{-1}(B_1), \ldots, X_n \in h_n^{-1}(B_n)\big) \\ &= \prod_{i=1}^{n} \mathbf{P}\big(X_i \in h_i^{-1}(B_i)\big) = \prod_{i=1}^{n} \mathbf{P}\big(h_i(X_i) \in B_i\big). \end{aligned}$$

□

Bemerkung: (i) Die Aussage gilt auch für Vektoren X_i.
(ii) Man braucht in obigem Beweis, dass die $h_i^{-1}(B_i)$ wieder in $\mathcal{B}$ liegen. Dass besagt genau die Annahme der Messbarkeit.

Satz 8.16 (Mehrdimensionale Version von Satz 7.5) *Seien $X_1, \ldots, X_n$ ZVAs mit gemeinsamer Verteilung und $g : \mathbb{R}^n \to \mathbb{R}$.*

(i) Sind X_i diskret mit gemeinsamer Zähldichte $p(x_1, \ldots, x_n)$ so gilt

$$\mathbf{E}g(X_1, \ldots, X_n) = \sum_{x_1, \ldots, x_n} g(x_1, \ldots, x_n)\, p(x_1, \ldots, x_n),$$

falls $\sum |g(x_1, \ldots, x_n)|\, p(x_1, \ldots, x_n) < \infty$.

(ii) Sind X_i stetig mit gemeinsamer Wahrscheinlichkeitsdichte $f(x_1, \ldots, x_n)$ so gilt

$$\mathbf{E}\, g(X_1, \ldots, X_n) = \int \cdots \int g(x_1, \ldots, x_n)\, f(x_1, \ldots, x_n)\, dx_n \cdots dx_1,$$

falls $\int \cdots \int\, |g(x_1, \ldots, x_n)|\, f(x_1, \ldots, x_n)\, dx_n \cdots dx_1 < \infty$.

Beweis. Analog zu Satz 7.5. □

Satz 8.17 *Seien X und Y stochastisch unabhängig mit $\mathbf{E}X^2 < \infty$, $\mathbf{E}Y^2 < \infty$. Dann gilt*

$$\mathbf{E}XY \;=\; \mathbf{E}X\;\mathbf{E}Y.$$

Beweis. (für stetige ZVAs - diskreter Fall analog)

$$\begin{aligned}\mathbf{E}XY &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} xy\, f_{X,Y}(x,y)\,dx\,dy = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} xy\, f_X(x)\, f_Y(y)\,dx\,dy \\ &= \int_{-\infty}^{\infty} x f_X(x)\,dx \int_{-\infty}^{\infty} y f_Y(y)\,dy = \mathbf{E}X\;\mathbf{E}Y.\end{aligned}$$

□

Satz 8.18 *Seien $X_1, \dots X_n$ stochastisch unabhängig mit $\mathbf{E}X_i^2 < \infty$. Dann gilt*

$$\mathrm{Var}\big(\sum_{i=1}^{n} X_i\big) = \sum_{i=1}^{n} \mathrm{Var}(X_i).$$

Beweis. Wegen $\mathrm{Var}(X) = \mathrm{Var}(X - \mathbf{E}X)$ (Satz 7.11 (ii)) kann man Œ annehmen, dass $\mathbf{E}X_i = 0\ \forall i$. Damit gilt

$$\begin{aligned}\mathrm{Var}\big(\sum_{i=1}^{n} X_i\big) = \mathbf{E}\Big[\big(\sum_{i=1}^{n} X_i\big)^2\Big] &= \sum_{i,j=1}^{n} \mathbf{E}(X_i\, X_j) \\ &= \sum_{i=1}^{n} \mathbf{E}(X_i^2) + \sum_{i\neq j} \mathbf{E}X_i\, \mathbf{E}X_j \\ &= \sum_{i=1}^{n} \mathrm{Var}(X_i).\end{aligned}$$

□

Anwendung 8.19 Seien $X_1, \dots, X_n$ unabhängig und identisch verteilt (iid: "independently and identically distributed") mit $\mu := \mathbf{E}X_i$ und $\sigma^2 := \mathrm{Var}(X_i)$ $(\mathbf{E}X_i^2 < \infty)$. Betrachte den Mittelwert $\overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$. Es gilt $\mathbf{E}\overline{X}_n = \mu$ und

$$\mathrm{Var}(\overline{X}_n) = \frac{1}{n^2}\mathrm{Var}(\sum_{i=1}^{n} X_i) = \frac{1}{n^2}\sum_{i=1}^{n} \mathrm{Var}X_i = \frac{\sigma^2}{n}.$$

In Kapitel 10 werden wir zeigen, dass $\overline{X}_n$ sogar in einem bestimmten Sinne gegen μ konvergiert. Für unabhängige $X_1, \ldots, X_n \sim \mathcal{N}(\mu, \sigma^2)$ folgt aus Bemerkung 8.14 sogar, dass $\overline{X}_n$ wieder Normal-verteilt ist:

$$\overline{X}_n := \frac{1}{n} \sum_{i=1}^{n} X_i \sim \mathcal{N}\big(\mu, \frac{\sigma^2}{n}\big).$$

□

Mit obigen Resultaten können wir nun das Testproblem $H_0 : \mu \leq \mu_0$ gegen $H_A : \mu > \mu_0$ basierend auf n unabhängigen Beobachtungen einer $\mathcal{N}(\mu, \sigma^2)$ - Verteilung betrachten.

Satz 8.20 (Gleichmäßig bester Test bei Normalverteilungen)
Seien $X_1, \ldots, X_n$ unabhängig und identisch $\mathcal{N}(\mu, \sigma^2)$ - verteilte Zufallsvariable. Dann ist

$$\phi^*(X_1, \ldots, X_n) = \begin{cases} 1 & , \quad \overline{X}_n > c^* \\ 0 & , \quad \overline{X}_n \leq c^* \end{cases}$$

mit $\mathbf{P}_{\mu_0}(\phi^* = 1) = \mathbf{P}_{\mu_0}\big(\overline{X}_n > c^*\big) \overset{(*)}{=} \alpha$ ein gleichmäßig bester Test für $H_0 : \mu \leq \mu_0$ gegen $H_A : \mu > \mu_0$ zum Niveau α; d.h. ϕ^* minimiert $\mathbf{P}_\mu(\phi = 0)$ gleichmäßig für alle $\mu > \mu_0$ über alle $\phi : \mathbb{R}^n \to \{0, 1\}$ mit $\mathbf{P}_\mu(\phi = 1) \leq \alpha$ für alle $\mu \leq \mu_0$. Es gilt $c^* = \mu_0 + \frac{\sigma}{\sqrt{n}} u_{1-\alpha}$ mit $\Phi(u_{1-\alpha}) = 1 - \alpha$ [Tabelle!].

Beweis. Wie in Beispiel 6.19 betrachten wir zunächst das Problem des Testens von $H_0 = \{\mathcal{N}(\mu_0, \sigma^2)\}$ gegen $H_A = \{\mathcal{N}(\mu_A, \sigma^2)\}$ mit festem $\mu_A > \mu_0$, d.h. wir testen die Verteilung $\mathbf{P}_{\mu_0}^{X_1,\ldots,X_n}$ gegen die Verteilung $\mathbf{P}_{\mu_A}^{X_1,\ldots,X_n}$. Das Neyman-Pearson-Lemma in Satz 6.18 gilt identisch auch für multivariate Verteilungen - der Likelihood-Quotient $L(x_1, \ldots, x_n)$ beträgt jetzt wegen der Unabhängigkeit der Beobachtungen

$$L(x_1,\ldots,x_n) = \frac{f_{\mu_A}(x_1,\ldots,x_n)}{f_{\mu_0}(x_1,\ldots,x_n)} = \frac{\prod_{i=1}^n e^{-\frac{1}{2\sigma^2}(x_i-\mu_A)^2}}{\prod_{i=1}^n e^{-\frac{1}{2\sigma^2}(x_i-\mu_0)^2}}$$

$$= e^{\sum_{i=1}^n \frac{1}{2\sigma^2}[(x_i-\mu_0)^2-(x_i-\mu_A)^2]}$$

$$= e^{\sum_{i=1}^n \frac{1}{2\sigma^2}[x_i^2-2\mu_0 x_i+\mu_0^2-x_i^2+2\mu_A x_i-\mu_A^2]}$$

$$= e^{\sum_{i=1}^n \frac{2}{2\sigma^2}(\mu_A-\mu_0)x_i}\, e^{n\frac{\mu_0^2-\mu_A^2}{2\sigma^2}} \;\geq\; \gamma^*$$

$$\Leftrightarrow \quad \frac{1}{n}\sum_{i=1}^n x_i \geq c^* \qquad \text{(monotoner Dichte-Quotient).}$$

wobei c^* gerade $(*)$ erfüllt, d.h. ϕ^* ist ein bester Test für $H_0 : \mu = \mu_0$ gegen $H_A : \mu = \mu_A$ und wegen der Monotonie des Dichtequotienten auch für $H_0 : \mu = \mu_0$ gegen $H_A : \mu > \mu_0$ (da c^* nicht von μ_A abhängt). Wir zeigen nun, dass $\mathbf{P}_\mu(\phi^* = 1) \leq \alpha$ für alle $\mu \leq \mu_0$ gilt. Wegen $\mathbf{P}_\mu^{\overline{X}_n} = \mathcal{N}\big(\mu, \frac{\sigma^2}{n}\big)$ gilt

$$\mathbf{P}_\mu(\phi^* = 1) = \mathbf{P}_\mu(\overline{X}_n > c^*) = \mathbf{P}_\mu^{\overline{X}_n}\big((c^*,\infty)\big) \leq \mathbf{P}_{\mu_0}^{\overline{X}_n}\big((c^*,\infty)\big) = \mathbf{P}_{\mu_0}(\phi^* = 1) = \alpha$$

[mit einer Skizze unmittelbar einsichtig!]. Da die Menge $\big\{\phi : \mathbb{R}^n \to \{0,1\} \,\big|\, \mathbf{P}_\mu(\phi = 1) \leq \alpha$ für alle $\mu \leq \mu_0\big\}$ kleiner ist als die Menge $\big\{\phi : \mathbb{R}^n \to \{0,1\} \,\big|\, \mathbf{P}_{\mu_0}(\phi = 1) \leq \alpha\big\}$ [letztere hat weniger Restriktionen] folgt die Behauptung. Ferner gilt unter $\mu = \mu_0$
$\sqrt{n}\,\frac{\overline{X}_n-\mu_0}{\sigma} \sim \mathcal{N}(0,1)$, d.h. $(*)$ führt zu der Bestimmungsgleichung

$$\mathbf{P}_{\mu_0}(\overline{X}_n > c^*) = \mathbf{P}_{\mu_0}\big(\sqrt{n}\,\frac{\overline{X}_n - \mu_0}{\sigma} \geq \sqrt{n}\,\frac{c^*-\mu_0}{\sigma}\big) = 1 - \Phi\left(\sqrt{n}\,\frac{c^*-\mu_0}{\sigma}\right) \stackrel{!}{=} \alpha.$$

und damit zu $c^* = \mu_0 + \frac{\sigma}{\sqrt{n}}\,u_{1-\alpha}$. □

Bemerkungen: (i) Der obige beste Test hängt von (dem meistens unbekannten) σ^2 ab. Ist σ^2 unbekannt, so kann man zeigen, dass dann der sogenannte t-Test optimal ist (s. Kapitel 13 und Statistik I).
(ii) Ähnliche Resultate gelten auch für andere Verteilungen mit monotonen Dichtequotienten (insbesondere bei stetigen Verteilungen).

9 Konfidenzintervalle

In diesem Kapitel werden Konfidenzintervalle eingeführt. Es wird gezeigt, wie man optimale Konfidenzintervalle durch Umformung des Annahmebereichs von optimalen Tests erhalten kann.

Bemerkung 9.1 (Konfidenzintervalle) Basierend auf Beobachtungen $X = (X_1, \ldots, X_n)$ möchten wir ein Intervall $S(X)$ angeben, in dem ein bestimmter Parameter mit hoher Wahrscheinlichkeit liegt. Genauer heißt $S(X) \subset \Theta$ ein $(1-\alpha)$-Konfidenzintervall für θ, falls

$$\mathbf{P}_\theta\big(\theta \in S(X)\big) \geq 1 - \alpha \qquad \forall \theta \in \Theta.$$

Ein solches Intervall kann man praktisch immer aus einem Schätzer für θ konstruieren, falls dessen Verteilung bekannt ist. Beispiel: Es gilt für $X_i \overset{\text{iid}}{\sim} \mathcal{N}(\mu, \sigma^2)$

$$\begin{aligned}
&\overline{X}_n \sim \mathcal{N}\big(\mu, \frac{\sigma^2}{n}\big) \quad \Rightarrow \quad \sqrt{n}\,\frac{\overline{X}_n - \mu}{\sigma} \sim \mathcal{N}(0,1) \\
&\Rightarrow\ \mathbf{P}\big(u_{\alpha/2} \leq \sqrt{n}\,\frac{\overline{X}_n - \mu}{\sigma} \leq u_{1-\alpha/2}\big) = \Phi(u_{1-\alpha/2}) - \Phi(u_{\alpha/2}) = 1 - \alpha/2 - \alpha/2 = 1 - \alpha \\
&\Rightarrow\ \mathbf{P}\big(\overline{X}_n - \frac{\sigma}{\sqrt{n}}\,u_{1-\alpha/2} \leq \mu \leq \overline{X}_n + \frac{\sigma}{\sqrt{n}}\,u_{1-\alpha/2}\big) = 1 - \alpha \quad \big(\text{wegen } u_{\alpha/2} = -u_{1-\alpha/2}\big).
\end{aligned}$$

d.h. $[\overline{X}_n - \frac{\sigma}{\sqrt{n}}\,u_{1-\alpha/2}, \overline{X}_n + \frac{\sigma}{\sqrt{n}}\,u_{1-\alpha/2}]$ ist ein $(1-\alpha)$-Konfidenzintervall für μ (bei bekanntem σ). Analog rechnet man nach, dass auch $(-\infty, \overline{X}_n + \frac{\sigma}{\sqrt{n}}\,u_{1-\alpha}]$ und $[\overline{X}_n - \frac{\sigma}{\sqrt{n}}\,u_{1-\alpha}, \infty)$ $(1-\alpha)$-Konfidenzintervalle sind. Weitere Konfidenzintervalle für μ kann man z.B. aus anderen Schätzern für μ [wie dem bisher noch nicht behandelten Median] konstruieren.

Damit stellt sich die Frage nach einem optimalen Konfidenzintervall.

Beispiel 9.2

Modell für die Emission von α-Teilchen (ionisierende Strahlung bei radioaktivem Zerfall):

Zeit X_i zwischen Emission des i-ten und $(i+1)$-ten Teilchens $X_i \sim \mathcal{E}(\lambda)$, X_i iid,
$f(t) = \lambda e^{-\lambda t}$, $f(x_1, \ldots, x_n) = \lambda^n e^{-\lambda \sum_{i=1}^n x_i}$, $\mathbf{E}X_i = \frac{1}{\lambda}$ (erwartete Zeit zwischen Zerfällen).

Die Radioaktivität des Materials wird in Becquerel gemessen

= Anzahl der Zerfälle pro Sekunde

$\approx \frac{1}{\text{erwartete Zeit zwischen Zerfällen}} = \lambda$ [Achtung: $\mathbf{E}\big(\frac{1}{X}\big) \neq \frac{1}{\mathbf{E}X}$]

Problem: Man ist aus Sicherheitsgründen an einer oberen Grenze $\overline{\theta}(X)$ für $\theta := \lambda$ (berechnet aus Beobachtungen $X = (X_1, \ldots, X_n)'$) interessiert mit

(i) $\mathbf{P}_\theta\big(\theta \leq \overline{\theta}(X)\big) = \mathbf{P}_\theta\big(\theta \in [0, \overline{\theta}(X)]\big) \geq 1 - \alpha \quad \forall \theta \in \Theta,$

d.h. an einem $(1-\alpha)$-Konfidenzintervall von der Form $[0, \overline{\theta}(X)]$. Andererseits ist klar, dass $\overline{\theta}(X)$ möglichst klein sein sollte [$\overline{\theta}(X) = \infty$ wäre nicht sinnvoll!]. Man fordert dieses in der Form

(ii) $\mathbf{P}_\theta\big(\theta' \leq \overline{\theta}(X)\big) = \mathbf{P}_\theta\big(\theta' \in [0, \overline{\theta}(X)]\big) = \min \quad \forall \theta' > \theta \quad \forall \theta.$

Erfüllt $\overline{\theta}(X)$ (i) und (ii), dann heißt $[0, \overline{\theta}(X)]$ gleichmäßig bestes $(1-\alpha)$-Konfidenzintervall bei falschen Parametern $\theta' \in \overline{K}(\theta) = \{\overline{\theta} \,|\, \overline{\theta} > \theta\}$. Allgemein heißt $S(X) \subset \Theta$ ein gleichmäßig bestes $(1-\alpha)$-Konfidenzintervall bei falschen Parametern $\theta' \in \overline{K}(\theta)$, falls

$$\mathbf{P}_\theta\big(\theta \in S(X)\big) \geq 1 - \alpha \qquad \forall \theta \in \Theta$$

und

$$\mathbf{P}_\theta\big(\theta' \in S(X)\big) = \min \quad \forall \theta' \in \overline{K}(\theta) \quad \forall \theta \in \Theta.$$

Satz 9.3 *(i) Für alle $\theta' \in \Theta$ sei $\phi_{\theta'}(x)$ ein nicht randomisierter Test zum Niveau α für $H : \theta \in H(\theta')$ [z.B. $H(\theta') = (-\infty, \theta']$, d.h. θ' entspricht θ_0 bzw. μ_0, z.B. in Satz 8.20] gegen $K : \theta \in K(\theta')$ und $A(\theta') = \{x | \phi_{\theta'}(x) = 0\}$ sei der Annahmebereich des Tests. Sei*

$$S(X) := \{\theta \in \Theta | X \in A(\theta)\}.$$

Dann ist $S(X)$ ein $(1-\alpha)$-Konfidenzintervall für θ.

(ii) Ist $\phi^*_{\theta'}$ für alle θ' ein nicht randomisierter gleichmäßig bester Test für $H(\theta')$ gegen $K(\theta')$ und $S^*(X)$ wie oben, so gilt

$$\mathbf{P}_\theta\big(\theta' \in S^*(X)\big) = \min \quad \forall \theta \in K(\theta') \quad \forall \theta' \in \Theta \tag{5}$$

unter der Nebenbedingung

$$\mathbf{P}_\theta\big(\theta \in S^*(X)\big) \geq 1-\alpha \quad \forall \theta \in \Theta, \tag{6}$$

d.h. $S^*(X)$ ist ein gleichmäßig bestes $(1-\alpha)$-Konfidenzintervall bei falschen Parametern $\theta' \in \overline{K}(\theta) = \{\overline{\theta} | \theta \in K(\overline{\theta})\}$.

Beweis.

(i) Es gilt

$$\begin{aligned} \mathbf{P}_\theta\big(\theta \in S(X)\big) &= \mathbf{P}_\theta\big(X \in A(\theta)\big) = \mathbf{P}_\theta\big(\phi_\theta(X) = 0\big) \\ &= 1 - \mathbf{P}_\theta\big(\phi_\theta(X) = 1\big) \geq 1-\alpha. \end{aligned}$$

(ii) Sei $S(X)$ ein anderes $(1-\alpha)$ Konfidenzintervall, und

$$\phi_\theta(x) = \begin{cases} 1, & \theta \notin S(X) \\ 0, & \theta \in S(X) \end{cases}.$$

Dann gilt

$$\mathbf{P}_\theta\big(\phi_\theta(X) = 1\big) = 1 - \mathbf{P}_\theta\big(\theta \in S(X)\big) \leq \alpha$$

d.h. ϕ_θ ist Test zum Niveau α. Daraus folgt $\forall \theta \in K(\theta')$

$$\mathbf{P}_\theta\big(\theta' \in S^*(X)\big) \overset{(i)}{=} 1 - \mathbf{P}_\theta\big(\phi^*_{\theta'}(X) = 1\big) \leq 1 - \mathbf{P}_\theta\big(\phi_{\theta'}(X) = 1\big) = \mathbf{P}_\theta\big(\theta' \in S(X)\big).$$

Damit gilt (5) $\forall \theta \in K(\theta')\ \forall \theta' \in \Theta$. Das ist aber dasselbe wie $\forall \theta' \in \overline{K}(\theta)\ \forall \theta \in \Theta$ mit $\overline{K}(\theta) = \{\theta' | \theta \in K(\theta')\}$.

□

Bemerkung: **(MT)** Man braucht außerdem, dass die Mengen $\{x | \theta' \in S(x)\}$ in $\mathcal{B}^n$ liegen.

Beispiel 9.4 (Glm. beste untere Konfidenzschranke bei Normalverteilungen) Seien $X_1, \ldots, X_n$ unabhängig und identisch $\mathcal{N}(\mu, \sigma^2)$ - verteilte Zufallsvariable und $\overline{X}_n = \frac{1}{n}\sum_{i=1}^n X_i$. In Satz 8.20 haben wir gezeigt, dass

$$\phi^*(X_1, \ldots, X_n) = \begin{cases} 1 & , \quad \overline{X}_n > c^* \\ 0 & , \quad \overline{X}_n \leq c^* \end{cases}$$

mit $c^* = \mu' + \frac{\sigma}{\sqrt{n}}\, u_{1-\alpha}$ [Bem.: $\Phi(u_{1-\alpha}) = 1 - \alpha$!] ein gleichmäßig bester Test für $H_0 = \{\mathcal{N}(\mu, \sigma^2) \mid \mu \leq \mu'\}$ gegen $H_A = \{\mathcal{N}(\mu, \sigma^2) \mid \mu > \mu'\}$ zum Niveau α ist (d.h. $K(\mu') = (\mu', \infty)$). Damit gilt ($\theta = \mu$)

$$\begin{aligned} X \in A(\mu) &\Leftrightarrow \overline{X}_n \leq \mu + \frac{\sigma}{\sqrt{n}}\, u_{1-\alpha} \\ &\Leftrightarrow \mu \geq \overline{X}_n - \frac{\sigma}{\sqrt{n}}\, u_{1-\alpha} \Leftrightarrow \mu \in \Big[\overline{X}_n - \frac{\sigma}{\sqrt{n}}\, u_{1-\alpha}, \infty\Big), \end{aligned}$$

d.h. $\big[\overline{X}_n - \frac{\sigma}{\sqrt{n}}\, u_{1-\alpha}, \infty\big)$ ist nach Satz 9.3 ein gleichmäßig bestes $(1-\alpha)$-Konfidenzintervall bei falschen Parametern $\overline{K}(\mu) = \{\mu' \,|\, \mu \in K(\mu')\} = \{\mu' \,|\, \mu \in (\mu', \infty)\} = \{\mu' \,|\, \mu' < \mu\} = (-\infty, \mu)$. Für die gleiche Fragestellung mit unbekanntem σ^2 werden wir in Kapitel 13 ein leicht modifiziertes Konfidenzintervall aus dem t-Test konstruieren.

Bemerkung 9.5 Die letzte Umformung zeigt noch einmal deutlich, wie der Ablehnbereich des Tests $K(\theta')$ und die falsche Parametermenge $\overline{K}(\theta) = \{\theta' | \theta \in K(\theta')\}$ zusammenhängen. Beispiele:

$$\begin{array}{ll} K(\theta') = \{\theta \,|\, \theta > \theta'\} = (\theta', \infty) & \overline{K}(\theta) = \{\theta' \,|\, \theta > \theta'\} = (-\infty, \theta) \\ K(\theta') = \{\theta \,|\, \theta < \theta'\} = (-\infty, \theta') & \overline{K}(\theta) = \{\theta' \,|\, \theta < \theta'\} = (\theta, \infty) \\ K(\theta') = \{\theta \,|\, \theta \neq \theta'\} & \overline{K}(\theta) = \{\theta' \,|\, \theta \neq \theta'\} \end{array}$$

In Beispiel 9.4 erhält man z.B. dann aus dem optimalen Test für $H_0 : \mu = \mu'$ gegen $H_A : \mu \neq \mu'$ (d.h. $K(\mu') = \{\mu \,|\, \mu \neq \mu'\}$) das gleichmäßig beste $(1-\alpha)$-Konfidenzintervall $\big[\overline{X}_n - \frac{\sigma}{\sqrt{n}}\, u_{1-\alpha/2}, \overline{X}_n + \frac{\sigma}{\sqrt{n}}\, u_{1-\alpha/2}\big]$ für μ (s. Statistik I).

10 Stochastische Konvergenz und das schwache Gesetz der großen Zahlen

In der Stochastik gibt es 4 wichtige Konvergenzbegriffe für Zufallsvariable, darunter die fast sichere Konvergenz (s. Wahrscheinlichkeitstheorie I) und die Konvergenz der Verteilungen (siehe Kapitel 14). In diesem Kapitel definieren wir die stochastische Konvergenz und die Konvergenz im quadratischen Mittel. Wir zeigen dann das schwache Gesetz der großen Zahlen, d.h. die stochastische Konvergenz des empirischen Mittelwerts gegen den Erwartungswert. In einem Beispiel betrachten wir ferner verschiedene Schätzer für den rechten Eckpunkt einer Rechteckverteilung und zeigen für alle Schätzer die Konsistenz. Außerdem betrachten wir den 'mean squared error' MSE als Gütekriterium für Schätzer und berechnen in dem Beispiel für alle Schätzer diesen MSE.

Definition 10.1 (Stochastische Konvergenz / konsistenter Schätzer)
Eine Folge (Z_n) von ZVAs konvergiert stochastisch (oder in Wahrscheinlichkeit) gegen eine ZVA Z ($Z_n \xrightarrow{P} Z$) falls

$$\mathbf{P}(|Z_n - Z| \geq \varepsilon) \overset{n\to\infty}{\longrightarrow} 0 \quad \text{für alle } \varepsilon > 0.$$

Ist Z_n ein auf Daten basierender Schätzer und $Z \equiv \theta$ ein zu schätzender Parameter (einer Verteilung), so heißt Z_n auch konsistenter Schätzer für θ.

Satz 10.2 (Schwaches Gesetz der großen Zahlen)
Seien $X_1, \ldots, X_n$ unabhängig und identisch verteilt mit $\mathbf{E}X_i^2 < \infty$, $\mu := \mathbf{E}X_i$ und $\sigma^2 := \mathrm{Var}(X_i)$. Dann gilt

$$\overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i \xrightarrow{P} \mu,$$

d.h. $\overline{X}_n$ ist ein konsistenter Schätzer für μ.

Beweis. Tschebyscheff-Ungleichung:

$$\mathbf{P}(|\overline{X}_n - \mu| \geq \varepsilon) \leq \frac{\mathrm{Var}(\overline{X}_n)}{\varepsilon^2} = \frac{\frac{1}{n^2}\mathrm{Var}\big(\sum_{i=1}^{n} X_i\big)}{\varepsilon^2} = \frac{\sigma^2}{n\varepsilon^2} \to 0.$$

□

Bemerkung: Der Satz sagt nichts darüber aus, wie nah $\overline{X}_n$ für feste n an μ liegt (der Beweis gibt nur eine (schlechte) Approximation). Wir wissen auch nicht, ob es nicht noch "bessere" Schätzer für μ gibt als $\overline{X}_n$.

Beispiel 10.3 (Schätzung von a bei der $\mathcal{R}[0,a]$ - Verteilung)

Wir beobachten $X_1 \ldots, X_n \overset{\text{iid}}{\sim} \mathcal{R}[0,a]$, wobei $a > 0$ unbekannt ist. Wir wollen a schätzen.

1. Idee: $\mathbf{E}X_1 = \int_0^a x \frac{1}{a}\, dx = \frac{a}{2}$.
   Satz 10.2 $\Rightarrow$ $\overline{X}_n \xrightarrow{P} \frac{a}{2}$.
   Wähle also $S_{1,n} = S_1(X_1, \ldots, X_n) = 2\,\overline{X}_n$ als Schätzer.
   Es gilt $S_{1,n} \xrightarrow{P} a$, d.h. $S_{1,n}$ ist konsistent.

2. Idee:

   Wähle $S_{2,n} = \gamma \cdot \max(X_1, \ldots, X_n) \qquad \gamma = ?$

Gütekriterium für einen Schätzer S eines Parameters a:

$$R(S,a) := \mathbf{E}_a(S-a)^2$$

heißt mittlerer quadratischer Fehler (MSE: "mean squared error"). Es gilt

$$\begin{aligned}
\mathbf{E}_a(S-a)^2 &= \mathbf{E}_a(S - \mathbf{E}_a S + \mathbf{E}_a S - a)^2 \\
&= \mathbf{E}_a(S - \mathbf{E}_a S)^2 + 2\,\mathbf{E}_a\big[(\mathbf{E}_a S - a)\,(S - \mathbf{E}_a S)\big] + (\mathbf{E}_a S - a)^2 \\
&= \mathbf{E}_a(S - \mathbf{E}_a S)^2 + 2\,(\mathbf{E}_a S - a)(\mathbf{E}_a S - \mathbf{E}_a S) \; + (\mathbf{E}_a S - a)^2 \\
&= \mathrm{Var}_a(S) + (\underbrace{\mathbf{E}_a S - a}_{\text{BIAS}})^2.
\end{aligned}$$

Schätzer 1:

$$\mathbf{E}_a S_{1,n} = 2\,\mathbf{E}_a \overline{X}_n = \frac{2}{n}\sum_{i=1}^{n} \mathbf{E}_a X_i = a \;\Rightarrow\; BIAS = 0,$$

$$\mathrm{Var}_a(S_{1,n}) = \frac{4}{n^2}\sum_{i=1}^{n} \mathrm{Var} X_i = \frac{a^2}{3\,n},$$

$$\left[\begin{array}{c} X_i \sim R[0,a], \qquad \mathbf{E}X_i = \frac{a}{2}, \qquad \mathbf{E}X_i^2 = \int_0^a x^2 \frac{1}{a}\,dx = \frac{a^2}{3} \\ \Rightarrow \mathrm{Var} X_i = \mathbf{E}X_i^2 - (\mathbf{E}X_i)^2 = \frac{a^2}{3} - \frac{a^2}{4} = \frac{a^2}{12} \end{array}\right]$$

$$\Rightarrow R(S_{1,n,}a) = \frac{a^2}{3\,n}.$$

Schätzer 2: Wir benötigen die Verteilung von $M = \max\{X_1, \ldots, X_n\}$:

$$F_M(x) = \mathbf{P}(M \le x) = \mathbf{P}(X_1 \le x) \cdot \ldots \cdot \mathbf{P}(X_n \le x) = \big(\frac{x}{a}\big)^n$$

$$\Rightarrow \quad \text{Dichte: } f_M(x) = \frac{d}{dx} F_M(x) = \frac{n}{a^n} x^{n-1} \qquad \text{für } 0 \le x \le a$$

$$\Rightarrow \quad \mathbf{E}\,S_{2,n} = \mathbf{E}\,\gamma M = \gamma \int_0^a x f_M(x)\,dx = \frac{\gamma\,n}{a^n}\int_0^a x^n\,dx = \gamma\,a\frac{n}{n+1}$$

Damit gilt: Der Schätzer ist unverfälscht (d.h. BIAS = 0) $\Leftrightarrow \gamma = \frac{n+1}{n}$ (plausibel!). $R(S,a)$ wird aber nicht notwendigerweise für dieses γ minimal.

$$\begin{aligned} \mathrm{Var}\,S_{2,n} = \mathbf{E}\,S_{2,n}^2 - (\mathbf{E}S_{2,n})^2 &= \gamma^2 \int_0^a x^2\,\frac{n}{a^n} x^{n-1}\,dx - \big(\gamma\,a\,\frac{n}{n+1}\big)^2 \\ &= \frac{\gamma^2\,n}{a^n}\frac{a^{n+2}}{n+2} - \big(\gamma\,a\,\frac{n}{n+1}\big)^2 \\ &= \gamma^2\,a^2\frac{n}{n+2} - \gamma^2\,a^2\frac{n^2}{(n+1)^2} \\ &= \gamma^2\,a^2\,\frac{n(n+1)^2 - n^2(n+2)}{(n+2)(n+1)^2} \\ &= \gamma^2\,a^2\,\frac{n}{(n+1)^2(n+2)}. \end{aligned}$$

Für den unverfälschten Schätzer $S^*_{2,n}$ ($\gamma = \frac{n+1}{n}$) erhält man damit als MSE

$$R(S^*_{2,n}, a) = \frac{a^2}{n(n+2)} < \frac{a^2}{3\,n} = R(S_{1,n}, a) \quad \text{für } n > 1.$$

Wir wollen nun den MSE von $S_{2,n}$ bezüglich γ minimieren:

$$R(S^{(\gamma)}_{2,n}, a) = \gamma^2 a^2 \frac{n}{(n+1)^2(n+2)} + \big(\gamma\, a\, \frac{n}{n+1} - a\big)^2$$

$$\begin{aligned}
\frac{\partial}{\partial\gamma} R(S^{(\gamma)}_{2,n}, a) &= 2\gamma a^2 \frac{n}{(n+1)^2(n+2)} + 2\big(\gamma\, a\, \frac{n}{n+1} - a\big)\, a \frac{n}{n+1} = 0 \\
&\Leftrightarrow 2\gamma\big(\underbrace{\frac{1}{(n+1)(n+2)} + \frac{n}{n+1}}_{\frac{n^2+2n+1}{(n+1)(n+2)} = \frac{n+1}{n+2}} \big) - 2 = 0 \\
&\Leftrightarrow \gamma = \frac{n+2}{n+1}\,.
\end{aligned}$$

Sei also $\tilde{S}_{2,n} := S^{(\frac{n+2}{n+1})}_{2,n}$.

$$\begin{aligned}
\Rightarrow R(\tilde{S}_{2,n}, a) &= a^2 \left(\frac{n(n+2)}{(n+1)^4} + \Big[\frac{n(n+2)}{(n+1)^2} - 1\Big]^2 \right) \\
&= a^2 \left(\frac{n(n+2)}{(n+1)^4} + \frac{1}{(n+1)^4} \right) = a^2 \frac{1}{(n+1)^2}\,.
\end{aligned}$$

Es gilt

$$\frac{a^2}{(n+1)^2} < \frac{a^2}{n(n+2)} < \frac{a^2}{3\,n},$$

d.h. der Schätzer mit $\gamma = \frac{n+2}{n+1}$ ist besser als der unverfälschte Schätzer mit $\gamma = \frac{n+1}{n}$. $\square$

Definition 10.4 (Konvergenz im quadratischen Mittel)
Eine Folge (Z_n) von ZVAs konvergiert im quadratischen Mittel gegen eine ZVA Z ($Z_n \xrightarrow{(2)} Z$) falls

$$\mathbf{E}\,(Z_n - Z)^2 \overset{n\to\infty}{\longrightarrow} 0.$$

Proposition 10.5 *Es gilt*

$$Z_n \xrightarrow{(2)} Z \quad \Rightarrow \quad Z_n \xrightarrow{P} Z\,.$$

Beweis. Tschebyscheff-Ungleichung:

$$\mathbf{P}(|Z_n - Z| \geq \varepsilon) \leq \frac{\mathbf{E}\,(Z_n - Z)^2}{\varepsilon^2} \to 0.$$

□

Bemerkung: Die Umkehrung gilt nicht. Beispiel: Sei $X \sim \mathcal{R}[0,1]$, $Z_n = n\,\mathrm{I}_{[0,1/n]}(X)$ und $Z \equiv 0$. Dann gilt $Z_n \xrightarrow{P} Z$ aber $\mathbf{E}\,(Z_n - Z)^2 = n \to \infty$ [bitte nachrechnen!].

Beispiel 10.6 In Beispiel 10.3 haben wir für alle drei Schätzer $\mathbf{E}_a\,(S_n - a)^2 \to 0$ bewiesen. Damit konvergieren alle Schätzer im quadratischen Mittel und damit auch stochastisch gegen a, d.h. alle drei Schätzer sind konsistent.

11 Kovarianz und Korrelation

Nach dem Erwartungswert und der Varianz führen wir mit der Kovarianz die dritte Kenngröße ein. Sie ist ein Maß für die Übereinstimmung zweier ZVA. Die Korrelation ist eine auf das Intervall $[-1,1]$ standardisierte Variante der Kovarianz.

Definition 11.1 *Seien X und Y gemeinsam verteilt mit $\mathbf{E}X^2 < \infty$, $\mathbf{E}Y^2 < \infty$. Dann heißt*

$$\mathrm{Kov}(X,Y) := \mathbf{E}\big[(X - \mathbf{E}X)(Y - \mathbf{E}Y)\big]$$

die Kovarianz von X und Y und

$$\rho(X,Y) := \frac{\mathrm{Kov}(X,Y)}{\sqrt{\mathrm{Var}X \cdot \mathrm{Var}Y}} \qquad \mathrm{Var}X, \mathrm{Var}Y \neq 0$$

der Korrelationskoeffizient von X und Y.

Bemerkung 11.2

(i) Es gilt

$$\begin{aligned}\mathrm{Kov}(X,Y) &= \mathbf{E}XY - \mathbf{E}(X\mathbf{E}Y) - \mathbf{E}(Y\mathbf{E}X) + \mathbf{E}X\mathbf{E}Y \\ &= \mathbf{E}XY - \mathbf{E}X\mathbf{E}Y\,,\end{aligned}$$

d.h. insbesondere $\mathrm{Kov}(X,X) = \mathrm{Var}(X)$ und

$$X, Y \text{ stoch. unabh.} \quad \Rightarrow \quad \mathrm{Kov}(X,Y) = 0;\ \rho(X,Y) = 0\,.$$

(ii) $Y = X \ \Rightarrow\ \mathrm{Kov}(X,Y) = \mathrm{Var}X \ \Rightarrow\ \rho(X,Y) = 1\,.$

(iii) $Y = -X \ \Rightarrow\ \mathrm{Kov}(X,Y) = -\mathrm{Var}X \ \Rightarrow\ \rho(X,Y) = -1\,.$

(iv) Allgemein gilt:

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

Beweis: (für X, Y stetig)

$$|\mathrm{Kov}(X,Y)| = \Big| \iint (x - \mathbf{E}X)(y - \mathbf{E}Y) f_{XY}(x,y)\, dxdy \Big|$$

$$\overset{\text{Cauchy-Schwarz}}{\leq} \left(\iint (x - \mathbf{E}X)^2 f_{XY}(x,y)\, dxdy \right)^{1/2} \left(\iint (y - \mathbf{E}Y)^2 f_{XY}(x,y)\, dxdy \right)^{1/2}$$

$$= \left(\int (x - \mathbf{E}X)^2 f_X(x) dx \right)^{1/2} \left(\int (y - \mathbf{E}Y)^2 f_Y(y) dy \right)^{1/2}$$

$\Rightarrow$ Beh.

Beispiel 11.3 (Fortsetzung von Beispiel 8.7)

(i)

$$f_{X,Y}(x,y) = \begin{cases} 2x + 2y - 4xy & , \quad x, y \in [0,1] \\ 0 & , \quad \text{sonst} \end{cases}$$

$$f_X(x) = \mathrm{I}_{[0,1]}(x) \quad \Rightarrow \quad \mathbf{E}X = \frac{1}{2}, \quad \mathbf{E}X^2 = \int_0^1 x^2\, dx = \frac{1}{3}, \quad \mathrm{Var}\, X = \frac{1}{3} - \frac{1}{4} = \frac{1}{12};$$

$$f_Y(y) = \mathrm{I}_{[0,1]}(y) \quad \Rightarrow \quad \mathbf{E}Y = \frac{1}{2}, \quad \mathrm{Var}\, Y = \frac{1}{12};$$

$$\begin{aligned} \mathrm{Kov}(X,Y) &= \mathbf{E}XY - \mathbf{E}X\, \mathbf{E}Y \\ &= \int_0^1 \int_0^1 xy\,(2x + 2y - 4xy)\, dx\, dy - \frac{1}{4} \\ &= \ldots = \frac{2}{9} - \frac{1}{4} = -\frac{1}{36} \end{aligned}$$

$$\Rightarrow \quad \rho(X,Y) = -\frac{\frac{1}{36}}{\sqrt{\frac{1}{12} \cdot \frac{1}{12}}} = -\frac{1}{3}.$$

(ii)

$$f_{X,Y}(x,y) = \begin{cases} 2 - 2x - 2y + 4xy & , \quad x, y \in [0,1] \\ 0 & \text{sonst} \end{cases}$$

$$\mathbf{E}X = \mathbf{E}Y = \frac{1}{2}, \quad \mathrm{Var}X = \mathrm{Var}Y = \frac{1}{12},$$

$$\begin{aligned}\mathrm{Kov}(X,Y) &= \mathbf{E}XY - \mathbf{E}X\,\mathbf{E}Y \\ &= \int_0^1 \int_0^1 xy(2-2x-2y+4xy)\,dx\,dy - \frac{1}{4} \\ &= \ldots = \frac{5}{18} - \frac{1}{4} = \frac{1}{36}\end{aligned}$$

$$\Rightarrow \quad \rho(X,Y) = \frac{\frac{1}{36}}{\sqrt{\frac{1}{12}\cdot\frac{1}{12}}} = \frac{1}{3}.$$

(iii)

$$\begin{aligned}f_{XY}(x,y) &= f_X(x)\cdot f_Y(y) \\ &\Rightarrow X, Y \text{ unabhängig} \\ &\Rightarrow \mathrm{Kov}(X,Y) = 0 \\ &\Rightarrow \rho(X,Y) = 0.\end{aligned}$$

Beachte die Plots der Dichten in Kapitel 8. Wir vertiefen den Zusammenhang zwischen $\rho(X,Y)$ und einem linearen Zusammenhang zwischen X und Y in Satz 11.6 unten. □

Bemerkung: Das wichtigste Beispiel (auch für den Rest dieses Kapitels) ist wiederum die multivariate Normalverteilung, die ausführlich im folgenden Kapitel behandelt wird. Man beachte schon einmal die Plots für verschiedene ρ in Kapitel 12.

Proposition 11.4 $\mathrm{Kov}(X,Y)$ $\left[\textit{nicht } \rho(X,Y)\textit{!}\right]$ *ist bilinear, d.h.*

$$\mathrm{Kov}(a + \sum_{i=1}^{n} b_i X_i\ ,\ c + \sum_{j=1}^{m} d_j Y_j) = \sum_{i,j} b_i\, d_j\, \mathrm{Kov}(X_i, Y_j).$$

Beweis. Folgende Beziehungen sind leicht nachzurechnen:

(i) $\mathrm{Kov}(X+Y, Z) = \mathrm{Kov}(X,Z) + \mathrm{Kov}(Y,Z)$;
(ii) $\mathrm{Kov}(aX, Z) = a\,\mathrm{Kov}(X,Z)$;
(iii) $\mathrm{Kov}(1, Z) = 0$.

Analoges beweist man für die 2. Komponente $\Rightarrow$ Beh. $\square$

Korollar 11.5 *Seien $X_1, \ldots, X_n$ gemeinsam verteilt. Dann gilt*

(i) $\mathrm{Var}\,(\sum_{i=1}^n X_i) = \sum_{i,j=1}^n \mathrm{Kov}(X_i, X_j)$;
(ii) $X_1, \ldots X_n$ *stoch. unabh.* $\Rightarrow$ $\mathrm{Var}\,(\sum_{i=1}^n X_i) = \sum_{i=1}^n \mathrm{Var}X_i$ *(vgl. Satz 8.18).*

Satz 11.6 *Seien $\mathbf{E}X^2, \mathbf{E}Y^2 < \infty$. Es gilt $|\rho| \leq 1$ und $\rho = \pm 1$ genau dann, wenn $a, b \in \mathbb{R}, a \neq 0$ existieren mit $\mathbf{P}(Y = aX + b) = 1$, wobei*

$$\begin{array}{ll} a > 0 & , \quad \textit{falls } \rho = +1 \\ a < 0 & , \quad \textit{falls } \rho = -1 \end{array} .$$

Beweis. $|\rho| \leq 1$ wurde bereits bewiesen.
Sei $\underline{\rho = 1}$.

$$\mathrm{Var}\big(\frac{X}{\sqrt{\mathrm{Var}X}} - \frac{Y}{\sqrt{\mathrm{Var}Y}}\big) = \mathrm{Var}\big(\frac{X}{\sqrt{\mathrm{Var}X}}\big) + \mathrm{Var}\big(\frac{Y}{\sqrt{\mathrm{Var}Y}}\big) - 2\,\mathrm{Kov}\big(\frac{X}{\sqrt{\mathrm{Var}X}}, \frac{Y}{\sqrt{\mathrm{Var}Y}}\big)$$
$$= 1 + 1 - 2\frac{\mathrm{Kov}(X,Y)}{\sqrt{\mathrm{Var}X\,\mathrm{Var}Y}} = 2 - 2\rho(X,Y) = 0$$

$$\overset{Korollar 7.16(ii)}{\Leftrightarrow} \quad \mathbf{P}\left(\frac{X}{\sqrt{\mathrm{Var}X}} - \frac{Y}{\sqrt{\mathrm{Var}Y}} = c\right) = 1$$

$$\Rightarrow \quad \mathbf{P}(Y = aX + b) = 1 \quad \text{mit } a = \frac{\sqrt{\mathrm{Var}Y}}{\sqrt{\mathrm{Var}X}} > 0.$$

Sei $\underline{\rho = -1}$. Analoge Rechnung wie eben mit $\mathrm{Var}(\frac{X}{\sqrt{\mathrm{Var}X}} + \frac{Y}{\sqrt{\mathrm{Var}Y}}) = \ldots = 0$.
Die andere Richtung rechnet man leicht nach. $\square$

Definition/Bemerkung 11.7 Die Matrix

$$\Sigma = \Sigma(X) := \big(\mathrm{Kov}(X_i, X_j)\big)_{i,j=1,\ldots,n}$$

heißt <u>Kovarianzmatrix</u> von $X = (X_1, \ldots, X_n)'$. Sei $\gamma = (\gamma_1, \ldots, \gamma_n)'$. Dann gilt

$$\mathrm{Var}(\gamma' X) = \gamma'\,\Sigma(X)\,\gamma \quad \text{(folgt aus Korrolar 11.5).}$$

Die Matrix Σ ist symmetrisch und nicht-negativ definit.

12 Die multivariate Normalverteilung und die Hauptkomponentenanalyse

Die multivariate Normalverteilung wird eingeführt und die Hauptachsentransformation mehrdimensionaler Verteilungen diskutiert. Als Anwendung wird die Hauptkomponentenanalyse (principal component analysis) kurz behandelt. Sie ist eine der wichtigsten Analyse-Methoden der multivariaten Statistik. Mathematisch spielt in diesem Kapitel die Hauptachsentransformation aus der Linearen Algebra eine wichtige Rolle (siehe z.B. Koecher, Lineare Algebra und analytische Geometrie, Kapitel 6.2)

Definition 12.1 (Die multivariate Normalverteilung) Sei $\underline{\mu} \in \mathbb{R}^d$, Σ eine positiv definite, symmetrische $d \times d$ Matrix und

$$f(x_1, \ldots, x_d) = \frac{1}{(2\pi)^{\frac{d}{2}} |\Sigma|^{\frac{1}{2}}} \exp\big\{-\frac{1}{2} (\underline{x} - \underline{\mu})' \Sigma^{-1} (\underline{x} - \underline{\mu})\big\},$$

wobei $\underline{x} = (x_1, \ldots, x_d)'$. Dann heißt eine stetige Verteilung auf $\mathbb{R}^d$ mit Dichte f multivariate Normalverteilung $\mathcal{N}(\underline{\mu}, \Sigma)$. Ist $\underline{X} = (X_1, \ldots, X_d)'$ Zufallsvektor mit induzierter Verteilung $\mathbf{P}^X = \mathcal{N}(\underline{\mu}, \Sigma)$ so schreibt man $\underline{X} \sim \mathcal{N}(\underline{\mu}, \Sigma)$.

Lemma 12.2 *Es gilt*

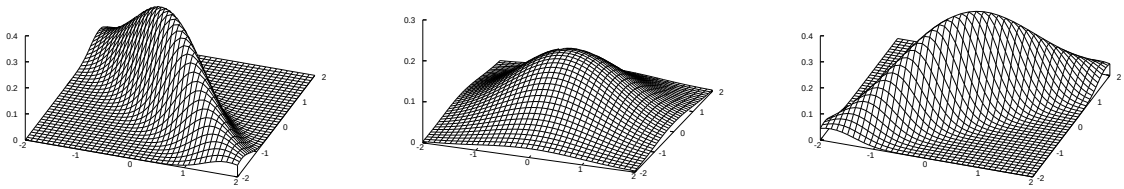
(i) $\int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} f(x_1, \ldots, x_d)\, dx_1 \cdots dx_d = 1$, *d.h. f ist eine Wahrscheinlichkeitsdichte.*

(ii) Sei $\underline{X} = (X_1, \ldots, X_d)'$ *stetig verteilt mit obiger Dichte* $f(x_1, \ldots, x_d)$. *Dann gilt* $\mathbf{E}X_i = \mu_i$ *und* $\mathrm{Kov}(X_i, X_j) = \Sigma_{ij}$, *d.h. insbesondere* $\mathrm{Var}(X_i) = \Sigma_{ii}$.

(iii) Sei A eine $(k \times d)$-Matrix mit Rang $A = k$ und $\underline{b} = (b_1, \ldots, b_k)'$, sowie $\underline{X} \sim \mathcal{N}(\underline{\mu}, \Sigma)$. Dann gilt $A\underline{X} + \underline{b} \sim \mathcal{N}(A\underline{\mu} + \underline{b}, A\Sigma A')$.

Beweis. Übungsaufgabe. □

Notation: Im Rest dieses Kapitels schreiben wir häufig X, μ, b, x anstelle von $\underline{X}, \underline{\mu}, \underline{b}, \underline{x}$.



$\rho = -0,9$ $\rho = 0,0$ $\rho = 0,9$

Figur: Dichte der 2-dim. Normalverteilung für verschiedene Korrelationskoeffizienten

Bemerkung 12.3 (Hauptachsentransformation)

$\underline{d=1}$: Dieser Fall ist identisch mit den bisherigen Ergebnissen.

$\underline{d=2}$: Wie sehen die Höhenlinien der Funktion f aus? Untersuche dafür die Kurven mit konstanter Dichte

$$\frac{1}{(2\pi)^{\frac{d}{2}}\,|\Sigma|^{\frac{1}{2}}}\exp\{-\frac{1}{2}\,c\},$$

also die Kurven mit $(\underline{x}-\underline{\mu})'\,\Sigma^{-1}\,(\underline{x}-\underline{\mu})=c$ in $\mathbb{R}^2$.

Σ symmetrisch und positiv definit

$\Rightarrow$ $\exists$ Orthonormalmatrix P (d.h. $PP'=P'P=I_d$) und Diagonalmatrix Λ mit

$$\Sigma=P\Lambda P',\qquad P=(P_1,P_2),\qquad \Lambda=\begin{pmatrix}\lambda_1 & 0\\ 0 & \lambda_2\end{pmatrix}.$$

$$\Rightarrow\quad P'=\begin{pmatrix}\cos\varphi & \sin\varphi\\ -\sin\varphi & \cos\varphi\end{pmatrix}\quad\text{oder}\quad P'=\begin{pmatrix}1 & 0\\ 0 & -1\end{pmatrix}\begin{pmatrix}\cos\varphi & \sin\varphi\\ -\sin\varphi & \cos\varphi\end{pmatrix}.$$

mit einem $\varphi\in[0,2\pi)$ (vgl. z.B. Koecher, Lineare Algebra und analytische Geometrie, Kapitel 4.2). Wegen

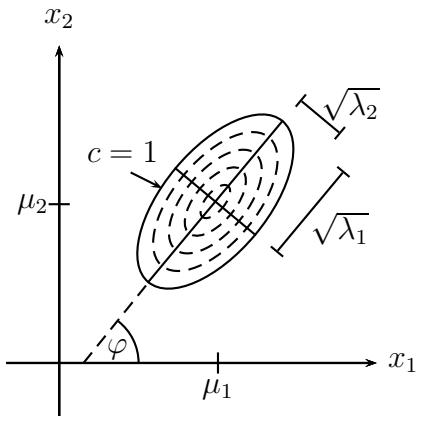
$$\begin{pmatrix}1 & 0\\ 0 & -1\end{pmatrix}\Lambda\begin{pmatrix}1 & 0\\ 0 & -1\end{pmatrix}=\Lambda$$

haben diese beiden Fälle dieselbe Dichte und wir können uns Œ auf den ersten Fall beschränken. Es gilt $\Sigma^{-1}=(P')^{-1}\Lambda^{-1}P^{-1}=P\Lambda^{-1}P'$.

$$\begin{aligned}\Rightarrow f(x_1,x_2)&=\frac{1}{2\pi\,|\Lambda|^{\frac{1}{2}}}\exp\big\{-\frac{1}{2}\,(\underline{x}-\underline{\mu})\,P\Lambda^{-1}P'\,(\underline{x}-\underline{\mu})\big\}\\ &=\frac{1}{2\pi\,|\Lambda|^{\frac{1}{2}}}\exp\{-\frac{1}{2}\,\underline{y}'\Lambda^{-1}\underline{y}\}\end{aligned}$$

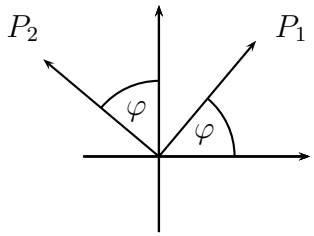
mit $\underline{y}=P'(\underline{x}-\underline{\mu})$ (Verschiebung + Drehung um den Winkel $-\varphi$).
Folgendes ist anzumerken:

- Wegen $\underline{y}'\Lambda^{-1}\underline{y} = \frac{y_1^2}{\lambda_1} + \frac{y_2^2}{\lambda_2}$ sind die Höhenlinien von f Ellipsen, wobei die Hauptachsen um den Winkel φ zur x_1-, bzw. x_2-Achse gedreht sind.



  [Verbal: $\underline{x} - \underline{\mu}$ wird um den Winkel $-\varphi$ gedreht und muss dann die elliptische Gleichung erfüllen.]

- Wegen $(\underline{x} - \underline{\mu}) = PP'(\underline{x} - \underline{\mu}) = P\,\underline{y} = y_1P_1 + y_2P_2$ sind (y_1, y_2) die Koordinaten von $(\underline{x} - \underline{\mu})$ in dem Koordinatensystem mit Achsen P_1 und P_2:

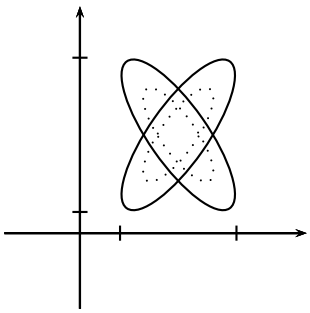


- Wir werden später sehen, dass die transformierten Zufallsvariablen $Y_j := P_j'(X - \mu)$ ebenfalls normal verteilt und stochastisch unabhängig sind [heuristisch klar, wenn man sich die Grafik ansieht].

Die Vektoren P_1 und P_2 heißen Hauptachsen von Σ, die Transformation in das y-Koordinatensystem Hauptachsentransformation und die Zufallsvariablen $Y_j := P_j'(X - \mu)$

Hauptkomponenten von X. Völlig analog erhält man die Hauptachsen $P_1, \ldots, P_n$ für d-dimensionale Normalverteilungen - diese sind dann die Hauptachsen der d-dimensionalen Ellipsoide.

Der eigentliche Vorteil der Projektion auf die Hauptkomponenten ergibt sich erst im Fall $d \geq 3$: Man gewinnt eine bessere Vorstellung von der Verteilung aus einem Plot der Projektion auf die Hauptkomponenten [die d-dimensionale Verteilungen kann man grafisch nicht darstellen].



In diesem Beispiel wird die Struktur der Verteilung durch einen Plot der Hauptkomponenten deutlich, aber nicht durch einen Plot, der x_1 und x_2 - Koordinaten (analog für Daten).

Wir merken noch an, dass man für obige Zerlegung von Σ nicht die Normalverteilungsannahme benötigt (allerdings geht die elliptische Form der Höhenlinien dann idR verloren).

Bemerkung 12.4 (Hauptkomponentenanalyse)
Wir wollen nun eine einzelne (multivariate) Beobachtung auf die Hauptkomponenten projizieren [d.h. in dem y-Koordinatensystem darstellen].

Sei $X = (X_1, \ldots, X_d)'$ ZVA mit Kovarianzmatrix Σ (idR ist Σ unbekannt und muss geschätzt werden - s. unten)

$$\Rightarrow \quad \Sigma \text{ symmetrisch und positiv semi-definit (idR sogar positiv definit)}$$

$$\Rightarrow \quad \Sigma = P\Lambda P' \text{ mit } PP' = P'P = I_d \text{ und } \Lambda = \begin{pmatrix} \lambda_1 & & 0 \\ & \ddots & \\ 0 & & \lambda_d \end{pmatrix}.$$

Seien Œ $\lambda_1 \geq \ldots \geq \lambda_d \geq 0$ und $P = (P_1, \ldots, P_d)$ mit $P_i \in \mathbb{R}^d$. Es gilt

$$X = PP'X = \sum_{j=1}^{d} (P_j' X) P_j = \sum_{j=1}^{d} Y_j P_j$$

mit den neuen Koordinaten $Y_j := P_j' X$ [man kann das "$-\mu$" weglassen]. Für diese gilt

$$\begin{aligned} \mathrm{Var}(Y_j) &= \mathrm{Var}(P_j' X) = P_j' \Sigma P_j = P_j' P \Lambda P' P_j = \lambda_j, \\ \mathrm{Kov}(Y_i, Y_j) &= \mathrm{Kov}(P_i' X, P_j' X) = P_i' \Sigma P_j = 0 \quad \text{für } i \neq j, \end{aligned}$$

d.h. die Y-Koordinaten sind unkorreliert (bei Normalverteilungen sogar unabhängig - Beweis später) und bzgl. der Streuung geordnet. Die Variable Y_1 enthält die meiste Information über X im folgenden Sinne: Gesucht werde diejenige Linearkombination $Y = \gamma_0 + \gamma' X$ und diejenigen $\alpha_j, \beta_j \in \mathbb{R}$, für die

$$\sum_{j=1}^{d} \mathbf{E}\big[X_j - (\alpha_j + \beta_j Y)\big]^2 = (*) \tag{7}$$

minimal wird.

Satz 12.5 *Der mittlere Prädiktionsfehler (7) wird für* $\gamma_0 = 0$, $\beta_j = \frac{\mathrm{Kov}(X_j, Y)}{\mathrm{Var}(Y)}$, $\alpha_j = \mathbf{E}X_j - \beta_j \, \mathbf{E}Y$ *und* $Y = \frac{1}{\sqrt{\lambda_1}} P_1' X = \frac{1}{\sqrt{\lambda_1}} Y_1$ *minimal (mit* P_1, λ_1 *wie oben).*

Bemerkung. Ein analoger Satz gilt für die Approximation von $X_1, \dots, X_d$ durch k Variable, nämlich durch $\frac{Y_1}{\sqrt{\lambda_1}}, \dots, \frac{Y_k}{\sqrt{\lambda_k}}$ $\big($genauer durch $\mathrm{Lin}(Y_1, \dots, Y_k)\big)$.

Beweis.

$$(*) \;= \sum_j \mathbf{E}\Big[(X_j - \mathbf{E}X_j) - \alpha_j^* - \beta_j(Y - \mathbf{E}Y)\Big]^2 \qquad \text{mit } \alpha_j^* = \alpha_j - \mathbf{E}X_j + \beta_j\,\mathbf{E}Y$$
$$= \sum_j \Big[\mathrm{Var}(X_j) - 2\,\beta_j\,\mathrm{Kov}(X_j, Y) + \alpha_j^{*2} + \beta_j^2\,\mathrm{Var}(Y)\Big], \quad \text{d.h. } \alpha^*_{j,\min} = 0.$$

Ferner gilt

$$\frac{\partial}{\partial \beta_j}\Big[\dots\Big] = -2\,\mathrm{Kov}(X_j, Y) + 2\,\beta_j\,\mathrm{Var}(Y) = 0 \;\Leftrightarrow\; \beta_j = \frac{\mathrm{Kov}(X_j, Y)}{\mathrm{Var}(Y)}$$

und damit

$$(*)_{\min} = \sum_j \Big[\mathrm{Var}(X_j) - \frac{\mathrm{Kov}(X_j, Y)^2}{\mathrm{Var}(Y)}\Big] = \sum_j \Big[\mathrm{Var}(X_j) - \frac{(\gamma'\Sigma)_j{}^2}{\gamma'\Sigma\gamma}\Big]$$
$$= \sum_j \mathrm{Var}(X_j) - \frac{\gamma'\Sigma^2\gamma}{\gamma'\Sigma\gamma}.$$

Maximiere also $\frac{\gamma'\Sigma^2\gamma}{\gamma'\Sigma\gamma}$ bzgl. γ mit $\Sigma = P\Lambda P'$ wie oben.

$$\frac{\gamma'\Sigma^2\gamma}{\gamma'\Sigma\gamma} = \frac{\gamma' P\Lambda^2 P'\gamma}{\gamma' P\Lambda P'\gamma} = \frac{\delta'\Lambda\delta}{\delta'\delta} \qquad \text{mit } \delta = \Lambda^{\frac{1}{2}} P'\gamma$$

wird maximal für $\delta = (1, 0, \dots, 0)'$, d.h. für $\gamma = P\Lambda^{-\frac{1}{2}}\delta = \frac{1}{\sqrt{\lambda_1}}\,P_1$. □

Beispiel 12.6 (Anwendung in der Datenanalyse: Hauptkomponentenanalyse)
Seien nun n stochastisch unabhängige Realisierungen $\underline{X}_1, \dots, \underline{X}_n$ der Zufallsvariablen $\underline{X} = (X_1, \dots, X_d)'$ gegeben. Wir wollen alle Daten auf die Hauptachsen projizieren, dh. die Hauptkomponenten berechnen [genauer: schätzen (da Σ unbekannt ist)]. Vorgehen:

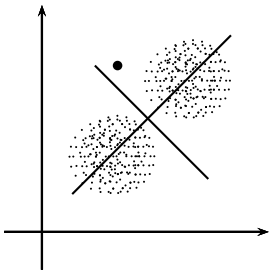
1. Schätze die unbekannte Kovarianzmatrix Σ durch

$$\hat{\Sigma}_{ij} := \frac{1}{n-1}\sum_{\ell=1}^{n}(X_{\ell i} - X_{\cdot i})(X_{\ell j} - X_{\cdot j}) \quad \text{wobei } X_{\cdot i} = \frac{1}{n}\sum_{\ell=1}^{n} X_{\ell i}\,.$$

Übungsaufgabe: Zeige, dass $\mathbf{E}\hat{\Sigma}_{ij} = \Sigma_{ij}$.

2. Berechne die λ_j und P_j sowie die Koordinaten $Y_{\ell j} = P_j' \underline{X}_\ell$, bzw. $\frac{1}{\sqrt{\lambda_j}} P_j' \underline{X}_\ell$
3. Plotte die $\underline{Y}_\ell$, $\ell = 1, \ldots, n$.

Beispiel:



2 Cluster, 1 Ausreißer: Bei der Projektion auf die Hauptachsen wird die Struktur deutlich.

Es bleibt, die Eigenschaften, dieses statistischen Verfahrens zu untersuchen (und weiterer Verfahren, die darauf aufbauen, wie z.B. der Clusteranalyse oder der Ausreißer - Erkennung). Hierzu reicht an dieser Stelle aber die Zeit nicht aus. □

Bezeichnung: Sei nun $a \in \mathbb{R}^d$ mit $\|a\| = 1$. Dann heißt $x \mapsto a'x$ standardisierte Linearform (SLF).

Proposition 12.7 *Für jede SLF gilt*

$$\mathrm{Var}(a'X) \leq \lambda_1.$$

Beweis. Sei $a = \sum_{i=1}^d c_i P_i$ mit $1 = \|a\|^2 = \sum_{i=1}^d c_i^2$. Dann gilt

$$\begin{aligned}
\mathrm{Var}(a'X) = a'\Sigma a &= a'\, P\, \Lambda\, P'\, a \\
&= \sum_{j=1}^d \lambda_j (P_j' a)^2 = \sum_{j=1}^d \lambda_j\, c_j^2 \\
&\leq \max_j\{\lambda_j\} \sum_{j=1}^d c_j^2 = \lambda_1.
\end{aligned}$$

□

Damit ist die Richtung $Y_1 = P_1'X$ die Richtung mit der größten Streuung. Analog zeigt man, dass $\mathrm{Var}(a'X) \geq \lambda_d$ für alle SLF gilt, d.h. $Y_d = P_d'X$ ist die Richtung mit der kleinsten Streuung. Auch die anderen Hauptkomponenten haben eine maximale Varianzeigenschaft:

Proposition 12.8 *Für jede SLF mit* $\mathrm{Kov}(a'X, Y_j) = 0 \quad (j = 1, \ldots, k)$ *gilt*

$$\mathrm{Var}(a'X) \leq \lambda_{k+1},$$

d.h. die Richtung $Y_{k+1} = P_{k+1}'X$ *hat maximale Varianz unter allen Richtungen, die senkrecht auf den* $Y_1, \ldots, Y_k$ *stehen.*

Beweis. Es gilt mit $a = \sum_{i=1}^{d} c_i P_i$

$$\begin{aligned} \mathrm{Kov}(a'X, Y_j) = a'\Sigma P_j = \lambda_j\, a'P_j = c_j\, \lambda_j \stackrel{!}{=} 0 \quad (j = 1, \ldots, k) \\ \Rightarrow \mathrm{Var}(a'X) = \sum_{j=1}^{d} \lambda_j\, c_j^2 = \sum_{j=k+1}^{d} \lambda_j\, c_j^2 \leq \lambda_{k+1} \sum_{j=1}^{d} c_j^2 = \lambda_{k+1}. \end{aligned}$$

□

13 Verteilungseigenschaften von Mittelwert und Varianz bei Normalverteilungen und der t-Test

In diesem Kapitel wird gezeigt, dass der Mittelwert und die empirische Varianz bei normalverteilten Beobachtungen stochastisch unabhängig sind. Daraus folgt insbesondere auch die Verteilung der Teststatistik beim t-Test. Die Optimalität des t-Tests wird erst in einer späteren Vorlesung hergeleitet.

Bemerkung 13.1 (Orthogonalzerlegung)
Seien $X_1, \dots, X_n$ mit $X_i \overset{\text{iid}}{\sim} \mathcal{N}(\mu, \sigma^2)$, d.h. $\underline{X} = (X_1, \dots, X_n)' \sim \mathcal{N}(\underline{\mu}, \sigma^2 I)$.
Betrachte den Mittelwert $\overline{X}_n := \frac{1}{n}\sum_{i=1}^n X_i$ und die empirische Varianz

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

Es gilt

$$\overline{X}_n = \frac{1}{n}(1, \dots, 1)\,\underline{X} \;\sim\; \mathcal{N}\left(\mu, \frac{\sigma^2}{n^2}\,(1, \dots, 1)\, I \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix}\right) = \mathcal{N}\Big(\mu, \frac{\sigma^2}{n}\Big),$$

d.h. insbesondere $\mathbf{E}\overline{X}_n = \mu,\ \mathrm{Var}(\overline{X}_n) = \frac{\sigma^2}{n}$ und

$$\begin{aligned}\mathbf{E}S^2 &= \frac{n}{n-1}\,\mathbf{E}\,\frac{1}{n}\sum_{i=1}^n \big((X_i-\mu)^2 + 2(X_i-\mu)(\mu-\overline{X}_n) + (\mu-\overline{X}_n)^2\big) \\ &= \frac{n}{n-1}\mathbf{E}\left[\frac{1}{n}\sum_{i=1}^n (X_i-\mu)^2 - (\mu-\overline{X}_n)^2\right] = \frac{n}{n-1}\left(\sigma^2 - \frac{\sigma^2}{n}\right) = \sigma^2.\end{aligned}$$

Berechne nun die Verteilung von S^2 und $\frac{\overline{X}_n - \mu}{S/\sqrt{n}}$.

$\left[\frac{\overline{X}_n-\mu}{S/\sqrt{n}}\right.$ kommt als Testgröße eines Tests für den Erwartungswert in Frage falls σ^2 unbekannt ist$\left.\right]$

Trick: Sei $P_n = (\frac{1}{\sqrt{n}}, \dots, \frac{1}{\sqrt{n}})'$. Ergänze P_n zu einer Orthonormalbasis (ONB) $\{P_1, \dots, P_n\}$ von $\mathbb{R}^n$. Sei $P = (P_1, \dots, P_n)$. Es gilt $PP' = P'P = I$.

Lemma 12.2(iii): $P'(\underline{X}-\underline{\mu}) \sim \mathcal{N}(0, \sigma^2 P'P) = \mathcal{N}(0, \sigma^2 I)$

$\Rightarrow$ die einzelnen $P_i'(\underline{X}-\underline{\mu})$ sind stoch. unabhängig.

Es gilt nun $\overline{X}_n - \mu = \frac{1}{\sqrt{n}} P_n' (\underline{X}-\underline{\mu})$ und

$$\begin{aligned} S^2 &= \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X}_n)^2 = \frac{n}{n-1}\left[\frac{1}{n}\sum_{i=1}^{n}(X_i-\mu)^2 - (\overline{X}_n-\mu)^2\right] \\ &= \frac{n}{n-1}\left[\frac{1}{n}(\underline{X}-\underline{\mu})'(\underline{X}-\underline{\mu}) - (\overline{X}_n-\mu)^2\right] \\ &= \frac{n}{n-1}\left[\frac{1}{n}(\underline{X}-\underline{\mu})'PP'(\underline{X}-\underline{\mu}) - \frac{1}{n}\left(P_n'(\underline{X}-\underline{\mu})\right)^2\right] \\ &= \frac{1}{n-1}\sum_{i=1}^{n-1}\left(P_i'(\underline{X}-\underline{\mu})\right)^2. \end{aligned}$$

Aus dieser Darstellung von Mittelwert und empirischer Varianz werden wir unten auf die Verteilung von $\frac{\overline{X}_n-\mu}{S/\sqrt{n}}$ schließen. Zunächst halten wir fest:

Proposition 13.2 *$\overline{X}_n$ und S^2 sind stochastisch unabhängig.*

Definition 13.3 (χ^2- und t-Verteilung)

(i) *Sei $Y \sim \mathcal{N}(0,1)$. Dann heißt die Verteilung von $Z = Y^2$ Chi-Quadrat-Verteilung mit einem Freiheitsgrad. Man schreibt $Z \sim \chi_1^2$.*

(ii) *Seien $Z_1, \ldots, Z_n \overset{iid}{\sim} \chi_1^2$. Dann heißt die Verteilung von $Z = \sum_{i=1}^n Z_i$ Chi-Quadrat-Verteilung mit n Freiheitsgraden. Man schreibt $Z \sim \chi_n^2$.*

(iii) *Seien $Y \sim \mathcal{N}(0,1), Z \sim \chi_n^2$ stoch. unabhängig. Dann heißt die Verteilung von $T := \frac{Y}{\sqrt{\frac{Z}{n}}}$ t-Verteilung (Student-Verteilung) mit n Freiheitsgraden. Man schreibt $T \sim t_n$.*

Satz 13.4 *Seien $X_1, \ldots X_n \overset{iid}{\sim} \mathcal{N}(\mu, \sigma^2)$. Dann gilt*

(i) $(n-1)\frac{S^2}{\sigma^2} \sim \chi_{n-1}^2$ *und*

(ii) $\frac{\overline{X}_n-\mu}{S/\sqrt{n}} \sim t_{n-1}$.

Beweis. Es gilt

$$\underbrace{(n-1)\,\frac{S^2}{\sigma^2}}_{=:\,(a)} = \sum_{i=1}^{n-1} \Big(\frac{1}{\sigma} P_i'(\underline{X} - \underline{\mu})\Big)^2 \sim \chi^2_{n-1}$$

und

$$\overline{X}_n - \mu \sim \mathcal{N}\big(0, \frac{\sigma^2}{n}\big) \quad \Rightarrow \quad \underbrace{\frac{\overline{X}_n - \mu}{\sigma/\sqrt{n}}}_{=:\,(b)} \sim \mathcal{N}(0,1).$$

Da (a) und (b) stochastisch unabhängig sind, folgt

$$\frac{\overline{X}_n - \mu}{\frac{\sigma}{\sqrt{n}}} \Big/ \sqrt{\frac{S^2}{\sigma^2}} = \frac{\overline{X}_n - \mu}{\frac{S}{\sqrt{n}}} \sim t_{n-1} \qquad (*).$$

□

Proposition 13.5 *Sei $Y \sim \chi^2_n$. Dann gilt $\mathbf{E}Y = n$ und $\mathrm{Var}Y = 2n$ und damit*

$$\mathrm{Var}(S^2) = \frac{2\sigma^4}{n-1}.$$

Beweis. Seien $Z_1, \ldots, Z_n \overset{\text{iid}}{\sim} \mathcal{N}(0,1)$. Es folgt, dass

$$\mathbf{E}\,Y = \mathbf{E}\,\big(\sum_{i=1}^{n} Z_i^2\big) = n\,\mathbf{E}Z_1^2 = n$$

und

$$\mathrm{Var}(Y) = \sum_{i=1}^{n} \mathrm{Var}(Z_i^2) = n\,(\mathbf{E}Z_i^4 - 1) = 2n$$

$$\left[\begin{array}{ll} \int_{-\infty}^{\infty} z^4 \frac{1}{\sqrt{2\pi}} \exp(-\frac{1}{2}z^2)\,dz & = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} z^3 z \exp(-\frac{1}{2}z^2)\,dz \\ & = -\frac{1}{\sqrt{2\pi}}\, z^3\, \exp(-\frac{1}{2}z^2)\Big|_{-\infty}^{\infty} + \frac{3}{\sqrt{2\pi}} \int_{-\infty}^{\infty} z^2\, \exp(-\frac{1}{2}z^2)\,dz \\ & = 0 + 3\,\mathbf{E}Z^2 = 3 \end{array}\right]$$

$$\Rightarrow \mathrm{Var}(S^2) = \frac{\sigma^4}{(n-1)^2}\, 2(n-1) = \frac{2\sigma^4}{n-1}.$$

□

Sei nun $\Gamma(x) := \int_0^\infty \mathrm{e}^{-t}\, t^{x-1}\, dt$ die Gammafunktion.

Proposition 13.6 *(i) Die χ^2_n-Verteilung hat die Dichte*

$$f(x) = \frac{1}{2^{\frac{n}{2}}\Gamma(\frac{n}{2})}\, e^{-\frac{x}{2}}\, x^{\frac{n}{2}-1} \qquad \textit{für } x \geq 0 \quad \big(f(x) = 0 \textit{ für } x < 0\big).$$

(ii) Die t_n-Verteilung hat die Dichte

$$f(x) = \frac{\Gamma(\frac{n+1}{2})}{\sqrt{n\pi}\,\Gamma(\frac{n}{2})} \Big(1 + \frac{x^2}{n}\Big)^{-(n+1)/2}, \quad x \in \mathbb{R}.$$

Beweis.

(i) Sei zunächst $n = 1$, d.h. $X = Y^2$ mit $Y \sim \mathcal{N}(0,1)$. Dann gilt

$$\begin{aligned} F_X(x) = \mathbf{P}(X \leq x) = \mathbf{P}(-\sqrt{x} \leq Y \leq \sqrt{x}) &= \int_{-\sqrt{x}}^{\sqrt{x}} \frac{1}{\sqrt{2\pi}} \exp(-\frac{1}{2}z^2)\, dz \\ &= 2\int_0^{\sqrt{x}} \frac{1}{\sqrt{2\pi}} \exp(-\frac{1}{2}z^2)\, dz \\ \Rightarrow \quad f_X(x) = \frac{d}{dx}F_X(x) = \frac{2}{\sqrt{2\pi}} \exp(-\frac{1}{2}x)\, \frac{1}{2}\, x^{-\frac{1}{2}} &= \frac{1}{\sqrt{2\pi}}\, e^{-\frac{x}{2}}\, x^{-\frac{1}{2}} \end{aligned}$$

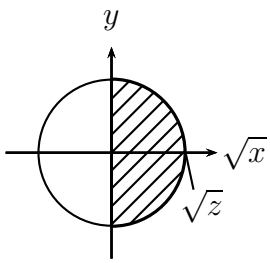
Wegen $\Gamma(\frac{1}{2}) = \sqrt{\pi}$ folgt Behauptung (i) für $n = 1$.

Induktion $n \mapsto n+1$: Seien nun $X \sim \chi^2_n$ und $Y \sim \mathcal{N}(0,1)$, X und Y stochastisch unabhängig. Per Definition gilt $Z = X + Y^2 \sim \chi^2_{n+1}$.

Sei nun $A_z = \big\{(x,y) \mid x + y^2 \leq z\big\}$.

$$\begin{aligned} F_Z(z) = \mathbf{P}(Z \leq z) = \mathbf{P}(X + Y^2 \leq z) &= \iint\limits_{A_z} f_{X,Y}(x,y)\, dx\, dy \\ &= \iint\limits_{A_z} c\, e^{-x/2}\, x^{\frac{n}{2}-1}\, e^{-\frac{y^2}{2}}\, dy\, dx \qquad \text{mit } c = \frac{1}{2^{\frac{n}{2}}\Gamma(\frac{n}{2})\sqrt{2\pi}} \qquad (8) \\ &= \int_{-\pi/2}^{\pi/2} \int_0^{\sqrt{z}} 2\, c\, e^{-r^2/2}\, r^n\, (\cos\varphi)^{n-1}\, dr\, d\varphi \end{aligned}$$

$$\left[\begin{array}{l} \text{Polarkoordinaten für } \sqrt{x} \text{ und } y \text{ (wegen } x+y^2 \le z): \\ \begin{array}{ll} y = r\sin\varphi & \\ x = r^2\cos^2\varphi & 0 < r < \sqrt{z},\ -\frac{\pi}{2} < \varphi < \frac{\pi}{2} \end{array} \\ \\ \Rightarrow \begin{pmatrix} \frac{\partial y}{\partial r} & \frac{\partial y}{\partial \varphi} \\ \frac{\partial x}{\partial r} & \frac{\partial x}{\partial \varphi} \end{pmatrix} = \begin{pmatrix} \sin\varphi & r\cos\varphi \\ 2r\cos^2\varphi & -2r^2\cos\varphi\,\sin\varphi \end{pmatrix} =: M \\ \\ \Rightarrow |M| = |2r^2\sin^2\varphi\,\cos\varphi + 2r^2\cos^3\varphi| = 2r^2\cos\varphi \end{array}\right]$$



Integrationsbereich Polarkoordinaten

$$\Rightarrow \frac{d}{dz}F_Z(z) = c' e^{-\frac{z}{2}} z^{\frac{n}{2}-\frac{1}{2}} = c' e^{-\frac{z}{2}} z^{\frac{n+1}{2}-1}.$$

Der Wert der Nominierungskonstanten folgt aus

$$\int_0^\infty z^{k-1} e^{-\frac{z}{2}}\, dz = 2^k \int_0^\infty z^{k-1} e^{-z} dz =: 2^k\,\Gamma(k).$$

(ii) Seien nun $X \sim \chi^2_n$ und $Y \sim \mathcal{N}(0,1)$. Per Definition gilt $W := \frac{Y}{\sqrt{X/n}} \sim t_n$.
Sei $A_w = \{(x,y) \mid \frac{y}{\sqrt{x/n}} \le w\}$.

$$\Rightarrow \quad F_W(w) \;=\; \mathbf{P}(W \le w) = \mathbf{P}\big(\frac{Y}{\sqrt{X/n}} \le w\big) = \int_{A_w} f_{X,Y}(x,y)\, dx\, dy$$

Wir führen die gleiche Transformation auf Polarkoordinaten durch wie oben. Wegen $y = r\sin\varphi$ und $x = r^2\cos^2\varphi$ gilt in Polarkoordinaten $A_w = \{(r,\varphi) \mid 0 < r < \infty, \varphi \in [-\frac{\pi}{2}, \frac{\pi}{2}], \tan\varphi \le \frac{w}{\sqrt{n}}\}$, d.h. man erhält

$$\begin{aligned} F_W(w) \;&=\; \int\limits_{\{\varphi\in[-\frac{\pi}{2},\frac{\pi}{2}]\,|\,\tan\varphi\le\frac{w}{\sqrt{n}}\}} \int_0^\infty 2\,c\,e^{-r^2/2}\, r^n (\cos\varphi)^{n-1}\, dr\, d\varphi \\ &=\; c' \int\limits_{-\pi/2}^{\arctan\frac{w}{\sqrt{n}}} (\cos\varphi)^{n-1}\, d\varphi = c' \int\limits_{-\pi/2}^{\arctan\frac{w}{\sqrt{n}}} \Big(\frac{1}{1+\tan^2\varphi}\Big)^{\frac{n-1}{2}}\, d\varphi \end{aligned}$$

da $\cos\varphi = \frac{1}{\sqrt{1+\tan^2\varphi}}$ für $\varphi \in [-\frac{\pi}{2}, \frac{\pi}{2}]$. Wegen $\frac{d}{dx}\arctan x = \frac{1}{1+x^2}$ folgt

$$f_W(w) = \frac{d}{dw}F_W(w) = \frac{c'}{\sqrt{n}}\left(\frac{1}{1+\frac{w^2}{n}}\right)^{\frac{n+1}{2}}.$$

Die Normierungskonstante beträgt mit (8) und der Substitution $t = \frac{r^2}{2}$, d.h. $dr = \frac{dt}{\sqrt{2t}}$

$$\frac{c'}{\sqrt{n}} = \frac{1}{\sqrt{n}}\int_0^\infty 2\,c\,e^{-r^2/2}\,r^n\,dr = \frac{2\int_0^\infty e^{-t}\,(2t)^{(n-1)/2}\,dt}{2^{\frac{n}{2}}\Gamma(\frac{n}{2})\sqrt{2\pi n}} = \frac{\Gamma(\frac{n+1}{2})}{\Gamma(\frac{n}{2})\sqrt{\pi n}}.$$

□

Bemerkung 13.7 Die Verteilungen sind vertafelt. Wegen $\left(1+\frac{x^2}{n}\right)^{-\frac{n+1}{2}} \to e^{-\frac{x^2}{2}}$ gilt

$$f_{t_n}(x) \to \varphi(x).$$

[Wegen $\frac{\overline{X}_n-\mu}{S/\sqrt{n}} \sim t_{n-1}$ war das zu erwarten!!]

Anwendung: Seien $X_1, \ldots, X_n \overset{\text{iid}}{\sim} \mathcal{N}(\mu, \sigma^2)$. Gesucht ist ein glm. bester Test zum Niveau α von $H_0 : \mu \le \mu_0$ gegen $H_A : \mu > \mu_0$. Satz 8.20:

$$\phi^*(X_1, \ldots, X_n) = \begin{cases} 1, & \text{falls } \overline{X}_n \ge c^* \\ 0, & \text{falls } \overline{X}_n < c^* \end{cases} \qquad \text{falls } \sigma^2 \text{ bekannt (unrealistisch).}$$

Satz 13.8 (Student t-Test, Ein-Stichproben-Problem) *Seien $X_1, \ldots, X_n \overset{iid}{\sim} \mathcal{N}(\mu, \sigma^2)$. Dann ist ein glm. bester Test zum Niveau α von $H_0 : \mu \le \mu_0$ gegen $H_A : \mu > \mu_0$ gegeben durch*

$$\phi^*(X_1, \ldots, X_n) = \begin{cases} 1, & \frac{\overline{X}_n-\mu_0}{S/\sqrt{n}} \ge c^* \\ 0, & \frac{\overline{X}_n-\mu_0}{S/\sqrt{n}} < c^* \end{cases}$$

mit $\mathbf{P}_{\mu_0,\sigma^2}(\frac{\overline{X}_n-\mu_0}{S/\sqrt{n}} \ge c^*) = \alpha$, d.h. $c^* = t_{n-1,1-\alpha}$ ist das $(1-\alpha)$ - Quantil der t_{n-1}-Verteilung. Das bedeutet, dass ϕ^* Lösung von

$$\begin{aligned}
&\mathbf{P}_{\mu,\sigma^2}(\phi^* = 1) = \alpha && \mu = \mu_0 && \underline{\forall\, \sigma > 0} \\
&\mathbf{P}_{\mu,\sigma^2}(\phi^* = 1) = \max && \mu > \mu_0 && \underline{\forall\, \sigma > 0} \\
&\mathbf{P}_{\mu,\sigma^2}(\phi^* = 0) = \max && \mu < \mu_0 && \underline{\forall\, \sigma > 0}
\end{aligned}$$

ist.

Beweis. → Mathematische Statistik □

Bemerkung 13.9 (Optimalität / Konfidenzintervalle)
(i) Die Kriterien für die Optimalität sind etwas anders als in Satz 8.20: Es werden die Fehler 1. und 2. Art gleichmäßig minimiert unter der Nebenbedingung, dass der Fehler 1. Art "zwischen Nullhypothese und Alternative" (dh. für $\mu = \mu_0$) gleich α ist.
(ii) Um aus diesem Test optimale Konfidenzintervalle zu konstruieren, kann man wie bisher den Annahmebereich des Tests umformen - aber auch hier ist eine etwas andere Definition für die Optimalität eines Konfidenzbereichs notwendig (s. Literatur). Auf jeden Fall folgt wie in Bemerkung 9.1 aus $\frac{\overline{X}_n - \mu}{S/\sqrt{n}} \sim t_{n-1}$, dass $[\overline{X}_n - \frac{S}{\sqrt{n}}\, t_{n-1,1-\alpha/2}, \overline{X}_n + \frac{S}{\sqrt{n}}\, t_{n-1,1-\alpha/2}]$ ein $(1-\alpha)$-Konfidenzintervall für μ (bei unbekanntem σ) ist.

Bemerkung 13.10 (Das Zwei-Stichproben-Problem)
Seien $X_1, \ldots, X_m \overset{\text{iid}}{\sim} \mathcal{N}(\mu_1, \sigma_1^2)$ und $Y_1, \ldots, Y_n \overset{\text{iid}}{\sim} \mathcal{N}(\mu_2, \sigma_2^2)$.
Teste $H_0 : \mu_1 \leq \mu_2$ gegen $H_A : \mu_1 > \mu_2$.

<u>3 Fälle</u>:

(i) <u>verbundene Stichproben</u> (vorher-nachher)
<u>Bsp.:</u> Blutwerte derselben Person vor und nach einer Behandlung;
$m = n$ und X_i und Y_i stochastisch abhängig;
typische Modellierung: $X_i - Y_i \sim \mathcal{N}(\mu, \tau^2)$;
führt zu Test $H_0 : \mu \leq 0$ gegen $H_A : \mu > 0$ bei einer Stichprobe.
(ii) <u>unverbundene Stichproben</u>, gleiche Varianzen
<u>Annahme</u>: Die X_i sind von den Y_i stochastisch unabhängig, $\sigma^2 := \sigma_1^2 = \sigma_2^2$.

Bsp.: Blutwerte von Leuten mit und ohne AIDS. Aufgrund der Annahme der gleichen Varianz σ^2 schätzt man σ^2 aus beiden Stichproben: Sei

$$S_X^2 = \frac{1}{m-1}\sum_{i=1}^{m}(X_i - \overline{X}_m)^2 \;\Rightarrow\; (m-1)\,\frac{S_X^2}{\sigma^2} \sim \chi^2_{m-1},$$

$$S_Y^2 = \frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \overline{Y}_n)^2 \;\Rightarrow\; (n-1)\,\frac{S_Y^2}{\sigma^2} \sim \chi^2_{n-1}.$$

Da $(m-1)\,\frac{S_X^2}{\sigma^2}$ und $(n-1)\,\frac{S_Y^2}{\sigma^2}$ stochastisch unabhängig sind folgt

$$\frac{1}{\sigma^2}\left[(m-1)S_X^2 + (n-1)S_Y^2\right] \sim \chi^2_{m+n-2}$$
$$\Rightarrow\; \mathbf{E}\,\underbrace{\frac{1}{m+n-2}\big((m-1)S_X^2 + (n-1)S_Y^2\big)}_{=:S^2} = \sigma^2.$$

$\overline{X}_m$ und $\overline{Y}_n$ sind stochastisch unabhängig von S^2 (aus obigem Beweis folgt, dass $\overline{X}_m$ und S_X^2 bzw. $\overline{Y}_n$ und S_Y^2 stochastisch unabhängig sind). Damit gilt für $\mu_1 = \mu_2$

$$\overline{X}_m \sim \mathcal{N}\big(\mu_1, \frac{\sigma^2}{m}\big), \quad \overline{Y}_n \sim \mathcal{N}\big(\mu_2, \frac{\sigma^2}{n}\big)$$

$$\Rightarrow\; \overline{X}_m - \overline{Y}_n \sim \mathcal{N}\big(0, \frac{\sigma^2}{m} + \frac{\sigma^2}{n}\big)$$
$$\Rightarrow\; \frac{1}{\sigma}\sqrt{\frac{mn}{m+n}}\,(\overline{X}_m - \overline{Y}_n) \sim \mathcal{N}(0,1)$$
$$\Rightarrow\; T_{m,n} := \frac{\sqrt{\frac{mn}{m+n}}\,(\overline{X}_m - \overline{Y}_n)}{S} \sim t_{m+n-2} \qquad \text{(analog zu oben).}$$

Optimaler Test: siehe nachfolgenden Satz

(iii) unverbundene Stichproben, verschiedene Varianzen
→ Behrens-Fischer Problem (ungelöst).

Satz 13.11 (Student t-Test, Zwei-Stichproben-Problem)
Seien $X_1, \ldots, X_m \overset{iid}{\sim} \mathcal{N}(\mu_1, \sigma^2)$, $Y_1, \ldots, Y_n \overset{iid}{\sim} \mathcal{N}(\mu_2, \sigma^2)$ *und die* X_i *stochastisch unabhängig von den* Y_j. *Dann ist der glm. beste Test zum Niveau* α *von* $H_0 : \mu_1 \leq \mu_2$ *gegen* $H_A : \mu_1 > \mu_2$ *gegeben durch*

$$\phi^*(\underline{X}, \underline{Y}) = \begin{cases} 1, & T_{m,n} \geq c^* \\ 0, & T_{m,n} < c^* \end{cases}$$

c^* *ist dabei das* $(1-\alpha)$ *- Quantil der* t_{m+n-2}*-Verteilung, d.h. der Verteilung von* $T_{m,n}$ *für* $\mu_1 = \mu_2$ *("Rand" der Nullhypothese).*

Beweis. $\rightarrow$ Mathematische Statistik $\square$

14 Der zentrale Grenzwertsatz

In diesem Kapitel wird die einfachste Form des zentralen Grenzwertsatzes bewiesen. Der Beweis ist durch die Verwendung eines "Teleskop-Argumentes" relativ elementar. Der zentrale Grenzwertsatz beinhaltet mit der schwachen Konvergenz eine weitere Art der Konvergenz von Zufallsvariablen. Wir studieren einige wichtige Zusammenhänge über die verschiedenen Konvergenzbegriffe für Zufallsvariable. Als Anwendung wird der χ^2-Test von Pearson behandelt, bei dem die Verteilung der Test-Statistik unter Verwendung des zentralen Grenzwertsatzes durch eine χ^2-Verteilung approximiert wird.

14.1 Motivation

Seien $X_1, \ldots, X_n$ iid, $\mathbf{E}X_i = \mu$, $\mathrm{Var}X_i = \sigma^2$, $\overline{X}_n = \frac{1}{n}\sum_{i=1}^n X_i$.

- $X_i \sim \mathcal{N}(\mu, \sigma^2)$. Dann gilt $\overline{X}_n \sim \mathcal{N}(\mu, \frac{\sigma^2}{n})$, und damit $\sqrt{n}\,\frac{\overline{X}_n - \mu}{\sigma} \sim \mathcal{N}(0,1)$.
- Im allgemeinen Fall werden wir zeigen, dass die Verteilung von $\sqrt{n}\,\frac{\overline{X}_n - \mu}{\sigma}$ gegen die $\mathcal{N}(0,1)$-Verteilung konvergiert, d.h. dass

$$\begin{array}{ccccc} P(\sqrt{n}\,\frac{\overline{X}_n-\mu}{\sigma} \le x) & \to & \Phi(x) & \forall x & \text{gilt} \\ || & & || & & \\ \mathbf{E}\,\mathrm{I}_{(-\infty,x]}\big(\sqrt{n}\,\frac{\overline{X}_n-\mu}{\sigma}\big) & & \mathbf{E}\,\mathrm{I}_{(-\infty,x]}(Z) & Z \sim \mathcal{N}(0,1). & \end{array}$$

Wir zeigen zunächst für glatte f

$$\mathbf{E}f\big(\sqrt{n}\,\frac{\overline{X}_n - \mu}{\sigma}\big) \to \mathbf{E}f(Z).$$

Wegen $\frac{\overline{X}_n-\mu}{\sigma} = \frac{1}{n}\sum_{i=1}^n(\frac{X_i-\mu}{\sigma})$ nehmen wir $\mathbb{E}\,\mathbf{E}X_i = 0, \mathrm{Var}X_i = 1$ an.

Definition 14.2 *(Schwache Konvergenz) Sei $(Z_n)_{n\in\mathbb{N}}$ eine Folge von ZVAs mit Verteilungsfunktionen F_n und Z eine ZVA mit Verteilungsfunktion F. Man sagt, dass Z_n schwach gegen Z konvergiert ($Z_n \xrightarrow{\mathcal{D}} Z$), falls*

$$F_n(z) = \mathbf{P}(Z_n \leq z) \to \mathbf{P}(Z \leq z) = F(z)$$

für alle Stetigkeitspunkte z von F gilt.

Bemerkung: (i) Da F monoton und beschränkt ist, gibt es höchstens abzählbar viele Unstetigkeitspunkte (Beweis einfach).
(ii) Bei der obigen Konvergenz handelt es sich im Grunde um die Konvergenz der Verteilungen $\mathbf{P}^{Z_n}$ gegen die Verteilung $\mathbf{P}^Z$ und nicht um die Konvergenz der ZVAs. Man sagt deshalb auch Konvergenz nach Verteilung (englisch: in distribution - daher $Z_n \xrightarrow{\mathcal{D}} Z$).

Lemma 14.3 *Sei $X_1, \ldots, X_n$ iid mit $\mathbf{E}X_i = 0$, $\mathrm{Var}X_i = 1$ und $f : \mathbb{R} \to \mathbb{R}$ zweimal stetig differenzierbar mit f'' stetig und beschränkt. Dann gilt für $Z_n = \sqrt{n}\,\overline{X}_n = \frac{1}{\sqrt{n}} \sum_{i=1}^n X_i$ und $Z \sim N(0,1)$:*

$$\mathbf{E}f(Z_n) \to \mathbf{E}f(Z).$$

Beweis. Seien $Y_1, \ldots, Y_n \overset{\text{iid}}{\sim} \mathcal{N}(0,1)$, so dass $X_1, \ldots, X_n, Y_1, \ldots, Y_n$ stoch. unabh.. Es gilt $\sqrt{n}\,\overline{Y}_n \sim \mathcal{N}(0,1)$, d.h. $\mathbf{E}(f(\sqrt{n}\,\overline{Y}_n)) = \mathbf{E}f(Z)$. Ersetze nacheinander X_i durch Y_i:

$$\begin{aligned}
f(Z_n) - f(\sqrt{n}\,\overline{Y}_n) &= f\left(\frac{X_1 + \ldots + X_n}{\sqrt{n}}\right) - f\left(\frac{Y_1 + X_2 + \ldots + X_n}{\sqrt{n}}\right) \\
&\quad + f\left(\frac{Y_1 + X_2 + \ldots + X_n}{\sqrt{n}}\right) - f\left(\frac{Y_1 + Y_2 + X_3 + \ldots + X_n}{\sqrt{n}}\right) \\
&\quad \vdots \\
&\quad + f\left(\frac{Y_1 + \ldots + Y_{n-1} + X_n}{\sqrt{n}}\right) - f\left(\frac{Y_1 + \ldots + Y_n}{\sqrt{n}}\right) \\
&= V_1 + \ldots + V_n
\end{aligned}$$

mit

$$V_i := f(U_i + \frac{X_i}{\sqrt{n}}) - f(U_i + \frac{Y_i}{\sqrt{n}})$$

und

$$U_i := \frac{Y_1 + \ldots + Y_{i-1} + X_{i+1} + \ldots + X_n}{\sqrt{n}} \, .$$

Die Taylor-Entwicklung von f um U_i ergibt:

$$V_i = \frac{X_i - Y_i}{\sqrt{n}} f'(U_i) + \frac{1}{2n} X_i^2 f''(U_i + \frac{\theta_1 X_i}{\sqrt{n}}) - \frac{1}{2n} Y_i^2 f''(U_i + \frac{\theta_2 Y_i}{\sqrt{n}})$$

mit θ_1, $\theta_2 \in [0,1]$ (Zufallsvariable!). Betrachte das Stetigkeitsmodul

$$\delta(h) := \sup_{|x-y| \le h} |f''(x) - f''(y)|$$

$$\Rightarrow V_i = \frac{X_i - Y_i}{\sqrt{n}} f'(U_i) + \frac{1}{2n}(X_i^2 - Y_i^2) f''(U_i) + R_i$$

mit $|R_i| \le \frac{1}{2n} X_i^2 \delta(\frac{|X_i|}{\sqrt{n}}) + \frac{1}{2n} Y_i^2 \delta(\frac{|Y_i|}{\sqrt{n}}) =: \frac{1}{2n} h_n(X_i) + \frac{1}{2n} h_n(Y_i)$. Satz 8.15 ergibt

$$\mathbf{E}V_i = \tfrac{1}{\sqrt{n}} \underbrace{\mathbf{E}\big((X_i - Y_i) f'(U_i)\big)}_{\begin{array}{c}\| \\ \mathbf{E}(X_i - Y_i)\, \mathbf{E} f'(U_i) \\ \| \\ 0\end{array}} + \tfrac{1}{2n} \underbrace{\mathbf{E}\big((X_i^2 - Y_i^2) f''(U_i)\big)}_{\begin{array}{c}\| \\ \mathbf{E}(X_i^2 - Y_i^2)\, \mathbf{E} f''(U_i) \\ \| \\ 0\end{array}} + \mathbf{E}R_i.$$

Es gilt nun

$$|\mathbf{E}R_i| \le \mathbf{E}|R_i| \le \frac{1}{2n} \mathbf{E}X_i^2 \, \delta\big(\frac{|X_i|}{\sqrt{n}}\big) + \frac{1}{2n} \mathbf{E}Y_i^2 \, \delta\big(\frac{|Y_i|}{\sqrt{n}}\big)$$

$$\begin{aligned} \mathbf{E}X_i^2 \, \delta\big(\frac{|X_i|}{\sqrt{n}}\big) &= \mathbf{E}X_i^2 \delta\big(\frac{|X_i|}{\sqrt{n}}\big) \big(\mathrm{I}_{\{|X_i| \le \lambda\sqrt{n}\}} + \mathrm{I}_{\{|X_i| > \lambda\sqrt{n}\}}\big) \\ &\le \underbrace{\mathbf{E}X_i^2 \, \delta(\lambda)}_{\text{bel. klein für } \lambda \text{ klein}} + C \cdot \underbrace{\mathbf{E}\big[X_i^2 \mathrm{I}_{\{|X_i| > \lambda\sqrt{n}\}}\big]}_{\to 0} \end{aligned}$$

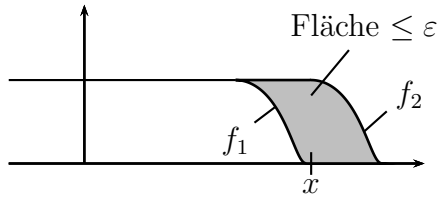
$$\Rightarrow \; \mathbf{E}R_i = \frac{1}{n} \, o(1) \quad \Rightarrow \quad \text{Beh.}$$

□

Satz 14.4 (Der zentrale Grenzwertsatz) *Seien $X_1, \ldots, X_n$ iid mit $\mathbf{E}X_i = \mu$ und $\mathrm{Var}X_i = \sigma^2 \in \mathbb{R}^+$. Dann gilt*

$$\sqrt{n}\,\frac{\overline{X}_n - \mu}{\sigma} \xrightarrow{\mathcal{D}} \mathcal{N}(0,1).$$

Beweis. Aufgrund der Definition und der Vorbetrachtung ist das Resultat bewiesen, falls wir die Aussage von Lemma 14.3 auch für $f = \mathrm{I}_{(-\infty,x]}$ zeigen. Idee: Wähle f_1, f_2 mit f_1'', f_2'' wie in Lemma 14.3 mit $f_1(t) \le \mathrm{I}_{(-\infty,x]}(t) \le f_2(t)$ und $\int (f_2(t) - f_1(t))\,dt \le \varepsilon$



Es gilt mit $Z_n = \sqrt{n}\bar{X}_n$ und $Z \sim N(0,1)$ ($\mathbb{E}\,\mathbf{E}X_i = 0, \mathrm{Var}X_i = 1$):

$$\begin{array}{ccccc} \mathbf{E}f_1(Z_n) & \le & \mathbf{E}\,\mathrm{I}_{(-\infty,x]}(Z_n) & \le & \mathbf{E}f_2(Z_n) \\ \downarrow & & & & \downarrow \\ \mathbf{E}f_1(Z) & \le & \mathbf{E}\,\mathrm{I}_{(-\infty,x]}(Z) & \le & \mathbf{E}f_2(Z) \end{array}$$

Wegen

$$\mathbf{E}\big(f_2(Z) - f_1(Z)\big) = \int \big(f_2(t) - f_1(t)\big)\,\phi(t)\,dt \le \phi(0)\,\varepsilon$$

folgt daraus die Behauptung. □

Bemerkungen 14.5 (i) Der ZGWS gilt auch, falls man die Voraussetzungen "stochastisch unabhängig" oder "identisch verteilt" etwas abschwächt ($\to$ Literatur).

(ii) Allgemeiner gilt für alle $A \in \mathcal{B}$ mit $\mathbf{P}(Z \in \partial A) = 0$, d.h. u.a. für alle Intervalle und Vereinigungen von Intervallen

$$\mathbf{P}\big(\sqrt{n}\,\frac{\overline{X}_n - \mu}{\sigma} \in A\big) \to \mathbf{P}(Z \in A).$$

14.6 Anwendung (Binomial-Approximation)
$X_i \overset{\text{iid}}{\sim} \mathcal{B}(1,p) \quad \Rightarrow \quad X := \sum_{i=1}^n X_i \sim \mathcal{B}(n,p)$

p klein, np "mittel": $P(X=k) = \binom{n}{k} p^k (1-p)^{n-k} \approx e^{-\lambda}\frac{\lambda^k}{k!},\ \lambda = np$ (vgl. Proposition 3.8)

p nicht klein: Verwende Normal-Approximation: $\mathbf{E}X_i = p\,,\ \mathrm{Var}X_i = p(1-p)$.

$$\begin{aligned}
\mathbf{P}(X=k) &= \mathbf{P}(k-0.5 < X \le k+0.5) \\
&= \mathbf{P}\Big(\frac{k-0.5-np}{\sqrt{np(1-p)}} < \underbrace{\frac{X-np}{\sqrt{np(1-p)}}}_{=\sqrt{n}\,\frac{\overline{X}_n - p}{\sqrt{p(1-p)}}} \le \frac{k+0.5-np}{\sqrt{np(1-p)}}\Big) \\
&\approx \mathbf{P}\left(\frac{k-0.5-np}{\sqrt{np(1-p)}} < Z \le \frac{k+0.5-np}{\sqrt{np(1-p)}}\right) \qquad \text{mit } Z \sim \mathcal{N}(0,1) \\
&= \Phi\left(\frac{k+0.5-np}{\sqrt{np(1-p)}}\right) - \Phi\left(\frac{k-0.5-np}{\sqrt{np(1-p)}}\right).
\end{aligned}$$

Bemerkung: Die Wahl des Intervalls $(k-0.5, k+0.5]$ erscheint etwas willkürlich. Bei dieser Wahl gilt $\sum_{k=0}^n \mathbf{P}\big(\frac{k-0.5-np}{\sqrt{np(1-p)}} < Z \le \frac{k+0.5-np}{\sqrt{np(1-p)}}\big) \to 1$, d.h. man approximiert eine normierte Folge asymptotisch auch durch eine normierte Folge.

Anwendung: Eine schöne Anwendung dieser Binomialapproximation ist in Anhang 4.1 zu finden. □

Satz 14.7 (Multivariater ZGWS) *Seien $X_1, \ldots, X_n$ iid Zufallsvektoren mit Werten in $\mathbb{R}^d$, Mittelwert-Vektor $\mu = \mathbf{E}X_i \in \mathbb{R}^d$ und Kovarianzmatrix $\Sigma = \Sigma(X_i)$. Dann gilt für alle $A \in \mathcal{B}^d$ (Borelsche σ-Algebra auf $\mathbb{R}^d$) mit $\mathbf{P}(Z \in \partial A) = 0$*

$$\mathbf{P}\big(\sqrt{n}\,(\overline{X}_n - \mu) \in A\big) \to P(Z \in A)$$

mit $Z \sim \mathcal{N}(0, \Sigma)$*, d.h.*

$$\sqrt{n}\,(\overline{X}_n - \mu) \xrightarrow{\mathcal{D}} \mathcal{N}(0, \Sigma)$$

$\big($*oder auch* $\sqrt{n}\,\Sigma^{-\frac{1}{2}}(\overline{X}_n - \mu) \xrightarrow{\mathcal{D}} \mathcal{N}(0, I)\big)$.

Beweis. s. Literatur. □

Bemerkung:
Die schwache Konvergenz im $\mathbb{R}^k$ kann dabei analog zu Definition 14.2 über die multivariaten Verteilungsfunktionen definiert werden.
[Definition $\Sigma^{-\frac{1}{2}}$: Gilt $\Sigma = P\Lambda P'$ (wie in Kapitel 12), so setzt man $\Sigma^{-1/2} := P\Lambda^{-1/2}P'$]

Jetzt wollen wir noch einige Zusammenhänge zwischen schwacher und stochastischer Konvergenz untersuchen.

Proposition 14.8 *Seien* (X_n)*,* (Y_n) *Folgen von ZVAs,* X *ZVA mit*

$$Y_n \xrightarrow{\mathcal{D}} X, \quad X_n - Y_n \xrightarrow{P} 0.$$

Dann gilt

$$X_n \xrightarrow{\mathcal{D}} X.$$

Beweis. Sei $Z_n := Y_n - X_n$. Zu zeigen ist, dass $F_{X_n}(x) \to F_X(x)$ für alle Stetigkeitspunkte x von $F_X(x)$. Seien $x \in \mathbb{R}$ und $\varepsilon > 0$ derart, dass $x, x \pm \varepsilon$ Stetigkeitspunkte von $F_X(x)$ sind.

$$\begin{aligned} F_{X_n}(x) &= \mathbf{P}(X_n \le x) = \mathbf{P}(Y_n \le x + Z_n) \\ &= \mathbf{P}(Y_n \le x + Z_n, Z_n < \varepsilon) + \mathbf{P}(Y_n \le x + Z_n,\ Z_n \ge \varepsilon) \\ &\le \mathbf{P}(Y_n \le x + \varepsilon) + \mathbf{P}(|Z_n| \ge \varepsilon) \\ \Rightarrow \quad &\limsup_{n\to\infty} F_{X_n}(x) \le F_X(x + \varepsilon). \end{aligned}$$

Analog folgt (setze $Z_n > -\varepsilon$ anstelle von $Z_n < \varepsilon$)

$$\liminf_{n\to\infty} F_{X_n}(x) \ge F_X(x - \varepsilon).$$

Da x Stetigkeitspunkt ist, folgt mit $\varepsilon \to 0$

$$\lim_{n\to\infty} F_{X_n}(x) = F_X(x).$$

□

Proposition 14.9 *Seien (X_n), (Y_n) Folgen von ZVAs, X ZVA und c eine Konstante. Dann gilt*

(i) $X_n \xrightarrow{P} X \quad \Rightarrow \quad X_n \xrightarrow{\mathcal{D}} X,$

(ii) $X_n \xrightarrow{\mathcal{D}} X,\ Y_n \xrightarrow{P} 0 \quad \Rightarrow \quad X_n Y_n \xrightarrow{P} 0,$

(iii) (Satz von Slutsky)

$$X_n \xrightarrow{\mathcal{D}} X,\ Y_n \xrightarrow{P} c \quad \Rightarrow \quad \begin{aligned} &X_n + Y_n \xrightarrow{\mathcal{D}} X + c, \\ &X_n \cdot Y_n \xrightarrow{\mathcal{D}} cX, \\ &X_n / Y_n \xrightarrow{\mathcal{D}} X/c \text{ , falls } \ c \neq 0, \end{aligned}$$

(iv) $X_n \xrightarrow{P} c$, *$h(\cdot)$ stetig in c* $\quad \Rightarrow \quad h(X_n) \xrightarrow{P} h(c).$

Beweis.

(i) Setze in Prop. 14.8 $Y_n \equiv X$.

(ii)

$$\begin{aligned} \mathbf{P}(|X_n Y_n| \geq \varepsilon) &= \mathbf{P}(|X_n Y_n| \geq \varepsilon, |Y_n| < \frac{\varepsilon}{k}) + \mathbf{P}(|X_n Y_n| \geq \varepsilon, |Y_n| \geq \frac{\varepsilon}{k}) \\ &\leq \mathbf{P}(|X_n| > k) + \mathbf{P}(|Y_n| \geq \frac{\varepsilon}{k}) \end{aligned}$$

$$\Rightarrow \limsup_{n\to\infty} \mathbf{P}(\ |X_n Y_n| \geq \varepsilon) \leq \mathbf{P}(|X| > k), \text{ falls } \ -k, k \text{ Stetigkeitspunkte von } F_X \text{ sind.}$$

Ersetze k durch eine monotone Folge von Stetigkeitspunkten $k_m \in \mathbb{R}$. Wegen $\sum_{m=1}^{\infty} \mathbf{P}\big(|X| \in (k_{m-1}, k_m]\big) < \infty$ folgt $\lim_{m\to\infty} \mathbf{P}(|X| > k_m) = 0$ und damit das Ergebnis .

(iii) a) $X_n \xrightarrow{\mathcal{D}} X \Rightarrow X_n + c \xrightarrow{\mathcal{D}} X + c$

Es gilt $(X_n + Y_n) - (X_n + c) = Y_n - c \xrightarrow{P} 0$

$\Rightarrow X_n + Y_n \xrightarrow{\mathcal{D}} X + c$ nach Prop. 14.8

b) Es gilt: $X_n \xrightarrow{\mathcal{D}} X \Rightarrow cX_n \xrightarrow{\mathcal{D}} cX$.
Andererseits: $X_nY_n - cX_n = X_n(Y_n - c) \xrightarrow{P} 0$ nach (ii)
$\Rightarrow X_nY_n \xrightarrow{\mathcal{D}} cX$ nach Prop. 14.8.

c) X_n/Y_n analog.

(iv) $h(\cdot)$ stetig $\Rightarrow \forall \varepsilon > 0\ \exists \delta > 0 : \mathbf{P}\big(|h(X_n) - h(c)| > \varepsilon\big) \leq \mathbf{P}\big(|X_n - c| > \delta\big) \to 0$.

□

Bemerkung 14.10 Analog zu (iv) gibt es auch ein sogenanntes "continuous mapping theorem" für die schwache Konvergenz (siehe W'Theorie). Man kann damit z.B. von der schwachen Konvergenz des Vektors $(X_n, Y_n) \xrightarrow{\mathcal{D}} (X, Y)$ auf $X_n + Y_n \xrightarrow{\mathcal{D}} X + Y$ schließen. Hierbei ist zu beachten, dass X und Y evtl. stochastisch abhängig sind.

14.11 Anwendung Seien X_i iid, $\mathbf{E}X_i = \mu$, $\mathrm{Var}X_i = \sigma^2$ und $\mathbf{E}(X_i - \mu)^4 = \mu_4$.
Dann gilt

$$\overline{X}_n \xrightarrow{P} \mu \quad \text{(schwaches Gesetz)}.$$

Setze $Y_i := (X_i - \mu)^2$. Damit ist $\mathbf{E}Y_i = \sigma^2$, $\mathrm{Var}\, Y_i = \mathbf{E}Y_i^2 - (\mathbf{E}Y_i)^2 = \mu_4 - \sigma^4$ und es folgt aus dem ZGWS

$$\sqrt{n}\left[\frac{1}{n}\sum_{i=1}^{n}(X_i - \mu)^2 - \sigma^2\right] \xrightarrow{\mathcal{D}} \mathcal{N}(0, \mu_4 - \sigma^4).$$

Wir wollen mit Proposition 14.8 folgern, dass auch

$$\sqrt{n}\,(S_n^2 - \sigma^2) \xrightarrow{\mathcal{D}} \mathcal{N}(0, \mu_4 - \sigma^4).$$

Differenz:

$$
\begin{aligned}
&\sqrt{n}\left[\frac{1}{n-1}\sum_{i=1}^{n}(X_i-\overline{X}_n)^2-\frac{1}{n}\sum_{i=1}^{n}(X_i-\mu)^2\right]\\
&=\sqrt{n}\left[(\frac{1}{n-1}-\frac{1}{n})\sum_{i=1}^{n}(X_i-\mu)^2-\frac{n}{n-1}(\overline{X}-\mu)^2\right]\\
&=\underbrace{\underbrace{\frac{\sqrt{n}}{n-1}}_{\to 0}\underbrace{\frac{1}{n}\sum_{i=1}^{n}(X_i-\mu)^2}_{\xrightarrow{P}\sigma^2\ \text{(S.G.)}}}_{\xrightarrow{P}0}-\underbrace{\underbrace{\frac{n}{n-1}}_{\to 1}\underbrace{(\overline{X}-\mu)}_{\xrightarrow{P}0\ \text{(S.G.)}}\underbrace{\sqrt{n}\,(\overline{X}-\mu)}_{\xrightarrow{\mathcal{D}}\mathcal{N}(0,\sigma^2)}}_{\xrightarrow{P}0}
\end{aligned}
$$

[S.G. = schwaches Gesetz der großen Zahlen]

Ähnlich: t-Statistik (vor allem im nicht Gaußschen Fall interessant)

$$
T_{n-1}=\frac{\overline{X}-\mu}{S/\sqrt{n}}=\underbrace{\frac{1}{S}}_{\xrightarrow{P}\frac{1}{\sigma}(*)}\underbrace{\sqrt{n}\,(\overline{X}-\mu)}_{\xrightarrow{\mathcal{D}}\mathcal{N}(0,\sigma^2)}\quad\xrightarrow{\mathcal{D}}\mathcal{N}(0,1)
$$

Zu $(*)$:

$$
\begin{aligned}
&\frac{1}{n}\sum_{i=1}^{n}(X_i-\mu)^2\xrightarrow{P}\sigma^2\quad\text{(schwaches Gesetz)}\\
&\Rightarrow S_n^2\xrightarrow{P}\sigma^2\quad\text{(s.o.: gleiche Abschätzung der Differenz)}\\
&\Rightarrow S_n\xrightarrow{P}\sigma\quad\text{(als Übung nachrechnen)}
\end{aligned}
$$

14.12 Pearson's χ^2-Test

In den Literaturwissenschaften werden häufig wahrscheinlichkeitstheoretische Modelle verwendet, um zu testen, ob ein Text einem bestimmten Autor zugeordnet werden kann. Man testet dann z.B. wie oft ein Text bestimmte für den jeweiligen Author charakteristische Wörter enthält (z.B. "ohne", "dieser", "Hoffnung" usw.).

Situation: N_j Anzahl von Beobachtungen des Typs j ($j=1,\dots,r$). Der Typ j trete mit

W't p_j auf, $\sum_{j=1}^r p_j = 1$. Wir wollen Hypothesen bzgl. der p_j testen.

Genauer: $X_1, \ldots, X_n$ iid mit

$$X_i = (X_{i1}, \ldots, X_{ir})' = \underset{\text{zufällige Stelle mit W't } p_j}{(0, \ldots, 0, \underset{\uparrow}{1}, 0, \ldots, 0)'}$$

$\sum_{i=1}^n X_i =: (N_1, \ldots, N_r)'$. Es gilt

$$p(n_1, \ldots, n_r) = \frac{n!}{n_1! \cdot \ldots \cdot n_r!}\, p_1^{n_1} \cdot \ldots \cdot p_r^{n_r} \text{ (Multinomialverteilung)}$$

Hypothese: $H_0 : p_j = \pi_j$ $(j = 1, \ldots, r)$, wobei π_j bekannt ist mit $\sum_{j=1}^r \pi_j = 1$.

Pearson's χ^2-Test:

$$\phi(\underline{x}) = \begin{cases} 0, & \sum_{i=1}^r \frac{(N_i - n\pi_i)^2}{n\pi_i} \leq c^* \\ 1, & \sum_{i=1}^r \frac{(N_i - n\pi_i)^2}{n\pi_i} > c^* \end{cases} .$$

Zur Bestimmung von c^* benötigt man die Verteilung von $\sum_{i=1}^r \frac{(N_i - n\pi_i)^2}{n\pi_i}$. Gibt es eine einfache Approximation durch den zentralen Grenzwertsatz?

Problem: N_i und N_j sind stochastisch abhängig.

Trick: Sei $U_i = \frac{N_i - n\pi_i}{\sqrt{n\pi_i}}$, $U = (U_1, \ldots, U_r)'$ und $P_1 = (\sqrt{\pi_1}, \ldots, \sqrt{\pi_r})'$. Ergänze P_1 zu einer ON-Basis $P_1, \ldots, P_r$ des $\mathbb{R}^r$. Es gilt $P_1'U = \sum_{i=1}^r \frac{N_i - n\pi_i}{\sqrt{n}} = 0$ und damit

$$\sum_{i=1}^r \frac{(N_i - n\pi_i)^2}{n\pi_i} = U'U = U'PP'U = \sum_{j=2}^r (P_j'U)^2 \quad \text{wobei}$$

$$\{P_j'U\}_{j=2,\ldots,r} = \sqrt{n}\, \frac{1}{n} \sum_{i=1}^n \underbrace{\left\{ P_j' \Big(\frac{X_{i1} - \pi_1}{\sqrt{\pi_1}}, \ldots, \frac{X_{ir} - \pi_r}{\sqrt{\pi_r}} \Big)' \right\}_{j=2,\ldots,r}}_{=: \underline{Y}_i} .$$

Unter H_0 gilt $\mathbf{E}\underline{Y}_i = 0$ und

$$\mathrm{Kov}(X_{ij}, X_{ik}) = \mathbf{E}X_{ij}X_{ik} - \pi_j\pi_k = \pi_j\delta_{jk} - \pi_j\pi_k \quad \text{, d.h.}$$

$$\Sigma(\underline{Y}_i)_{j,k} = P_{j+1}'\, \Sigma\Big(\big(\frac{X_{i1}}{\sqrt{\pi_1}}, \ldots, \frac{X_{ir}}{\sqrt{\pi_r}} \big)' \Big)\, P_{k+1}$$

$$= P_{j+1}'\, (I - P_1P_1')\, P_{k+1} = \delta_{jk}$$

Sei nun $C = \{x \in \mathbb{R}^{r-1} \mid \sum x_j^2 \leq c\}$

$$\Rightarrow \mathbf{P}\Big(\sum_{i=1}^{r} \frac{(N_i - n\pi_i)^2}{n\pi_i} \leq c\Big) = \mathbf{P}\Big(\sum_{j=2}^{r} (P_j'U)^2 \leq c\Big) = \mathbf{P}\Big((P_2'U, \ldots, P_r'U)' \in C\Big)$$

$$\overset{\text{multiv. ZGWS}}{\to} \mathbf{P}\big(\underbrace{(Z_1, \ldots, Z_{r-1})}_{\sim \mathcal{N}(0, I_{r-1})}' \in C\big) = \mathbf{P}\Big(\underbrace{\sum_{j=1}^{r-1} Z_j^2}_{\sim \chi^2_{r-1}} \leq c\Big) \qquad \text{, d.h.}$$

$$\sum_{i=1}^{r} \frac{(N_i - n\pi_i)^2}{n\pi_i} \xrightarrow{\mathcal{D}} \chi^2_{r-1}.$$

Damit liefert $c^* = \chi^2_{r-1,1-\alpha}$ einen Test, der (zumindest) asymptotisch das Niveau α einhält.

[In dem obigen Argument ist das "continuous mapping theorem" versteckt]

15 Maximum-Likelihood-Schätzer

In diesem Kapitel werden der Maximum-Likelihood-Schätzer als allgemeines Schätzprinzip eingeführt und seine Eigenschaften wie Konsistenz und asymptotische Normalität diskutiert. Wir führen die Beweise zunächst allgemein - beschränken uns aber an den entscheidenden Stellen dann auf Exponentialfamilien, um die Voraussetzungen und Beweise für eine einführende Vorlesung noch halbwegs übersichtlich zu halten. Ferner zeigen wir die Cramér-Rao Ungleichung als untere Schranke für die Varianz eines Schätzers und beweisen, dass der Maximum-Likelihood-Schätzer diese untere Schranke asymptotisch annimmt (Fisher-Effizienz).

Bemerkung 15.1 (Maximum-Likelihood-Schätzer (MLE)) Seien $X_1, \ldots, X_n$ ZVAs mit gemeinsamer Dichte $f_{\theta_0}^{(n)}(x)$ (bzw. Zähldichte $p_{\theta_0}^{(n)}(x)$) und einem Parameter $\theta_0 \in \Theta \subseteq \mathbb{R}^d$. Bekannt ist die Klasse von Verteilungen $\{f_\theta^{(n)} \mid \theta \in \Theta\}$, aber nicht der konkrete (wahre) Wert θ_0. Gesucht ist ein Schätzer $\widehat{\theta}_n$ für θ_0. Wir betrachten hier meistens nur den Spezialfall, dass die ZVAs iid mit eindimensionaler Dichte $f_{\theta_0}^{(1)}(x) = f_{\theta_0}(x)$ sind.

Maximum-Likelihood-Schätzer (MLE):

$$\begin{aligned}\widehat{\theta}_n &:= \operatorname*{argmax}_{\theta\in\Theta} f_\theta^{(n)}(X_1, \ldots, X_n) \\ &= \operatorname*{argmax}_{\theta\in\Theta} \prod_{i=1}^{n} f_\theta(X_i) \qquad \text{(falls die } X_i \text{ iid)}.\end{aligned}$$

Hierbei kann $f_\theta(x)$ sowohl die W-Dichte einer stetigen Verteilung als auch die Zähldichte $f_\theta(x) = p_\theta(x)$ einer diskreten Verteilung sein. Alle Ergebnisse dieses Kapitels gelten für beide Fälle.

Beispiele:

(i) $X_1, \ldots, X_n \overset{\text{iid}}{\sim} \mathcal{P}(\lambda)\,, \quad \theta = \lambda$

$$\Rightarrow \quad p_\theta^{(n)}(x_1, \ldots, x_n) = \prod_{i=1}^{n} e^{-\theta} \frac{\theta^{x_i}}{x_i!} = e^{-\theta n} \frac{\theta^{\Sigma x_i}}{\prod(x_i!)}$$

$$\Rightarrow \quad \frac{\partial}{\partial\theta}\, p_\theta^{(n)}(x_1, \ldots, x_n) = \frac{(-n)\, e^{-\theta n}\, \theta^{\Sigma\, x_i}}{\prod(x_i!)} + \frac{e^{-\theta n}\, (\Sigma\, x_i)\, \theta^{\Sigma\, x_i - 1}}{\prod(x_i!)} \overset{!}{=} 0$$

$$\Leftrightarrow \quad (-n) + \frac{1}{\theta} \Sigma\, x_i = 0 \;\Leftrightarrow\; \theta = \frac{1}{n} \Sigma\, x_i$$

d.h. $\widehat{\theta}_n = \widehat{\lambda}_n = \overline{X}_n$ [zur Erinnerung: es gilt $\mathbf{E}X_i = \lambda$].

(ii) $X_1, \ldots, X_n \overset{\text{iid}}{\sim} \mathcal{E}(\lambda)\,, \quad \theta = \lambda$

$$\Rightarrow \quad f_\theta^{(n)}(x_1, \ldots, x_n) = \prod_{i=1}^{n} \theta\, \mathrm{e}^{-\,\theta x_i} = \theta^n\, \mathrm{e}^{-\,\theta \sum_{i=1}^n x_i}$$

$$\Rightarrow \quad \frac{\partial}{\partial\theta}\, f_\theta^{(n)}(x_1, \ldots, x_n) = n\, \theta^{n-1}\, \mathrm{e}^{-\,\theta \sum_{i=1}^n x_i} - \theta^n \Big(\sum_{i=1}^{n} x_i \Big)\, \mathrm{e}^{-\,\theta \sum_{i=1}^n x_i} \overset{!}{=} 0$$

$$\Leftrightarrow \quad n - \theta\, \Sigma\, x_i = 0 \;\Leftrightarrow\; \theta = \frac{1}{\frac{1}{n} \Sigma\, x_i}$$

d.h. $\widehat{\theta}_n = \widehat{\lambda}_n = \frac{1}{\overline{X}_n}$ [zur Erinnerung: es gilt $\mathbf{E}X_i = \frac{1}{\lambda}$].

(iii) $X_1, \ldots, X_n \overset{\text{iid}}{\sim} \mathcal{N}(\mu, \sigma^2)\,, \quad \theta = (\mu, \sigma^2)$

$$\overset{\text{Ü-Aufgabe}}{\Rightarrow} \widehat{\theta}_n = (\hat{\mu}_n, \hat{\sigma}_n^2) \text{ mit } \widehat{\mu}_n = \overline{X}_n \text{ und } \widehat{\sigma}_n^2 = \frac{1}{n} \sum_{i=1}^{n} (X_i - \overline{X}_n)^2.$$

(iv) Wenn $X_1, \ldots, X_n$ stochastisch abhängig sind, faktorisiert $f_\theta^{(n)}$ nicht. Für MLEs in dieser Situation gibt es zahlreiche Beispiele aus dem Bereich der stochastischen Prozesse ($\rightarrow$ KV Statistik).

Da $\log(\cdot)$ monoton ist, maximiert $\widehat{\theta}_n$ auch $\log f_\theta^{(n)}(X_1, \ldots, X_n)$ bzw. minimiert

$$L_n(\theta) := -\frac{1}{n} \log f_\theta^{(n)}(X_1, \ldots, X_n) \quad \Big(= -\frac{1}{n} \sum_{i=1}^{n} \log f_\theta(X_i) \quad \text{im iid-Fall} \Big). \qquad (*)$$

Im iid-Fall ergibt das schwache Gesetz der großen Zahlen, dass $L_n(\theta)$ stochastisch gegen

$$L(\theta) := -\mathbf{E}_{\theta_0} \log f_\theta(X_1) \qquad (**)$$

konvergiert. Wir werden unten beweisen, dass daraus auch die stochastische Konvergenz von den minimierenden Werten

$$\widehat{\theta}_n := \operatorname{argmin}_{\theta\in\Theta} L_n(\theta)$$

gegen

$$\theta_0 \overset{(zz)}{=} \operatorname{argmin}_{\theta\in\Theta} L(\theta),$$

d.h. die Konsistenz von $\widehat{\theta}_n$, folgt. Zunächst zeigen wir aber, dass θ_0 die Funktion $L(\theta)$ minimiert.

Proposition 15.2 *θ_0 ist das eindeutig bestimmte Minimum der Funktion $L(\theta)$.*

Beweis. Es gilt

$$L(\theta) = -\mathbf{E}_{\theta_0} \log f_{\theta_0}(X) + \mathbf{E}_{\theta_0} \log \frac{f_{\theta_0}(X)}{f_\theta(X)}.$$

Wegen $\log x \le x - 1 \ \forall x$ gilt $\log \frac{1}{x} \ge 1 - x \ \forall x$ und damit [im stetigen Fall, diskret analog]

$$\mathbf{E}_{\theta_0} \log \frac{f_{\theta_0}(X)}{f_\theta(X)} = \int \log \frac{f_{\theta_0}(x)}{f_\theta(x)} \, f_{\theta_0}(x)\, dx \ge \int \Big(1 - \frac{f_\theta(x)}{f_{\theta_0}(x)}\Big) \, f_{\theta_0}(x)\, dx = 1 - 1 = 0$$

d.h.

$$L(\theta) = -\mathbf{E}_{\theta_0} \log f_\theta(X) \ge -\mathbf{E}_{\theta_0} \log f_{\theta_0}(X)$$

mit Gleichheit falls $\theta = \theta_0$, d.h. θ_0 ist Minimum von $L(\theta)$. Die Eindeutigkeit ist schwieriger [man braucht dafür sog. Identifizierbarkeitsbedingungen]. Heuristisch: Oben gilt die Gleichheit nur falls $f_\theta(x) = f_{\theta_0}(x)$ für alle x, d.h. das Minimum ist eindeutig. □

Der Beweis der Konsistenz erfolgt nun zunächst für allgemeine Funktionen $L_n(\theta)$ bzw. $L(\theta)$ [mit weiteren Anwendungen, z.B. für Minimum-Distanz-Schätzer, die wir hier aber nicht diskutieren wollen]. Die entsprechenden Annahmen werden wir anschließend für die speziellen Funktionen aus (*) und (**) nachrechnen. Außerdem werden wir später noch einen zentralen Grenzwertsatz für $\widehat{\theta}_n$ beweisen.

Proposition 15.3 *Sei $L_n(\theta)$ eine Folge von stochastischen Funktionen und $L(\theta)$ eine deterministische Funktion mit*

(i) L ist stetig und $\theta_0 = \arg\min_{\theta\in\Theta} L(\theta)$ ist eindeutig,

(ii) Θ ist kompakt und $\theta_0 \in \mathrm{Int}(\Theta)$,

(iii) $\sup_{\theta\in\Theta} |L_n(\theta) - L(\theta)| \xrightarrow{P} 0$.

Dann gilt für $\widehat{\theta}_n := \mathrm{argmin}_{\theta\in\Theta} L_n(\theta)$

$$\widehat{\theta}_n \xrightarrow{P} \theta_0.$$

Bemerkung: Im Fall von Vektoren bedeutet $\widehat{\theta}_n \xrightarrow{P} \theta_0$ ebenfalls $\mathbf{P}\big(|\widehat{\theta}_n - \theta_0| > \varepsilon\big) \to 0$, wobei jetzt $|\cdot|$ die ℓ_2 - Norm ist. Man kann leicht zeigen, dass dieses äquivalent zur stochastischen Konvergenz der einzelnen Komponenten ist.

Beweis. Wir behaupten zunächst, dass es für alle $\varepsilon > 0$ ein $\delta > 0$ gibt mit

$$|\theta - \theta_0| > \varepsilon \Rightarrow |L(\theta) - L(\theta_0)| > \delta. \qquad (*)$$

Beweis: Falls dieses nicht gelten würde, würde es ein $\varepsilon_0 > 0$ und eine Folge θ_n geben mit $|\theta_n - \theta_0| > \varepsilon_0\ \forall n$ und $|L(\theta_n) - L(\theta_0)| \to 0$. Da Θ kompakt ist, hat die Folge (θ_n) einen Häufungspunkt θ', d.h. es gibt eine Teilfolge $\theta_{n_k} \to \theta'$ mit $L(\theta') = \lim L(\theta_{n_k}) = L(\theta_0)$ (wegen der Stetigkeit von L). Wegen der Eindeutigkeit von θ_0 folgt daraus aber $\theta' = \theta_0$ und damit ein Widerspruch zu $|\theta_n - \theta_0| > \varepsilon_0\ \forall n$.

Wegen $L_n(\widehat{\theta}_n) \le L_n(\theta_0)$ und $L(\theta_0) \le L(\widehat{\theta}_n)$ folgt nun

$$0 \le L(\widehat{\theta}_n) - L(\theta_0) = \big(L(\widehat{\theta}_n) - L_n(\widehat{\theta}_n)\big) + \underbrace{\big(L_n(\widehat{\theta}_n) - L_n(\theta_0)\big)}_{\le 0} + \big(L_n(\theta_0) - L(\theta_0)\big)$$

und damit aus (*)

$$\begin{aligned}\mathbf{P}\big(|\widehat{\theta}_n - \theta_0| > \varepsilon\big) &\le \mathbf{P}\big(|L(\widehat{\theta}_n) - L(\theta_0)| > \delta\big) \\ &\le \mathbf{P}\big(|L_n(\widehat{\theta}_n) - L(\widehat{\theta}_n)| > \frac{\delta}{2}\big) + \mathbf{P}\big(|L_n(\theta_0) - L(\theta_0)| > \frac{\delta}{2}\big) \to 0.\end{aligned}$$

□

[Die Annahme der Kompaktheit entspricht oft nicht der Realität. Sie ist aber üblich und wird gemacht, damit die Beweise nicht zu schwierig werden. Um sie wegzulassen, müsste man z.B. (*) annehmen und die Beziehung $\mathbf{P}\big(|L_n(\widehat{\theta}_n) - L(\widehat{\theta}_n)| > \frac{\delta}{2}\big) \to 0$ anders beweisen. Der obige Beweis ist von der Grundstruktur analog zum Konsistenzbeweis von Wald (1949). Eine Alternative wäre z.B. die Einschränkung auf sog. Exponentialfamilien (s.u.), wo der Beweis einfacher wird, weil man den MLE explizit hinschreiben kann.]

Beispiel 15.4 Wir beweisen nun die Annahme (iii) im Fall $X_1, \ldots, X_n \overset{\text{iid}}{\sim} \mathcal{P}(\theta)$, d.h.

$$L_n(\theta) = -\frac{1}{n}\sum_{i=1}^{n} \log\Big(e^{-\theta}\,\frac{\theta^{X_i}}{X_i!}\Big) = \frac{1}{n}\sum_{i=1}^{n}\Big(\theta - X_i \log\theta + \log X_i!\Big),$$

$$L(\theta) = -\mathbf{E}_{\theta_0} \log f_\theta(X_1) = \Big(\theta - \theta_0 \log\theta + \mathbf{E}_{\theta_0} \log X_1!\Big)$$

und damit

$$\begin{aligned}\sup_{\theta\in\Theta} |L_n(\theta) - L(\theta)| &= \sup_{\theta\in\Theta} \Big| -\Big(\frac{1}{n}\sum_{i=1}^{n} X_i - \theta_0\Big)\log\theta + \Big(\frac{1}{n}\sum_{i=1}^{n}\log X_i! - \mathbf{E}_{\theta_0}\log X_1!\Big)\Big| \\ &\le \Big|\frac{1}{n}\sum_{i=1}^{n} X_i - \theta_0\Big| \sup_{\theta\in\Theta}\big|\log\theta\big| + \Big|\frac{1}{n}\sum_{i=1}^{n}\log X_i! - \mathbf{E}_{\theta_0}\log X_1!\Big| \xrightarrow{P} 0\end{aligned}$$

falls $\Theta = [c_1, c_2]$ mit $0 < c_1 < c_2$ [die Ausdrücke in den Betragsstrichen $|\cdot|$ konvergieren beide nach dem schwachen Gesetz der großen Zahlen stochastisch gegen 0]. □

Für viele andere Verteilungen kann man Annahme (iii) aus Proposition 15.3 völlig analog nachrechnen. Das liegt daran, dass diese Verteilungen eine sogenannte Exponentialfamilie bilden [z.B. $\mathcal{B}(n,p)$ oder $\mathcal{N}(\mu,\sigma^2)$ - s. unten].

Definition 15.5 (Exponentialfamilie) Eine Familie $\mathcal{P} = \{\mathbf{P}_\theta | \theta \in \Theta\}$ von Verteilungen heißt Exponentialfamilie falls die $\mathbf{P}_\theta$ eine Wahrscheinlichkeitsdichte oder eine Zähl-

dichte von der Form

$$f_\theta(x) = C(\theta)\, \exp\Big(\sum_{j=1}^{s} \eta_j(\theta)\, T_j(x) \Big)\, h(x)$$

haben, wobei $\eta_j(\theta)$ und $C(\theta)$ stetige reellwertige Funktionen und $T_j(x)$ reellwertige Statistiken sind.

Satz 15.6 *Seien* $X_1, \ldots, X_n \overset{iid}{\sim} \mathbf{P}_\theta$ *und* $\mathcal{P} = \{\mathbf{P}_\theta | \theta \in \Theta\}$ *eine Exponentialfamilie mit* $\eta_j(\cdot)$ *stetig auf* Θ*, und gelten die Annahmen (i) und (ii) aus Proposition 15.3. Dann ist auch Annahme (iii) erfüllt und der Maximum-Likelihood-Schätzer ist konsistent.*

Beweis. Es gilt

$$\log f_\theta(x) = \log C(\theta) + \sum_{j=1}^{s} \eta_j(\theta)\, T_j(x) + \log h(x). \qquad (\triangle)$$

und damit völlig analog zu obigem Beispiel

$$\begin{aligned}
&\sup_{\theta\in\Theta} \big| L_n(\theta) - L(\theta) \big| \\
&= \sup_{\theta\in\Theta} \Big| \sum_{j=1}^{s} \eta_j(\theta) \Big(\frac{1}{n} \sum_{i=1}^{n} T_j(X_i) - \mathbf{E}_{\theta_0} T_j(X_1) \Big) + \Big(\frac{1}{n} \sum_{i=1}^{n} \log h(X_i) - \mathbf{E}_{\theta_0} \log h(X_1) \Big) \Big| \\
&\leq \sum_{j=1}^{s} \underbrace{\sup_{\theta\in\Theta} \big|\eta_j(\theta)\big|}_{\leq K < \infty} \Big| \frac{1}{n} \sum_{i=1}^{n} T_j(X_i) - \mathbf{E}_{\theta_0} T_j(X_1) \Big| + \Big| \frac{1}{n} \sum_{i=1}^{n} \log h(X_i) - \mathbf{E}_{\theta_0} \log h(X_1) \Big| \overset{P}{\longrightarrow} 0
\end{aligned}$$

da Θ kompakt ist [die Ausdrücke in den Betragsstrichen $|\cdot|$ konvergieren alle nach dem schwachen Gesetz der großen Zahlen stochastisch gegen 0]. □

Bemerkung 15.7 (i) Die meisten Verteilungen bilden eine Exponentialfamilie, z.B. $\mathcal{P}(\lambda)$ (s.oben), $\mathcal{B}(n,p)$, $\mathcal{E}(\lambda)$ (s. oben), $\mathcal{N}(\mu, \sigma^2)$, $\mathcal{N}(\underline{\mu}, \Sigma)$ (Ü-Aufgabe) - aber auch die Familie der Beta- und Gamma-Verteilungen (s. Literatur). Ausnahme: $\mathcal{R}[a,b]$. Obiger Satz liefert für alle diese Verteilungen die Konsistenz des MLEs. Kleiner Schönheitsfehler dabei ist

die Kompaktheit von Θ, die jeweils zu leicht modifizierten ("abgeschnittenen") Schätzern führt. Für Exponentialfamilien kann man aber auf diese Annahme verzichten (s. z.B. Bickel und Doksum, Mathematical Statistics, sec. ed., Kap. 5.2.1).
(ii) Man kann die Bedingung (iii) auch alternativ unter Differenzierbarkeitsbedingungen an $\log f_\theta(x)$ nachrechnen. Dieses ist an dieser Stelle aber zu aufwendig.

Wir wollen jetzt die asymptotische Normalität des MLE nachweisen. Zunächst beweisen wir wie bei der Konsistenz ein allgemeines Resultat.

Proposition 15.8 *Gelten zusätzlich zu den Annahmen (i)-(iii) aus Proposition 15.3 noch*

(iv) L *und* L_n *sind zweimal stetig differenzierbar in* θ *und* $W := \nabla^2 L(\theta_0)$ *ist positiv definit,*

(v) $\sup_{\theta\in\Theta} |\nabla^2 L_n(\theta) - \nabla^2 L(\theta)| \xrightarrow{P} 0$,

(vi) $\sqrt{n}\,\nabla L_n(\theta_0) \xrightarrow{D} \mathcal{N}(0, V)$.

Dann folgt

$$\sqrt{n}\big(\widehat{\theta}_n - \theta_0\big) \xrightarrow{\mathcal{D}} \mathcal{N}\big(0, W^{-1}VW^{-1}\big).$$

Beweis. Wir beweisen den Satz nur für den Fall eines eindimensionalen Parameters θ - verwenden aber direkt die multivariaten Symbole $\nabla L(\theta) = \frac{\partial}{\partial\theta} L(\theta)$ und $\nabla^2 L(\theta) = \frac{\partial^2}{\partial\theta^2} L(\theta)$. Die Verallgemeinerung zum multivariaten Fall ist unten im Anhang zu finden. Wir erhalten mit dem Mittelwertsatz

$$\sqrt{n}\,\nabla L_n(\widehat{\theta}_n) - \sqrt{n}\,\nabla L_n(\theta_0) = \nabla^2 L_n(\bar{\theta}_n)\,\sqrt{n}\,(\widehat{\theta}_n - \theta_0) \qquad (*)$$

mit $|\bar{\theta}_n - \theta_0| \le |\widehat{\theta}_n - \theta_0|$. Aus Proposition 15.3 folgt $\widehat{\theta}_n \xrightarrow{P} \theta_0$ und damit auch $\bar{\theta}_n \xrightarrow{P} \theta_0$. Falls $\widehat{\theta}_n$ im Inneren von Θ liegt, haben wir $\nabla L_n(\widehat{\theta}_n) = 0$. Falls $\widehat{\theta}_n$ auf dem Rand von Θ liegt, dann impliziert die Annahme, dass θ_0 im Inneren liegt, $|\widehat{\theta}_n - \theta_0| \ge \delta$ für ein $\delta > 0$, d.h. wir erhalten (insgesamt) $\mathbf{P}\big(\sqrt{n}|\nabla L_n(\widehat{\theta}_n)| \ge \varepsilon\big) \le \mathbf{P}\big(|\widehat{\theta}_n - \theta_0| \ge \delta\big) \to 0$ für alle $\varepsilon > 0$ und damit wegen (vi) (vgl. Proposition 14.9 (iii))

$$\nabla^2 L_n(\bar{\theta}_n)\,\sqrt{n}\,(\widehat{\theta}_n - \theta_0) \xrightarrow{\mathcal{D}} \mathcal{N}(0, V)$$

Aus (v) und der Stetigkeit von $\nabla^2 L(\theta)$ folgt (vgl. Proposition 14.9 (iv))

$$\nabla^2 L_n(\bar{\theta}_n) - W = \big(\nabla^2 L_n(\bar{\theta}_n) - \nabla^2 L(\bar{\theta}_n)\big) + \big(\nabla^2 L(\bar{\theta}_n) - \nabla^2 L(\theta_0)\big) \xrightarrow{P} 0$$

und damit (wiederum mit Proposition 14.9 (iii)) die Behauptung. □

Bemerkung. Wir wollen noch einmal die Essenz des obigen Beweises festhalten: Durch eine Taylorentwicklung von $\nabla L_n(\theta)$ um θ_0 (das ist anders formuliert die obige Anwendung des Mittelwertsatzes) sieht man, dass das asymptotische Verhalten von $\sqrt{n}\,(\widehat{\theta}_n - \theta_0)$ durch das asymptotische Verhalten von $\sqrt{n}\,\nabla L_n(\theta_0)$ bestimmt wird ($\nabla^2 L_n(\bar{\theta}_n)$ ist praktisch eine Konstante). Dieser Term ist aber eine Summe von iid-Variablen und wir können den normalen zentralen Grenzwertsatz anwenden.

Wir wenden die Proposition jetzt wieder auf den MLE im iid-Fall an.

Satz 15.9 *Seien $X_1, \ldots, X_n \overset{iid}{\sim} \mathbf{P}_\theta$ und $\mathcal{P} = \{\mathbf{P}_\theta | \theta \in \Theta\}$ eine Exponentialfamilie mit zweimal stetig differenzierbaren $\eta_j(\theta)$ und $C(\theta)$ und gelten die Annahmen (i), (ii) und (iv) aus Proposition 15.3 bzw. Proposition 15.8. Dann sind auch die Annahmen (v) und (vi) erfüllt mit $V = W = I(\theta_0) := \mathbf{E}_{\theta_0} \nabla \log f_{\theta_0}(X_1) \nabla \log f_{\theta_0}(X_1)'$ (Fisher-Informations-Matrix), d.h. es gilt für den Maximum-Likelihood-Schätzer*

$$\sqrt{n}\big(\widehat{\theta}_n - \theta_0\big) \xrightarrow{\mathcal{D}} \mathcal{N}\big(0, I(\theta_0)^{-1}\big).$$

Bemerkung. Als Anwendung dieses Ergebnisses kann man z.B. ein approximatives Konfidenzintervall für θ_0 konstruieren. Ein Beispiel dafür ist unten im Anhang zu finden.

Beweis. (v) folgt wie oben aus der Form von $\log f_\theta(x)$ in ($\triangle$):

$$\begin{aligned}
&\sup_{\theta \in \Theta} \big|\nabla^2_{k,\ell} L_n(\theta) - \nabla^2_{k,\ell} L(\theta)\big| \\
&\leq \sum_{j=1}^{s} \underbrace{\sup_{\theta \in \Theta} \big|\nabla^2_{k,\ell} \eta_j(\theta)\big|}_{\leq K < \infty} \Big|\frac{1}{n}\sum_{i=1}^{n} T_j(X_i) - \mathbf{E}_{\theta_0} T_j(X_1)\Big| + \Big|\frac{1}{n}\sum_{i=1}^{n} \log h(X_i) - \mathbf{E}_{\theta_0} \log h(X_1)\Big| \xrightarrow{P} 0.
\end{aligned}$$

(vi) Wegen $\theta_0 \in \mathrm{Int}(\Theta)$ folgt $0 = \nabla L(\theta_0) = -\nabla \mathbf{E}_{\theta_0} \log f_\theta(X_1)\big|_{\theta=\theta_0} \overset{\mathbf{(MT)}}{=} -\mathbf{E}_{\theta_0} \nabla \log f_{\theta_0}(X_1)$ und mit dem (multivariaten) zentralen Grenzwertsatz

$$\sqrt{n}\, \nabla L_n(\theta_0) = -\sqrt{n}\, \Big(\frac{1}{n} \sum_{i=1}^{n} \nabla \log f_{\theta_0}(X_i) - \mathbf{E}_{\theta_0} \nabla \log f_{\theta_0}(X_1) \Big) \xrightarrow{\mathcal{D}} \mathcal{N}(0, V)$$

mit $V = \mathbf{E}_{\theta_0} \nabla \log f_{\theta_0}(X_1) \nabla \log f_{\theta_0}(X_1)' = I(\theta_0)$. Den Beweis von $W = I(\theta_0)$ formulieren wir nur eindimensional - der multivariate Fall ist völlig analog. Es gilt

$$W = -\frac{\partial^2}{\partial \theta^2} \mathbf{E}_{\theta_0} \log f_\theta(X)\big|_{\theta=\theta_0} \overset{\mathbf{(MT)}}{=} -\mathbf{E}_{\theta_0} \frac{\partial^2}{\partial \theta^2} \log f_\theta(X)\big|_{\theta=\theta_0}.$$

Wegen

$$\frac{\partial^2}{\partial \theta^2} \log f_\theta(x) = \frac{\partial}{\partial \theta} \Big[f_\theta(x)^{-1} \frac{\partial}{\partial \theta} f_\theta(x) \Big] = \frac{\partial}{\partial \theta} f_\theta(x)^{-1} \frac{\partial}{\partial \theta} f_\theta(x) + f_\theta(x)^{-1} \frac{\partial^2}{\partial \theta^2} f_\theta(x)$$

und

$$\mathbf{E}_{\theta_0} f_\theta(X)^{-1} \frac{\partial^2}{\partial \theta^2} f_\theta(X)\big|_{\theta=\theta_0} = \int \frac{\partial^2}{\partial \theta^2} f_\theta(x)\, dx\big|_{\theta=\theta_0} \overset{\mathbf{(MT)}}{=} \frac{\partial^2}{\partial \theta^2} \underbrace{\int f_\theta(x)\, dx}_{=1}\big|_{\theta=\theta_0} = 0$$

[im stetigen Fall, im diskreten anolog mit Summe] folgt daraus

$$\begin{aligned} W &= -\mathbf{E}_{\theta_0} \frac{\partial^2}{\partial \theta^2} \log f_\theta(X)\big|_{\theta=\theta_0} = -\mathbf{E}_{\theta_0} \Big(f_\theta(X) \frac{\partial}{\partial \theta} f_\theta(X)^{-1} \Big) \Big(f_\theta(X)^{-1} \frac{\partial}{\partial \theta} f_\theta(X) \Big)\big|_{\theta=\theta_0} \\ &= \mathbf{E}_{\theta_0} \Big(\frac{\partial}{\partial \theta} \log f_\theta(X) \Big)^2\big|_{\theta=\theta_0} = I(\theta_0). \end{aligned}$$

[In allen drei Fällen folgt das $\overset{\mathbf{(MT)}}{=}$ aus dem Satz von der dominierten Konvergenz der Maßtheorie oder alternativ aus Vertauschungssätzen der Analysis]. □

Beispiel 15.10 Man kann nun leicht für verschiedene Verteilungen den MLE und $I(\theta_0)$ ausrechnen und so eine Reihe von Beispielen angeben. Im Hinblick auf eine spätere Anwendung berechnen wir $I(\theta_0)$ für Normalverteilungen. Aus $f_\theta(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\big(-\frac{1}{2\sigma^2}(x-\mu)^2 \big)$ folgt

$$\log f_\theta(x) = -\frac{1}{2} \log\big(2\pi\sigma^2\big) - \frac{1}{2\sigma^2} (x-\mu)^2$$

d.h.

$$\frac{\partial}{\partial\mu}\log f_\theta(x) = \frac{1}{\sigma^2}(x-\mu) \quad \Rightarrow \quad \mathbf{E}_\theta\Big(\frac{\partial}{\partial\mu}\log f_\theta(X)\Big)^2 = \frac{\mathbf{E}_\theta(X-\mu)^2}{\sigma^4} = \frac{1}{\sigma^2}$$

und

$$\begin{aligned}\frac{\partial}{\partial\sigma^2}\log f_\theta(x) &= -\frac{1}{2\sigma^2} + \frac{1}{2\sigma^4}(x-\mu)^2 \\ \Rightarrow\ \mathbf{E}_\theta\Big(\frac{\partial}{\partial\sigma^2}\log f_\theta(X)\Big)^2 &= \frac{1}{4\sigma^4} - \frac{1}{2\sigma^6}\sigma^2 + \frac{1}{4\sigma^8}\underbrace{\mathbf{E}_\theta(X-\mu)^4}_{=3\sigma^4 \text{ (vgl. Prop.13.5)}} = \frac{1}{2\sigma^4}.\end{aligned}$$

Ferner gilt $\mathbf{E}_\theta\Big(\frac{\partial}{\partial\mu}\log f_\theta(X)\frac{\partial}{\partial\sigma^2}\log f_\theta(X)\Big) = 0$, d.h. $I(\theta_0) = \begin{pmatrix} 1/\sigma^2 & 0 \\ 0 & 1/(2\sigma^4)\end{pmatrix}$ und damit für den MLE aus Bemerkung 15.1

$$\sqrt{n}\Big((\hat\mu_n, \hat\sigma_n^2) - (\mu,\sigma^2)\Big) \xrightarrow{\mathcal{D}} \mathcal{N}\left(0, \begin{pmatrix}\sigma^2 & 0 \\ 0 & 2\sigma^4\end{pmatrix}\right).$$

Bemerkung 15.11 (Cramér-Rao Ungleichung / Asympt. Effizienz des MLE)
Man kann zeigen, dass $I(\theta_0)^{-1}$ die kleinstmögliche asymptotische Varianz (Kovarianzmatrix) unter allen Schätzern für θ ist, d.h. der MLE ist in diesem Sinne optimal (effizient). Ein rigoroser Beweis dieser Aussage ist sehr schwierig und man braucht viele Annahmen. Wir leiten dieses Ergebnis nun heuristisch im Fall von iid-Beobachtungen mit stetiger Verteilung und $d = \dim\Theta = 1$ her. ["Heuristisch" bedeutet dabei, dass wir beliebig Integration und Differentiation vertauschen. Dieses folgt jeweils aus zusätzlichen Differenzierbarkeitsbedingungen - z.B. mit Sätzen der Maßtheorie (dominierte Konvergenz) oder der Analysis.]

(i) (Cramér-Rao Ungleichung) Sei $f_\theta^{(n)}(x) = \prod_{i=1}^n f_\theta(x_i)$. Für einen beliebigen Schätzer $S = S(X)$ von θ gilt

$$\begin{aligned}\mathbf{E}_\theta S &= \int S(x) f_\theta^{(n)}(x)\,dx \\ \Rightarrow\ \frac{\partial}{\partial\theta}\mathbf{E}_\theta S &= \int S(x)\frac{\partial}{\partial\theta}f_\theta^{(n)}(x)\,dx = \int S(x)\Big(\frac{\partial}{\partial\theta}\log f_\theta^{(n)}(x)\Big)f_\theta^{(n)}(x)\,dx.\end{aligned}$$

Ferner gilt

$$\begin{aligned} 1 &= \int f_\theta^{(n)}(x)\,dx \\ &\Rightarrow\ 0 = \int \frac{\partial}{\partial\theta} f_\theta^{(n)}(x)\,dx = \int \Big(\frac{\partial}{\partial\theta}\log f_\theta^{(n)}(x)\Big) f_\theta^{(n)}(x)\,dx \\ &\Rightarrow\ \frac{\partial}{\partial\theta}\,\mathbf{E}_\theta S = \int \Big(S(x) - \mathbf{E}_\theta S\Big)\Big(\frac{\partial}{\partial\theta}\log f_\theta^{(n)}(x)\Big) f_\theta^{(n)}(x)\,dx. \end{aligned}$$

Mit der Cauchy-Schwarz Ungleichung [zur Erinnerung: $\big(\int fg\big)^2 \le \big(\int f^2\big)\big(\int g^2\big)$] folgt daraus

$$\begin{aligned} \Big(\frac{\partial}{\partial\theta}\,\mathbf{E}_\theta S\Big)^2 &\le \int \Big(S(x) - \mathbf{E}_\theta S\Big)^2 f_\theta^{(n)}(x)\,dx \int \Big(\frac{\partial}{\partial\theta}\log f_\theta^{(n)}(x)\Big)^2 f_\theta^{(n)}(x)\,dx \\ &= \mathrm{Var}_\theta S \quad \mathrm{Var}_\theta\Big(\frac{\partial}{\partial\theta}\log f_\theta^{(n)}(X)\Big). \end{aligned}$$

Wegen

$$\mathrm{Var}_\theta\Big(\frac{\partial}{\partial\theta}\log f_\theta^{(n)}(X)\Big) = \mathrm{Var}_\theta\Big(\frac{\partial}{\partial\theta}\sum_{i=1}^n \log f_\theta(X_i)\Big) = n\mathrm{Var}_\theta\Big(\frac{\partial}{\partial\theta}\log f_\theta(X_1)\Big) = nI(\theta)$$

folgt insgesamt mit der Bezeichnung $b(\theta) := \mathbf{E}_\theta S - \theta$ (Bias)

$$\mathrm{Var}_\theta S \ge \frac{\big(1 + b'(\theta)\big)^2}{nI(\theta)} \qquad \text{(Cramér-Rao Ungleichung).}$$

Beispiel:

Seien $X_1, \ldots, X_n \overset{\text{iid}}{\sim} \mathcal{N}(\mu, \sigma^2)$. Für den Mittelwert gilt $b(\mu) = 0$ und $\mathrm{Var}_\theta \overline{X}_n = \frac{\sigma^2}{n} = \frac{1}{nI(\mu)}$ (vgl. obiges Beispiel), d.h. der Schätzer ist optimal [dass es in diesem Beispiel noch σ^2 als zweiten Parameter gibt, ist nicht problematisch - die Cramér-Rao Ungleichung gilt trotzdem in der obigen Form].
Für die empirische Stichproben-Varianz S^2 gilt $b(\sigma^2) = 0$ und nach Proposition 13.5 $\mathrm{Var}_\theta S^2 = \frac{2\sigma^4}{n-1} > \frac{2\sigma^4}{n} = \frac{1}{nI(\sigma^2)}$, d.h. der Schätzer ist nicht optimal. Ebenso ist der MLE nicht optimal [ein Schätzer für σ^2, der die Cramér-Rao Schranke annimmt, ist mir nicht bekannt].

(ii) (Asymptotische Effizienz) Angenommen für einen Schätzer S_n gilt ein zentraler Grenzwertsatz. d.h. $\sqrt{n}(S_n - \theta) \xrightarrow{\mathcal{D}} \mathcal{N}(0, v(\theta))$. Unter Zusatzbedingungen folgt daraus $n \operatorname{Var}_\theta S_n \to v(\theta)$. Außerdem gilt i.d.R. $b_n'(\theta) := \frac{\partial}{\partial\theta}\mathrm{Bias}(S_n) \to 0$. Daraus folgt dann

$$v(\theta) = \lim_{n\to\infty} n \operatorname{Var}_\theta S_n \geq \lim_{n\to\infty} n \frac{\big(1 + b_n{}'(\theta)\big)^2}{nI(\theta)} = I(\theta)^{-1}.$$

Falls $v(\theta) = I(\theta)^{-1}$ nennt man den Schätzer S_n deshalb asymptotisch (Fisher) effizient [auch im multivariaten Fall, wenn $v(\theta)$ und $I(\theta)$ Matrizen sind]. Aus Satz 15.9 folgt, dass der MLE asymptotisch effizient ist [aber i.d.R. nicht für festes n wie obiges Beispiel zeigt].

Bemerkung 15.12 (Kritische Würdigung des MLE) Die Grundidee des MLE mag auf den allerersten Blick überzeugend sein - bereits beim zweiten Blick stellt sich aber die Frage, warum das Maximum-Likelihood-Prinzip zu einem guten Schätzer führen soll. Das wurde bereits von Gauß (1839) in einem Brief an Bessel angemerkt. Darüberhinaus gibt es eine Reihe von Situationen, in denen der MLE nicht konsistent ist. Trotzdem ist der MLE ein weitverbreiteter Schätzer. Der Wert des MLE liegt vor allem darin, dass er in vielen komplexen Situationen (wie z.B. bei stochastischen Prozessen) unmittelbar zu einem Schätzer führt, der idR zumindest mit numerischen Verfahren auch berechnet werden kann. Man vertraut dann darauf, dass der MLE bei niedrig dimensionalen Parameterräumen und "glatten" Wahrscheinlichkeitsdichten gute Eigenschaften hat - dieses sollte man im konkreten Fall aber nachrechnen. Im folgenden geben wir ein Beispiel für den MLE in so einer komplexen Situation.

Beispiel 15.13 (Logistische Regression) Um die Toxidität einer Substanz zu testen, werden jeweils m_i Mäusen die Dosis z_i verabreicht ($i = 1, \ldots, n$). Sei X_i die Anzahl der im Experiment gestorbenen Tiere. Offensichtlich gilt

$$X_i \sim \mathcal{B}(m_i, \lambda_i), \quad \lambda_i = \lambda_\theta(z_i).$$

Als Modell für die Abhängigkeit von z_i verwenden wir

$$\lambda_\theta(z) = \frac{1}{1 + \exp\big(-(\theta_1 + \theta_2 z)\big)}.$$

Es gilt $\lim_{z\to\infty}\lambda_\theta(z) = 1$ und $\lambda_\theta(0) = \big[1+\exp(-\theta_1)\big]^{-1}$ [d.h. für hinreichend kleine (negative) θ_1 wird $\lambda_\theta(0)$ beliebig klein]. Es folgt

$$\frac{\lambda_\theta(z)}{1-\lambda_\theta(z)} = \frac{\frac{1}{1+\exp\big(-(\theta_1+\theta_2 z)\big)}}{1-\frac{1}{1+\exp\big(-(\theta_1+\theta_2 z)\big)}} = \exp(\theta_1+\theta_2 z)$$

und wir erhalten damit für die Dichte

$$\begin{aligned} &f_\theta(x_1,\ldots,x_n) \\ &\quad= \prod_{i=1}^n \binom{m_i}{x_i}\,\lambda_\theta(z_i)^{x_i}\,\big(1-\lambda_\theta(z_i)\big)^{m_i-x_i} = \prod_{i=1}^n \big(1-\lambda_\theta(z_i)\big)^{m_i}\,\prod_{i=1}^n e^{x_i(\theta_1+\theta_2 z_i)}\,\prod_{i=1}^n \binom{m_i}{x_i} \end{aligned}$$

d.h.

$$\begin{aligned} L_n(\theta) &= -\frac{1}{n}\log f_\theta(x_1,\ldots,x_n) \\ &= -\frac{1}{n}\sum_{i=1}^n m_i \log\big(1-\lambda_\theta(z_i)\big) \;-\; \frac{1}{n}\sum_{i=1}^n x_i(\theta_1+\theta_2 z_i) \;- h(x) \end{aligned}$$

mit

$$h(x) = \frac{1}{n}\sum_{i=1}^n \log\binom{m_i}{x_i}.$$

Zur Berechnung des MLEs suchen wir nach lokalen Minima durch Nullsetzen der partiellen Ableitungen. Es gilt

$$\frac{\partial}{\partial\theta_1}L_n(\theta) = -\frac{1}{n}\sum_{i=1}^n m_i\,\big(1-\lambda_\theta(z_i)\big)^{-1}\,\frac{\partial}{\partial\theta_1}\big(1-\lambda_\theta(z_i)\big) \;-\frac{1}{n}\sum_{i=1}^n x_i.$$

Wegen

$$1-\lambda_\theta(z_i) = \frac{\exp\big(-(\theta_1+\theta_2 z)\big)}{1+\exp\big(-(\theta_1+\theta_2 z)\big)} \quad\text{und}\quad \frac{\partial}{\partial\theta_1}\lambda_\theta(z_i) = \frac{\exp\big(-(\theta_1+\theta_2 z_i)\big)}{\big[1+\exp\big(-(\theta_1+\theta_2 z_i)\big)\big]^2}$$

folgt $\big(1-\lambda_\theta(z_i)\big)^{-1}\frac{\partial}{\partial\theta_1}\big(1-\lambda_\theta(z_i)\big) = -\,\lambda_\theta(z_i)$ und damit

$$\frac{\partial}{\partial\theta_1}L_n(\theta) = \frac{1}{n}\sum_{i=1}^n m_i\lambda_\theta(z_i) \;-\frac{1}{n}\sum_{i=1}^n x_i \stackrel{!}{=} 0.$$

Völlig analog erhält man

$$\frac{\partial}{\partial\theta_2}L_n(\theta) = \frac{1}{n}\sum_{i=1}^{n} m_i z_i \lambda_\theta(z_i) \; - \frac{1}{n}\sum_{i=1}^{n} x_i z_i \stackrel{!}{=} 0.$$

Die Lösung dieser beiden Gleichungen bzgl. $\theta = (\theta_1, \theta_2)$ ergibt den MLE. Die Gleichungen müssen numerisch gelöst werden, z.B. mit dem Newton-Verfahren. Die erste Gleichung entspricht im übrigen der MLE-Gleichung bei Binomialveteilungen: $\frac{1}{n}\sum_{i=1}^{n} x_i = m\lambda$.

Dieser Schätzer ist ein Beispiel für einen MLE in einer komplexeren nicht-iid Situation. Man kann auch hier Proposition 15.3 und Proposition 15.8 verwenden, um Konsistenz und asymptotische Normalität nachzuweisen, z.B. für $m_i = m$ fest und $n \to \infty$. Man braucht dann noch eine Annahme über das asymptotische Verhalten der Folge z_i. Dieses ist an dieser Stelle zu aufwendig [vor allem weil man nicht unmittelbar das Schwache Gesetz der großen Zahlen und den ZGWS in der Form des letzten Kapitels anwenden kann].

15.1 Anhang: Asymptotische Normalität im multivariaten Fall

15.14 (Beweis von Proposition 15.8 im multivariaten Fall) Der Beweis verläuft analog zum eindimensionalen Fall. Wir benötigen aber eine multivariate Version von Proposition 14.9 (iii), insbesondere die Aussagen

(a) $X_n \xrightarrow{\mathcal{D}} X$, $Y_n \xrightarrow{P} c \quad \Rightarrow X_n + Y_n \xrightarrow{\mathcal{D}} X + c$ (für d-dimensionale Vektoren)

(b) $X_n \xrightarrow{\mathcal{D}} X$, $Y_n \xrightarrow{P} C \quad \Rightarrow Y_n X_n \xrightarrow{\mathcal{D}} CX$ (Y_n, C $d \times d$ - Matrizen)
$Y_n^{-1} X_n \xrightarrow{\mathcal{D}} C^{-1}X$, falls C regulär

[c, C sind Konstante]. Das erste Problem bei der Übertragung des Beweises ist, dass der Mittelwertsatz nicht für vektorwertige Funktionen gilt, sondern nur für $f : \mathbb{R}^d \to \mathbb{R}$, d.h.

wir erhalten

$$\sqrt{n}\,\nabla_i L_n(\widehat{\theta}_n) - \sqrt{n},\nabla_i L_n(\theta_0) = \sum_{j=1}^{d} \nabla^2_{i,j} L_n(\theta_n^{(i)})\ \sqrt{n}\,(\widehat{\theta}_n - \theta_0)_j$$

mit $|\theta_n^{(i)} - \theta_0| \le |\widehat{\theta}_n - \theta_0|$ $(i = 1, ..., p)$. Wie im eindimensionalen Fall gilt $\mathbf{P}(\sqrt{n}|\nabla_i L_n(\widehat{\theta}_n)| \ge \varepsilon) \le \mathbf{P}(|\widehat{\theta}_n - \theta_0| \ge \delta) \to 0$ für alle $\varepsilon > 0$ und damit

$$W_n\ \sqrt{n}\,(\widehat{\theta}_n - \theta_0) \xrightarrow{\mathcal{D}} \mathcal{N}(0, V)$$

mit der (stochastischen) Matrix $W_n := \nabla^2_{i,j} L_n(\theta_n^{(i)})_{i,j=1....,d}$. Wie im eindimensionalen Fall folgt aus (v) und der Stetigkeit von $\nabla^2 L(\theta)$ jetzt

$$W_{n\,i,j} - W_{i,j} = \left(\nabla^2_{i,j} L_n(\theta_n^{(i)}) - \nabla^2_{i,j} L(\theta_n^{(i)})\right) + \left(\nabla^2_{i,j} L(\theta_n^{(i)}) - \nabla^2_{i,j} L(\theta_0)\right) \xrightarrow{P} 0,$$

d.h. es gilt $W_n \xrightarrow{P} W$. Aus (b) oben folgt dann die Behauptung.

16 Bedingte Verteilungen und bedingte Erwartungswerte

In diesem Kapitel werden bedingte Verteilungen und bedingte Erwartungswerte definiert. Der bedingte Erwartungswert kann als Zufallsvariable aufgefasst werden, deren Eigenschaften wir untersuchen. Als Anwendung betrachten wir den besten Prädiktor und den besten linearen Prädiktor von einer Zufallsvariablen. Wir berechnen die bedingte Verteilung bei Normalverteilungen und zeigen, dass dort der beste Prädiktor und der beste lineare Prädiktor übereinstimmen.

16.1 Vorbetrachtung

Häufig interessiert man sich für die Verteilung einer Zufallsvariablen X, wenn der Wert einer anderen Zufallsvariablen Y bekannt ist. Beispiel:

- X Anzahl der Buben im Skat, Y Anzahl der Buben im eigenen Blatt. Gesucht: $\mathbf{P}(X = x \mid Y = y)$.
- X Gewicht, Y Größe. Gesucht: $\mathbf{P}(X \in B \mid Y = y)$.

Im diskreten Fall gilt für $p_Y(y) > 0$

$$\mathbf{P}(X = x \mid Y = y) = \frac{\mathbf{P}(X = x, Y = y)}{\mathbf{P}(Y = y)} = \frac{p_{XY}(x, y)}{p_Y(y)}$$

und im stetigen Fall (Problem: $\mathbf{P}(Y = y) = 0$!!)

$$\begin{aligned}
\mathbf{P}(X \in B \mid Y = y) &\approx \mathbf{P}\big(X \in B \mid Y \in [y, y + \Delta y]\big) \qquad \Delta y \text{ klein} \\
&= \frac{\mathbf{P}\big(X \in B, Y \in [y, y + \Delta y]\big)}{\mathbf{P}\big(Y \in [y, y + \Delta y]\big)} \\
&= \frac{\int_B \int_y^{y+\Delta y} f_{XY}(x, z)\, dz dx}{\int_y^{y+\Delta y} f_Y(z)\, dz} \\
&\approx \frac{\int_B f_{XY}(x, y)\, \Delta y\, dx}{f_Y(y)\, \Delta y} \\
&= \frac{\int_B f_{XY}(x, y)\, dx}{f_Y(y)}
\end{aligned}$$

Satz / Definition 16.2 (Bedingte diskrete Verteilungen)
Seien X und Y gemeinsam diskret verteilt mit Zähldichte $p_{XY}(x, y)$, sowie

$$p_{X|Y=y}(x) := \begin{cases} \frac{p_{XY}(x,y)}{p_Y(y)} & , \text{ falls } p_Y(y) > 0 \\ 0 & , \text{ sonst.} \end{cases}$$

Dann gilt $p_{X|Y=y}(x) \geq 0$, $\sum_x p_{X|Y=y}(x) = 1 \; \forall y$ mit $p_Y(y) > 0$ und

$$p_X(x) = \sum_y p_{X|Y=y}(x)\, p_Y(y).$$

$p_{X|Y=y}(x)$ heißt bedingte Zähldichte von X gegeben $Y = y$.
$\mathbf{E}(X \mid Y = y) := \sum_x x\, p_{X|Y=y}(x)$ heißt bedingter Erwartungswert von X gegeben $Y = y$.

Beweis. trivial. □

Beispiel 16.3 Dreimaliges Werfen einer Münze

$\Omega = \big\{(\omega_1, \omega_2, \omega_3) \mid \omega_i \in \{0, 1\}\big\}$; $\quad X(\omega) = \omega_1$; $\quad Y(\omega) = \sum_{i=1}^3 \omega_i$.

$p_{XY}(x_i, y_j)$:

x_i \ y_j	0	1	2	3
0	1/8	2/8	1/8	0
1	0	1/8	2/8	1/8

Hiermit erhält man

$$\begin{aligned} p_{X\,|\,Y=1}(0) &= \frac{p_{XY}(0,1)}{p_Y(1)} = \frac{\frac{2}{8}}{\frac{2}{8}+\frac{1}{8}} = \frac{2}{3}; \\ p_{X\,|\,Y=1}(1) &= \frac{p_{XY}(1,1)}{p_Y(1)} = \frac{\frac{1}{8}}{\frac{3}{8}} = \frac{1}{3}; \\ \mathbf{E}(X|Y=1) &= 0 \cdot \frac{2}{3} + 1 \cdot \frac{1}{3} = \frac{1}{3}. \end{aligned}$$

□

Beispiel 16.4 (Geigerzähler)

N Anzahl der vom Geigerzähler aufgenommenen Teilchen pro Zeiteinheit (Sekunde).
X Anzahl der gezählten Teilchen.
p Wahrscheinlichkeit, dass ein aufgenommenes Teilchen gezählt wird.

Die Ereignisse "das i-te Teilchen wird gezählt" sind unabhängig voneinander, d.h.

$$p_{X|N=n}(k) = \mathbf{P}(X = k \mid N = n) = \binom{n}{k} p^k \, (1-p)^{n-k}.$$

Annahme: $N \sim \mathcal{P}(\lambda)$, d.h. $p_N(n) = e^{-\lambda}\frac{\lambda^n}{n!}$. $\quad X \sim ?$

$$\begin{aligned}
\mathbf{P}(X = k) &= \sum_{n=0}^{\infty} p_{X|N=n}(k)\, p_N(n) \\
&= \sum_{n=0}^{\infty} \binom{n}{k} p^k (1-p)^{n-k}\, e^{-\lambda}\frac{\lambda^n}{n!} \\
&= e^{-\lambda}\, p^k\, \frac{1}{k!}\, \lambda^k \underbrace{\sum_{n=k}^{\infty} \frac{\lambda^{n-k}}{(n-k)!}(1-p)^{n-k}}_{e^{\lambda(1-p)}} \\
&= e^{-p\lambda}\frac{(p\lambda)^k}{k!}
\end{aligned}$$

$$\Rightarrow X \sim \mathcal{P}(p\lambda)$$ □

Satz / Definition 16.5 (Bedingte stetige Verteilungen)
Seien $X, Y : \Omega \to \mathbb{R}$ gemeinsam stetig verteilt mit Wahrscheinlichkeitsdichte $f_{XY}(x, y)$, sowie

$$f_{X|Y=y}(x) = \begin{cases} f_{XY}(x,y)/f_Y(y)\,, & \text{falls } f_Y(y) > 0 \\ 0\,, & \text{sonst.} \end{cases}$$

Dann gilt $f_{X|Y=y}(x) \geq 0$, $\int_{\mathbb{R}} f_{X|Y=y}(x)\, dx = 1\ \forall y$ mit $f_Y(y) > 0$ und

$$f_X(x) = \int_{\mathbb{R}} f_{X|Y=y}(x)\, f_Y(y)\, dy.$$

$f_{X|Y=y}(x)$ heißt bedingte W'dichte von X gegeben $Y = y$.
$\mathbf{E}(X \mid Y = y) := \int_{\mathbb{R}} x f_{X|Y=y}(x)\, dx$ heißt bedingter Erwartungswert von X gegeben $Y = y$ (falls $\int |x| f_{X|Y=y}(x)\, dx < \infty$).

Beweis. trivial. □

Bemerkung 16.6 Die Aussagen von Satz 16.2 bzw. 16.5 besagen unter anderem, dass $p_{X|Y=y}(x)$ für diejenigen y, für die $p_Y(y) > 0$ gilt, eine Verteilung, nämlich die bedingte

Verteilung von X gegeben $Y = y$, definiert [analog im stetigen Fall, falls $f_Y(y) > 0$]. Wegen

$$\mathbf{P}\big(\underbrace{\{\omega \mid p_Y(Y(\omega)) = 0\}}_{=:N}\big) = \mathbf{P}^Y\big(\{y \mid p_Y(y) = 0\}\big) = \sum_{\substack{y \\ p_Y(y)=0}} p_Y(y) = 0$$

ist dieses für alle $y = Y(\omega)$ außerhalb einer Menge N mit $\mathbf{P}(N) = 0$ richtig. Man sagt: die Aussage gilt $\mathbf{P}$-f.s. ($\mathbf{P}$-fast sicher). [Analog im stetigen Fall.] □

Mit der Definition $\mathbf{E}(X \mid Y)(\omega) := \mathbf{E}(X \mid Y = y)$ wobei $y = Y(w)$, ist $\mathbf{E}(X \mid Y) : \Omega \to \mathbb{R}$ ebenfalls eine Zufallsvariable.

Satz 16.7 (Eigenschaften von $\mathbf{E}(X \mid Y)$) *Gelte $\mathbf{E}|X| < \infty$. Dann folgt*

(i) $\mathbf{E}\big(\mathbf{E}(X \mid Y)\big) = \mathbf{E}X$;

(ii) X, Y *stoch. unabh.* $\Rightarrow$ $\mathbf{E}(X \mid Y) = \mathbf{E}X$ $\mathbf{P}$*-f.s.;*

(iii) $\mathbf{E}(X \cdot h(Y)|Y) = h(Y)\,\mathbf{E}(X \mid Y)$ *für jede (messbare) Funktion h, d.h. insbesondere*
- $\mathbf{E}(1 \mid Y) = 1$ *(folgt aus (ii))*
- $\mathbf{E}(h(Y) \mid Y) = h(Y)$

(iv) $\mathbf{E}(\alpha X + \beta Y \mid Z) = \alpha\,\mathbf{E}(X \mid Z) + \beta\,\mathbf{E}(Y \mid Z)$ *(Linearität)*

Beweis. Seien X, Y diskret.

(i)

$$\begin{aligned}
\mathbf{E}\big(\mathbf{E}(X \mid Y)\big) &= \sum_{j=1}^{\infty} \mathbf{E}(X \mid Y = y_j)\, p_Y(y_j) \\
&= \sum_{j=1}^{\infty} \sum_{i=1}^{\infty} x_i \underbrace{p_{X|Y=y_j}(x_i)\, p_Y(y_j)}_{p_{XY}(x_i,y_j)} \\
&= \mathbf{E}X.
\end{aligned}$$

(ii)

$$\mathbf{E}(X \mid Y = y) = \sum_{j=1}^{\infty} x_j \underbrace{p_{X|Y=y}(x_j)}_{p_X(x_j)} \text{ , falls } p_Y(y) > 0$$
$$= \mathbf{E}X.$$

(iii) Sei $\omega \in \Omega$ und $y := Y(\omega)$. Dann gilt (falls $p_Y(y) > 0$)

$$\mathbf{E}\big(X \cdot h(Y) \mid Y\big)(\omega) = \mathbf{E}\big(X \cdot h(Y) \mid Y = y\big) = \underset{\substack{\uparrow \\ \text{konstant}}}{h(y)}\ \mathbf{E}(X \mid Y = y)$$
$$= h\big(Y(\omega)\big)\, \mathbf{E}(X \mid Y)\,(\omega).$$

(iv) Nachrechnen.

Beweis im stetigen Fall analog (Ü-Aufgabe). □

Bemerkung 16.8 Die Sätze 16.2, 16.5 und 16.7 gelten analog auch für Zufallsvektoren X und Y.

Beispiel 16.9 (Zufallssummen) Sei X_i die Höhe des i-ten Schadens bei einer Versicherung (X_i iid) und N die Anzahl der Schadensfälle in einem Jahr (N sei unabhängig von X_i).

$$\Rightarrow \text{ Jahresschaden } S = \textstyle\sum_{i=1}^{N} X_i$$

Gesucht: $\mathbf{E}S$, $\mathrm{Var}S$.

Es gilt:

$$\mathbf{E}(S \mid N = n) = \mathbf{E}\Big(\sum_{i=1}^{N} X_i \,\Big|\, N = n\Big) = \sum_{i=1}^{n} \mathbf{E}X_i = n\, \mathbf{E}X_i$$
$$\Rightarrow\ \mathbf{E}(S \mid N) = N\, \mathbf{E}X_i$$
$$\Rightarrow\ \mathbf{E}(S) = \mathbf{E}\big(\mathbf{E}(S \mid N)\big) = \mathbf{E}N\, \mathbf{E}X_i\,;$$

$$
\begin{aligned}
\mathbf{E}(S^2 \mid N=n) &= \mathbf{E}\Big(\big(\sum_{i=1}^{N} X_i\big)^2 \;\Big|\; N=n \Big) = \mathbf{E}\Big(\big(\sum_{i=1}^{n} X_i\big)^2 \Big) \\
&= \operatorname{Var}\Big(\sum_{i=1}^{n} X_i\Big) + \Big(\mathbf{E}\big(\sum_{i=1}^{n} X_i\big)\Big)^2 \\
&= n \operatorname{Var}(X_1) + n^2 (\mathbf{E}X_1)^2 \\
\Rightarrow \qquad \mathbf{E}(S^2 \mid N) &= N \operatorname{Var}(X_1) + N^2 (\mathbf{E}X_1)^2 \\
\Rightarrow \qquad \mathbf{E}S^2 &= \mathbf{E}N \operatorname{Var}(X_1) + \mathbf{E}N^2 (\mathbf{E}X_1)^2 \\
\Rightarrow \qquad \operatorname{Var} S &= \mathbf{E}S^2 - (\mathbf{E}S)^2 \\
&= \mathbf{E}N \operatorname{Var}(X_1) + \operatorname{Var}N \, (\mathbf{E}X_1)^2 .
\end{aligned}
$$

Vergleich mit fester Summe $S_n = \sum_{i=1}^{n} X_i$:

$$\operatorname{Var} S_n = n \operatorname{Var}(X_1),$$

d.h. zu der Streuung der X_i kommt noch die Streuung von N. □

16.10 Vorhersage von Zufallsvariablen / bester Prädiktor

Seien X, Y ZVAs. Ziel: Vorhersage von Y aus X.

Beispiel:

Y:	Holzvolumen eines Baumes,	X:	Umfang des Stammes;
Y:	Baukosten eines Hauses,	X:	Wohnfläche.

Gesucht: Funktion $h(\cdot)$ die den 'mean squared error' (MSE) $\mathbf{E}\big(Y - h(X)\big)^2$ minimiert. Es gilt

$$
\begin{aligned}
\mathbf{E}\big(Y - h(X)\big)^2 &= \mathbf{E}\Big[\big(Y - \mathbf{E}(Y \mid X)\big) + \big(\mathbf{E}(Y \mid X) - h(X)\big)\Big]^2 \qquad (\Delta) \\
&= \mathbf{E}\big(Y - \underset{\geq 0}{\mathbf{E}}(Y|X)\big)^2 + 2 \underbrace{\mathbf{E}\Big[\big(Y - \mathbf{E}(Y|X)\big)\big(\mathbf{E}(Y|X) - h(X)\big)\Big]}_{(*)} + \underset{\geq 0}{\mathbf{E}\big(\mathbf{E}(Y|X) - h(X)\big)^2}
\end{aligned}
$$

$$
\begin{aligned}
(*) \quad &= \quad \mathbf{E}\Big\{\mathbf{E}\Big[\big(Y-\mathbf{E}(Y|X)\big)\big(\mathbf{E}(Y|X)-h(X)\big)\;\Big|\;X\Big]\Big\} \\
&\overset{\text{Satz 16.7 (iii)}}{=} \quad \mathbf{E}\Big\{\big(\mathbf{E}(Y|X)-h(X)\big)\underbrace{\mathbf{E}\big[Y-\mathbf{E}(Y|X)\;\big|\;X\big]}_{=\mathbf{E}(Y|X)-\mathbf{E}(Y|X)=0}\Big\} \\
&= \quad 0
\end{aligned}
$$

$\Rightarrow (\Delta)$ ist minimal für $h(X) = \mathbf{E}(Y \mid X)$.

Das Gleiche gilt, wenn X ein Zufallsvektor ist. □

16.11 Vorhersage von Zufallsvariablen / bester linearer Prädiktor

Seien Y, X ZVAs. Gesucht: a, b mit $\mathbf{E}\big[Y-(a+bX)\big]^2 = \min$. Es gilt

$$
\begin{aligned}
\mathbf{E}(Y-a-bX)^2 &= \mathrm{Var}(Y-a-bX) + (\mathbf{E}Y - a - b\,\mathbf{E}X)^2 \\
&= \mathrm{Var}\underset{\geq 0}{(Y-bX)} + (\mathbf{E}Y - \underset{\geq 0}{a} - b\,\mathbf{E}X)^2 \\
\Rightarrow\; a &= \mathbf{E}Y - b\,\mathbf{E}X.
\end{aligned}
$$

1. Term:

$$
\begin{aligned}
\mathrm{Var}(Y-bX) &= \mathrm{Var}Y + b^2\,\mathrm{Var}X - 2\,b\,\mathrm{Kov}(X,Y) \\
\tfrac{\partial}{\partial b}\mathrm{Var}(Y-bX) &= 2\,b\,\mathrm{Var}X - 2\,\mathrm{Kov}(X,Y) = 0 \\
\Leftrightarrow\; b &= \tfrac{\mathrm{Kov}(X,Y)}{\mathrm{Var}X}
\end{aligned}
$$

$$
\text{d.h. } b = \tfrac{\mathrm{Kov}(X,Y)}{\mathrm{Var}X}, \quad a = \mathbf{E}Y - \tfrac{\mathrm{Kov}(X,Y)}{\mathrm{Var}X}\,\mathbf{E}X.
$$

Prognosefehler:

$$
\mathbf{E}(Y-a-bX)^2 = \mathrm{Var}Y - \frac{\mathrm{Kov}(X,Y)^2}{\mathrm{Var}X} = \mathrm{Var}Y\,\big[1-\rho(X,Y)^2\big],
$$

d.h. der Prognosefehler ist klein, falls ρ nahe bei ± 1 ist (plausibel!).

Der beste lineare Prädiktor ist i.d.R. einfacher zu berechnen.

Ist X ein Zufallsvektor, so folgt völlig analog für den besten linearen Prädiktor $b = \Sigma(X)^{-1}\Sigma(X,Y)$ und $a = \mathbf{E}Y - b'\,\mathbf{E}X$. □

Beispiel 16.12

$Y = X^2 + X + Z, \qquad X, Z \overset{\text{iid}}{\sim} \mathcal{N}(0,1), \qquad$ gesucht: Prognose von Y gegeben X.

- Bester Prädiktor:

$$\mathbf{E}(Y \mid X) = X^2 + X + \mathbf{E}(Z \mid X) = X^2 + X\,;$$

MSE:

$$\mathbf{E}(Y - X^2 - X) = \mathbf{E}Z^2 = 1\,.$$

- Bester linearer Prädiktor:

$a + bX$ mit

$$b = \frac{\mathrm{Kov}(X,Y)}{\mathrm{Var}X} = \mathrm{Kov}(X, X^2 + X) = \mathbf{E}X^3 - \mathbf{E}X^2\mathbf{E}X + \mathbf{E}X^2 - (\mathbf{E}X)^2 = 1,$$
$$a = \mathbf{E}Y - \mathbf{E}X = \mathbf{E}X^2 = 1,$$

d.h. $1 + X$ ist bester linearer Prädiktor.

MSE:

$$\begin{aligned}\mathbf{E}\big(Y - (1+X)\big)^2 = \mathbf{E}\big(X^2 + Z - 1\big)^2 &= \mathbf{E}X^4 + 2\mathbf{E}X^2Z + \mathbf{E}Z^2 - 2\mathbf{E}X^2 - 2\mathbf{E}Z + 1\\ &= \mathbf{E}X^4 + \mathbf{E}Z^2 - 2\,\mathbf{E}X^2 + 1\\ &= 3 + 1 - 2 + 1 = 3.\end{aligned}$$

[$\mathbf{E}X^4 = 3$ wurde im Beweis von Proposition 13.5 hergeleitet]. □

16.1 Anhang: Die bedingte Verteilung bei Normalverteilungen

Als Ergänzung wollen wir die bedingte Verteilung bei Normalverteilungen explizit ausrechnen. Im Anschluss daran werden wir daraus folgern, dass der beste Prädiktor und der beste lineare Prädiktor bei Normalverteilungen identisch sind.

Satz 16.13 *(i) Seien Y und X Zufallsvektoren mit*

$$\begin{pmatrix} Y \\ X \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} \mu_Y \\ \mu_X \end{pmatrix}, \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix} \right).$$

Gilt $|\Sigma| > 0$, so ist die bedingte Verteilung von Y gegeben $X = x$

$$\mathcal{N}\big(\mu_Y + \Sigma_{12}\Sigma_{22}^{-1}(x - \mu_X), \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}\big).$$

(ii) Der beste Prädiktor von Y gegeben $X = x$ ist $\mathbf{E}(Y|X = x) = \mu_Y + \Sigma_{12}\Sigma_{22}^{-1}(x - \mu_X)$, d.h. der beste Prädiktor und der beste lineare Prädiktor stimmen überein.

Bemerkung:
Wegen $\mathbf{E}(Y|X) = \mu_Y + \Sigma_{12}\Sigma_{22}^{-1}(X - \mu_X) = a + b'X$ mit $b = \Sigma(X)^{-1}\Sigma(X, Y)$ und $a = \mathbf{E}Y - b'\,\mathbf{E}X$ stimmt der Prädiktor mit dem in Beispiel 9 hergeleiteten besten linearen Prädiktor (auch in der Form) überein.

Zum Beweis des Satzes benötigt man folgende Aussage, die man durch einfaches Nachrechnen von $\Sigma\,\Sigma^{-1} = \Sigma^{-1}\Sigma = I$ beweist:

Lemma

Ist $\Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}$ regulär, so gilt

$$\Sigma^{-1} = \begin{pmatrix} E^{-1} & -E^{-1}F \\ -F'E^{-1} & \Sigma_{22}^{-1} + F'\,E^{-1}\,F \end{pmatrix}$$

mit

$$E = \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21} \qquad \text{und} \qquad F = \Sigma_{12}\Sigma_{22}^{-1}.$$

Beweis von Satz 16.13 (i) Die bedingte Dichte von Y gegeben $X = x$ ist mit $z = \binom{y}{x}$ und $\mu = \binom{\mu_Y}{\mu_X}$

$$\begin{aligned}
\frac{f_{Y,X}(y,x)}{f_X(x)} &= c \exp\Big\{ -\frac{1}{2}(z-\mu)' \Sigma^{-1}(z-\mu) + \frac{1}{2}(x-\mu_X)' \Sigma_{22}^{-1}(x-\mu_X) \Big\} \\
&= c \exp\left\{ -\frac{1}{2}(z-\mu)' \begin{pmatrix} E^{-1} & -E^{-1}F \\ -F'E^{-1} & F'\,E^{-1}\,F \end{pmatrix} (z-\mu) \right\} \\
&= c \exp\left\{ -\frac{1}{2} \begin{pmatrix} y-\mu_Y \\ -F(x-\mu_X) \end{pmatrix}' \begin{pmatrix} E^{-1} & E^{-1} \\ E^{-1} & E^{-1} \end{pmatrix} \begin{pmatrix} y-\mu_Y \\ -F(x-\mu_X) \end{pmatrix} \right\} \\
&= c \exp\Big\{ -\frac{1}{2} \Big(y - [\mu_Y + F(x-\mu_X)] \Big)' E^{-1} \Big(y - [\mu_Y + F(x-\mu_X)] \Big) \Big\}.
\end{aligned}$$

Daraus folgt die Behauptung (die Konstante ergibt sich zwangsläufig durch die Normierung). [Es sei noch bemerkt, dass man aus der Linearen Algebra weiß, dass aus $|\Sigma| > 0$ auch $|\Sigma_{22}| > 0$ folgt.]

(ii) Die Aussage folgt, da der bedingte Erwartungswert der Erwartungswert der bedingten Verteilung ist. □

17 Varianz- und Regressionsanalyse / Das lineare Modell

In diesem Kapitel wird das lineare Modell eingeführt. Es ist der mathematische Rahmen, in dem die Regressions- und die Varianzanalyse behandelt werden. Es werden elementare Eigenschaften des kleinste-Quadrate-Schätzers diskutiert. Ferner wird gezeigt, dass die Theorie weitgehend durch lineare Projektionen des Beobachtungsvektors auf den Raum der Regressionsvariablen beschrieben werden kann. Zum Schluss wird die Optimalität des Gauß-Markov-Schätzers gezeigt.

17.1 Grundlagen

Gegeben seien Beobachtungen $Y = (Y_1, \ldots, Y_n)'$ und weitere (erklärende) Variable $X_i = (X_{i1}, \ldots, X_{in})'$ (Regressorvariable).

Lineares Modell:

Das lineare Modell lautet dann mit $X = (X_1, \ldots, X_k)$ $\big($häufig ist $X_1 = (1, \ldots, 1)'\big)$:

Y	$=$	X	β	$+$	ε
$n \times 1$		$n \times k$	$k \times 1$		$n \times 1$
Beobachtung		Designmatrix	Parametervektor		Fehlervektor
stochastisch		deterministisch	deterministisch		stochastisch
bekannt		bekannt	unbekannt		unbekannt

Alternativ wird X manchmal auch als stochastisch angenommen.

Annahmen:

Die ε_i seien iid verteilt mit $\mathbf{E}(\varepsilon_i) = 0,\ \mathrm{Var}(\varepsilon_i) = \sigma^2$ [manchmal auch $\varepsilon_i \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma^2)$].

Ziele: (z.B.)

- Prognose von Y für weitere Werte von X;
- Testen spezieller Hypothesen bzgl. β.

Beispiele:

(i) Y_i sind die Gewinne eines Unternehmens und X_i sind verschiedene erklärende Variable wie zum Beispiel der Umsatz, die Anzahl der Beschäftigten, verschiedene Konjunkturdaten, usw. Ziel ist es, die funktionale Form der Regressionsgeraden zu schätzen (Schätzung von β) oder z.B. eine Prognose des Umsatzes für veränderte Konjunkturdaten durchzuführen. Dabei interessiert auch die Genauigkeit der Prognose (Konfidenzintervall). Diese Fragen werden in diesem und dem nächsten Kapitel untersucht.

(ii) Modellierung eines zeitlichen Verlaufs:
Y_i ist der Umsatz am Tag i und $X_{ij} = (X_j)_i = (i)^{j-1}$. Modell:

$$Y_i = \beta_1 + \beta_2 i + \beta_3 i^2 + \beta_4 i^3 + \varepsilon_i.$$

(iii) Zweifache Varianzanalyse: Y_{ijk} sei die Anzahl der an Maschine $i \in \{1, \ldots, I\}$ von Arbeiter $j \in \{1, \ldots, J\}$ am Tag $k \in \{1, \ldots, K\}$ produzierten Stücke. Modell:

$$Y_{ijk} = \mu_{ij} + \varepsilon_{ijk}$$

[man kann das Modell in der Form $Y = X\beta + \varepsilon$ schreiben!]. Von Bedeutung ist z.B. ein Test, ob die Effekte additiv sind, d.h. der Hypothese

$$\mu_{ij} = \alpha_i + \gamma_j$$

(falls man diese Hypothese ablehnt, würde es Sinn machen, die einzelnen Arbeiter bestimmten Maschinen zuzuordnen, um die Produktivität zu erhöhen). Eine andere mögliche Hypothese ist, dass Arbeiter 1 und Arbeiter 2 gemittelt über alle Maschinen gleich produktiv sind, d.h.

$$\frac{1}{I}\sum_{i=1}^{I} \mu_{i1} = \frac{1}{I}\sum_{i=1}^{I} \mu_{i2}.$$

Ziel: Vordergründig ist oft das Ziel, β zu schätzen. In fast allen Fällen sucht man aber eigentlich nach einem $\widehat{\beta}$, so dass $X\widehat{\beta}$ eine gute Prognose von Y aus X liefert.

Kleinste Quadrate Kriterium: Minimiere

$$R^2(\beta) := (Y - X\beta)'(Y - X\beta) = Y'Y - 2Y'X\beta + \beta'X'X\beta$$

[das ist die Summe der Fehlerquadrate]. Es gilt

$$\nabla R^2(\beta) = -2X'Y + 2X'X\beta \stackrel{!}{=} 0 \quad \Leftrightarrow \quad X'X\beta = X'Y.$$

[d.h. $\widehat{\beta} = (X'X)^{-1}X'Y$ falls $X'X$ regulär ist, d.h. falls Rang$X = k$].

$\widehat{\beta}$ heißt KQ-Schätzer (kleinste Quadrate-Schätzer).

Projektionen: Als Prognose von Y aus X ergibt sich damit $X\widehat{\beta}$. Wir werden im folgenden zeigen, dass dies die (Orthogonal-) Projektion von Y auf den von den Spalten von X aufgespannten Raum ist. Sei dafür $\mathcal{X} = \text{Lin}(X)$ die lineare Hülle des Spaltenraums von X mit $r = \dim \mathcal{X} = \text{Rang}X$ und $P_{\mathcal{X}}$ die Projektionsmatrix auf $\mathcal{X}$. Dann ist $P_{\mathcal{X}}$ symmetrisch mit

$$P_{\mathcal{X}}^2 = P_{\mathcal{X}} \quad \text{und} \quad P_{\mathcal{X}}X = X.$$

Ferner gilt

$$(I - P_{\mathcal{X}})^2 = I - 2P_{\mathcal{X}} + P_{\mathcal{X}}^2 = (I - P_{\mathcal{X}}) \quad \text{und} \quad (I - P_{\mathcal{X}})X = 0,$$

d.h. wir haben die Orthogonal-Zerlegung

$$\mathbb{R}^n = \mathcal{X} \oplus \mathcal{X}_\perp \quad \text{mit} \quad P_{\mathcal{X}_\perp} = I - P_{\mathcal{X}}.$$

Satz 17.2 *Folgende Aussagen sind äquivalent:*

(i) $R^2(\beta)$ *wird minimal für* $\widehat{\beta}$*;*

(ii) $X\widehat{\beta} = P_{\mathcal{X}}Y$*;*

(iii) $X'X\widehat{\beta} = X'Y$ *(Normalengleichung).*

Im Fall $r = \text{Rang}X = k$ *ist* $\widehat{\beta}$ *eindeutig bestimmt und es gilt*

$$\widehat{\beta} = (X'X)^{-1}X'Y \quad \textit{und} \quad P_{\mathcal{X}} = X(X'X)^{-1}X'.$$

Beweis. Sei $P = P_{\mathcal{X}}$. Es gilt

$$\begin{aligned}(Y - PY)'(PY - X\beta) &= Y'PY - Y'P'PY - Y'X\beta + Y'P'X\beta = 0\\ \Rightarrow \quad \|Y - X\beta\|^2 &= \|Y - PY + PY - X\beta\|^2 = \|Y - PY\|^2 + \|PY - X\beta\|^2,\end{aligned}$$

d.h.

$$R^2(\beta) \text{ minimal für } \widehat{\beta} \;\Leftrightarrow\; PY = X\widehat{\beta} \;\Leftrightarrow\; Y - X\widehat{\beta} = (I - P)Y \;\Rightarrow\; X'(Y - X\widehat{\beta}) = 0$$

und damit (i) $\Leftrightarrow$ (ii) $\Rightarrow$ (iii). (ii) $\Leftrightarrow$ (iii) wurde schon oben bewiesen. Die Form von $\widehat{\beta}$ und P im Fall $\text{Rang}X = k$, folgt aus (iii) und (ii). □

Bemerkung: Ist $\text{Rang}X < k$, so hat die Gleichung $X'X\widehat{\beta} = X'Y$ mehrere Lösungen. Die Projektion auf $\mathcal{X}$, d.h. $P_{\mathcal{X}}Y = X\widehat{\beta}$, ist aber immer eindeutig.

Bemerkung 17.3 (Deterministische Eigenschaften von $\widehat{Y} := X\widehat{\beta} = P_{\mathcal{X}}Y$)
Seien $\widehat{\varepsilon} := Y - \widehat{Y} = Y - X\widehat{\beta} = (I - P_{\mathcal{X}})Y$ die geschätzten Residuen.

(i) Es gilt $\widehat{\varepsilon} \perp \text{Lin}(X)$, d.h. insbesondere $\widehat{\varepsilon} \perp \widehat{Y}$.

$$\Rightarrow \quad \frac{1}{n}\sum_{j=1}^{n} \widehat{\varepsilon}_j \widehat{Y}_j = 0, \quad \text{d.h. } \widehat{\varepsilon} \text{ und } \widehat{Y} \text{ sind (empirisch) unkorreliert.}$$

(ii) Ist $\mathbf{1} \in \text{Lin}(X)$ so folgt:

(a) $\widehat{\varepsilon} \perp \mathbf{1}$, d.h. $\frac{1}{n}\sum_{i=1}^{n} \widehat{\varepsilon}_i = \frac{1}{n}\mathbf{1}'\widehat{\varepsilon} = 0$.

(b) Die Regressionsgerade geht durch den Schwerpunkt (d.h. durch das Mittel) der Beobachtungen:

$$Y = X\widehat{\beta} + \widehat{\varepsilon} \quad \Rightarrow \quad \frac{1}{n}\mathbf{1}'Y = \frac{1}{n}\mathbf{1}'X\widehat{\beta}.$$

Beispiel 17.4 (Lineare Regression) Wir betrachten die beste lineare Prädiktion aus 16.11 erneut, diesmal basierend auf empirischen Daten $(X_1, Y_1), \ldots, (X_n, Y_n)$. Zur Erinnerung: In 16.11 wurde die Lösung von $\mathbf{E}\big[Y - (a + bX)\big]^2 = \min$ gegeben durch $b = \Sigma(X,X)^{-1}\Sigma(X,Y)$ und $a = \mathbf{E}Y - b'\,\mathbf{E}X$.

Das Modell $Y_i = a + bX_i + \varepsilon_i \, (i = 1, \ldots, n)$ kann geschrieben werden als

$$Y = X\beta + \varepsilon \quad \text{mit} \quad X = \begin{pmatrix} 1 & X_1 \\ \vdots & \vdots \\ 1 & X_n \end{pmatrix} \quad \text{und} \quad \beta = \begin{pmatrix} a \\ b \end{pmatrix}.$$

Da X in der Regressionsanalyse (in der Regel) vollen Rang hat, folgt für den KQ-Schätzer $\widehat{\beta} = (X'X)^{-1}X'Y$. Aufschlußreicher ist die Schreibweise

$$Y_{\mathrm{z}} = X_{\mathrm{z}}\begin{pmatrix} a' \\ b \end{pmatrix} + \varepsilon \quad \text{mit} \quad X_{\mathrm{z}} = \begin{pmatrix} 1 & X_1 - \overline{X}_n \\ \vdots & \vdots \\ 1 & X_n - \overline{X}_n \end{pmatrix}, \quad Y_{\mathrm{z}} = \begin{pmatrix} Y_1 - \overline{Y}_n \\ \vdots \\ Y_n - \overline{Y}_n \end{pmatrix}$$

[der Index "z" steht für "zentriert"] und $a' = a - \overline{Y}_n + b\overline{X}_n$. Es folgt

$$X_{\mathrm{z}}'X_{\mathrm{z}} = \begin{pmatrix} n & 0 \\ 0 & \sum_i (X_i - \overline{X}_n)^2 \end{pmatrix} \quad \text{und} \quad X_{\mathrm{z}}'Y_{\mathrm{z}} = \begin{pmatrix} 0 \\ \sum_i (X_i - \overline{X}_n)(Y_i - \overline{Y}_n) \end{pmatrix}$$

und damit

$$\widehat{\beta}_z = \begin{pmatrix} \widehat{a'} \\ \widehat{b} \end{pmatrix} = (X_{\mathrm{z}}'X_{\mathrm{z}})^{-1}X_{\mathrm{z}}'Y_{\mathrm{z}} = \begin{pmatrix} 0 \\ \frac{\sum_i (X_i - \overline{X}_n)(Y_i - \overline{Y}_n)}{\sum_i (X_i - \overline{X}_n)^2} \end{pmatrix},$$

d.h. wir erhalten analog zu 16.11 die Schätzer

$$\widehat{b} = S_{X,X}^{-1} S_{X,Y} \quad \text{und} \quad \widehat{a} = \overline{Y}_n - \widehat{b}\,\overline{X}_n,$$

wobei

$$S_{X,X} := \frac{1}{n-1}\sum_i (X_i - \overline{X}_n)^2 \quad \text{und} \quad S_{X,Y} := \frac{1}{n-1}\sum_i (X_i - \overline{X}_n)(Y_i - \overline{Y}_n).$$

Numerisch sind die Schätzer völlig identisch, d.h. es gilt $\binom{\widehat{a}}{\widehat{b}} = \widehat{\beta}$. Dieses folgt, weil in beiden Fällen dasselbe Minimierungsproblem gelöst wird und das Minimum eindeutig ist [man kann die Gleichheit aber auch explizit nachrechnen].

Bemerkung: Alle bisherigen Aussagen gelten sowohl für den Fall, dass X deterministisch, als auch für den Fall, dass X stochastisch ist.

Beispiel 17.5 (Varianzanalyse / Identifizierbarkeit) Ein Fall, in dem X nicht den vollen Rang hat, ist die zweifache Varianzanalyse mit additiven Effekten

$$Y_{ijk} = \alpha_i + \gamma_j + \varepsilon_{ijk} \qquad i = 1, \ldots, I;\ j = 1, \ldots, J;\ k = 1, \ldots, m.$$

Wir behandeln hier nur den Fall $I = J = 2$. Sei $\mathbf{1} = (\underbrace{1, \ldots, 1}_{\text{m - mal}})'$ und $\mathbf{0} = (\underbrace{0, \ldots, 0}_{\text{m - mal}})'$. Dann gilt

$$Y = X\beta + \varepsilon \quad \text{mit} \quad X = \begin{pmatrix} \mathbf{1} & \mathbf{0} & \mathbf{1} & \mathbf{0} \\ \mathbf{1} & \mathbf{0} & \mathbf{0} & \mathbf{1} \\ \mathbf{0} & \mathbf{1} & \mathbf{1} & \mathbf{0} \\ \mathbf{0} & \mathbf{1} & \mathbf{0} & \mathbf{1} \end{pmatrix} \quad \text{und} \quad \beta = \begin{pmatrix} \alpha_1 \\ \alpha_2 \\ \gamma_1 \\ \gamma_2 \end{pmatrix}.$$

Hier gilt offensichtlich Rang$X = 3$. Da die Parametervektoren $\beta = (\alpha_1, \alpha_2, \gamma_1, \gamma_2)'$ und $\beta = (\alpha_1 + c, \alpha_2 + c, \gamma_1 - c, \gamma_2 - c)'$ dasselbe Modell liefern, ist klar, dass z.B. α_1 nicht "identifizierbar" ist. Allerdings ist $\alpha_1 - \alpha_2$ "identifizierbar" (solche Differenzen bezeichnet man als Kontraste) oder z.B. auch $\alpha_1 + \gamma_1$. Wir wollen die Identifizierbarkeit jetzt definieren und genauer untersuchen.

Definition 17.6 Sei C ein Vektor oder eine Matrix. $C\beta$ heißt identifizierbar, falls

$$\forall \beta, \beta^* \in \mathbb{R}^k : \quad X\beta = X\beta^* \ \Rightarrow\ C\beta = C\beta^*.$$

Proposition 17.7 *$C\beta$ ist identifizierbar genau dann wenn*

$$\exists\, C_0 :\ C = C_0 X \quad \big(\Leftrightarrow\ \mathrm{Lin}(C') \subset \mathrm{Lin}(X') \big).$$

Beweis. „⇐ “: Sei $C = C_0 X$ und $X\beta = X\beta^*$. Dann folgt

$$C\beta = C_0 X\beta = C_0 X\beta^* = C\beta^*.$$

„⇒ “: Für $\mathrm{Lin}(X') = \mathbb{R}^k$ ist die Aussage erfüllt. Sei also $\mathrm{Lin}(X') \subset \mathbb{R}^k$ und $x \in \mathrm{Lin}(X')_\perp$.

$$\Rightarrow \quad Xx = 0 = X0 \quad \Rightarrow \quad Cx = C0 = 0\,,$$

d.h. die Zeilen von C sind orthogonal zu x. Da x beliebig aus $\mathrm{Lin}(X')_\perp$ ist, folgt
$\mathrm{Lin}(C') \subseteq \mathrm{Lin}(X')$ □

Bemerkung 17.8 In der Regressionsanalyse gilt (idR) $\mathrm{Rang}X = k$ und damit auch $\mathrm{Lin}(C') \subset \mathbb{R}^k = \mathrm{Lin}(X')$, d.h. alle $C\beta$ und insbesondere β selbst sind identifizierbar. Ferner ist $\widehat{\beta}$ eindeutig bestimmt. In der Varianzanalyse oder Kovarianzanalyse [die hier nicht behandelt wird] stimmt das oft nicht: Im obigen Beispiel 17.5 ist $\alpha_1 = (1,0,0,0)\beta$ nicht identifizierbar [wäre $(1,0,0,0)' \in \mathrm{Lin}(X')$, so würden aus Symmetriegründen alle Einheitsvektoren in $\mathrm{Lin}(X')$ liegen und es müßte $\mathrm{Rang}X = 4$ gelten]. Andererseits gilt für $C = (1,-1,0,0)$

$$C' = \begin{pmatrix} 1 \\ -1 \\ 0 \\ 0 \end{pmatrix} = \begin{pmatrix} 1 \\ 0 \\ 1 \\ 0 \end{pmatrix} - \begin{pmatrix} 0 \\ 1 \\ 1 \\ 0 \end{pmatrix} \in \mathrm{Lin}(X').$$

Das heißt $C\beta = \alpha_1 - \alpha_2$ ist identifizierbar. Ebenso ist $(1,0,1,0)\beta = \alpha_1 + \gamma_1$ identifizierbar.

17.9 (Gauß-Markov-Schätzer) Ist $\psi = C\beta$ identifizierbar, dann ist ein naheliegender Schätzer für ψ der Gauß-Markov-Schätzer (GM-Schätzer) $\widehat{\psi} = C\widehat{\beta}$, wobei $\widehat{\beta}$ eine Lösung der Normalengleichung $X'X\widehat{\beta} = X'Y$ ist. Ist $\mathrm{Rang}X < k$ dann ist $\widehat{\psi}$ trotzdem eindeutig bestimmt, da nach Satz 17.2 gilt

$$\widehat{\psi} = C\widehat{\beta} = C_0 X\widehat{\beta} = C_0 P_{\mathcal{X}} Y$$

und die Projektion $P_{\mathcal{X}} Y$ eindeutig ist.

Satz 17.10 (Gauß-Markov-Theorem) *Sei $\psi = C\beta$ identifizierbar und $\widehat{\psi}$ der Gauß-Markov-Schätzer. Dann gilt*

(i) $$\mathbf{E}\widehat{\psi} = \psi$$

und

$$\mathrm{Var}(\widehat{\psi}) = \sigma^2 C_0 P_{\mathcal{X}} C_0' \quad (= \sigma^2 C(X'X)^{-1}C' \text{ falls } \mathrm{Rang}X = k\,);$$

(ii) $\widehat{\psi}$ ist der eindeutig bestimmte, beste lineare unverfälschte Schätzer für ψ, d.h.

$$\mathrm{Var}(\widehat{\psi}) \le \mathrm{Var}(\widetilde{\psi}) \quad \forall \text{ lineare } \widetilde{\psi} \text{ mit } \mathbf{E}\widetilde{\psi} = \psi.$$

Beweis. Sei wiederum $P = P_{\mathcal{X}}$ und $\Sigma(Y)$ die Kovarianzmatrix von Y.
(i) Wegen $\mathbf{E}\varepsilon = 0$ gilt $\mathbf{E}Y = X\beta$ und damit

$$\mathbf{E}(\widehat{\psi}) = C_0 P\,\mathbf{E}Y = C_0 PX\beta = C_0 X\beta = C\beta = \psi.$$

Wir setzen nun zur Abkürzung $a_0 := PC_0' \in \mathcal{X}$, d.h. $\widehat{\psi} = a_0'Y$. Damit gilt

$$\begin{aligned}\mathrm{Var}(\widehat{\psi}) &= a_0'\,\Sigma(Y)\,a_0 = a_0'\,\Sigma(\varepsilon)\,a_0 = \sigma^2 a_0'a_0 = \sigma^2 C_0 P^2 C_0' \\ &= \sigma^2 C(X'X)^{-1}C' \ \text{ falls } \mathrm{Rang}X = k\,.\end{aligned}$$

(ii) Sei nun $\widetilde{\psi} = a'Y$ ein anderer linearer unverfälschter Schätzer für ψ. Dann gilt

$$\begin{aligned}&\mathbf{E}\big((Pa)'Y\big) = a'P\,\mathbf{E}Y = a'PX\beta = a'X\beta = \mathbf{E}(a'Y) = \psi \\ &\Rightarrow 0 = \mathbf{E}\big((Pa)'Y - \widehat{\psi}\big) = \mathbf{E}(Pa - a_0)'Y = (Pa - a_0)'X\beta \quad \forall\beta\,,\end{aligned}$$

d.h. $Pa - a_0 \perp \mathcal{X}$. Andererseits gilt wegen $a_0 := PC_0'$ auch $Pa - a_0 \in \mathcal{X}$ d.h. $Pa - a_0 = 0$. Daraus folgt

$$\begin{aligned}\mathrm{Var}(\widetilde{\psi}) &= a'\Sigma(Y)a = a'\Sigma(\varepsilon)a \\ &= \sigma^2\|a\|^2 = \sigma^2\big(\|Pa\|^2 + \|(I-P)a\|^2\big) \\ &= \sigma^2\|a_0\|^2 + \sigma^2\|(I-P)a\|^2 \\ &= \mathrm{Var}(\widehat{\psi}) + \sigma^2\|(I-P)a\|^2 \\ &\ge \mathrm{Var}(\widehat{\psi})\end{aligned}$$

mit Gleichheit genau dann wenn $(I-P)a = 0$, d.h. $a = Pa = a_0$. □

Bemerkung: Es gibt auch wichtige nichtlineare Schätzer, zum Beispiel in der "robusten" Statistik, wo Schätzer betrachtet werden, die unempfindlich gegenüber dem Einfluss möglicher Ausreißer sind.

Beispiel 17.11 (Zweifache Varianzanalyse / Fortsetzung von Beispiel 17.5)
Zur Erinnerung: Wir betrachten das Modell

$$Y = X\beta + \varepsilon \quad \text{mit} \quad X = \begin{pmatrix} \mathbf{1} & \mathbf{0} & \mathbf{1} & \mathbf{0} \\ \mathbf{1} & \mathbf{0} & \mathbf{0} & \mathbf{1} \\ \mathbf{0} & \mathbf{1} & \mathbf{1} & \mathbf{0} \\ \mathbf{0} & \mathbf{1} & \mathbf{0} & \mathbf{1} \end{pmatrix} \quad \text{und} \quad \beta = \begin{pmatrix} \alpha_1 \\ \alpha_2 \\ \gamma_1 \\ \gamma_2 \end{pmatrix}.$$

Wie wir oben gesehen haben ist $\alpha_1 - \alpha_2 = (1\ {-1}\ 0\ 0)\beta =: C\beta$ identifizierbar. Genauer gilt

$$C = (1\ {-1}\ 0\ 0) = \frac{1}{2m}\big(\mathbf{1}'\ \mathbf{1}'\ {-\mathbf{1}'}\ {-\mathbf{1}'}\big)\ X =: C_0\, X.$$

Der Gauß-Markov-Schätzer für $\psi = C\beta = \alpha_1 - \alpha_2$ beträgt damit $\widehat{\psi} = C\widehat{\beta} = C_0 P_{\mathcal{X}} Y$. Das Problem ist nun, die Projektion $P_{\mathcal{X}} Y$ zu berechnen (insbesondere, weil $\mathrm{Rang} X = 3$ ist). Man kann aber leicht sehen, dass die Vektoren

$$\tilde{x_1} := \frac{1}{\sqrt{2m}} \begin{pmatrix} \mathbf{1} \\ \mathbf{1} \\ \mathbf{0} \\ \mathbf{0} \end{pmatrix}, \quad \tilde{x_2} := \frac{1}{\sqrt{2m}} \begin{pmatrix} \mathbf{0} \\ \mathbf{0} \\ \mathbf{1} \\ \mathbf{1} \end{pmatrix}, \quad \tilde{x_3} := \frac{1}{\sqrt{4m}} \begin{pmatrix} \mathbf{1} \\ \mathbf{-1} \\ \mathbf{1} \\ \mathbf{-1} \end{pmatrix}$$

eine Orthonormalbasis von $\mathcal{X} = \mathrm{Lin}(X)$ bilden, d.h. es gilt mit $\tilde{X} := (\tilde{x_1}, \tilde{x_2}, \tilde{x_3})$

$$\widehat{\psi} = C_0\, P_{\mathcal{X}}\, Y = C_0\, \tilde{X}\big(\tilde{X}'\tilde{X}\big)^{-1}\tilde{X}'\, Y = C_0\, \tilde{X} I_3 \tilde{X}'\, Y = \frac{1}{\sqrt{2m}}\, \tilde{x_1}'Y - \frac{1}{\sqrt{2m}}\, \tilde{x_2}'Y = Y_{1\cdot\cdot} - Y_{2\cdot\cdot}$$

wobei wir die in der Varianzanalyse übliche Notation verwenden, nach der ein Punkt die Mittelung über die entsprechende Komponente bedeutet, d.h. $Y_{i\cdot\cdot} := \frac{1}{2}\sum_{j=1}^{2} \frac{1}{m} \sum_{k=1}^{m} Y_{ijk}$.
□

Ein weiteres Beispiel zum Gauß-Markov-Schätzer ist die Prognose bei der linearen Regression (vgl. Beispiel 18.9).

18 Der F-Test als Likelihood-Quotienten Test und Konfidenzintervalle im linearen Modell

In diesem Kapitel werden der F-Test und Konfidenzintervalle im linearen Modell unter der Annahme einer Normalverteilung hergeleitet. Der F-Test ist dabei auch ein prominentes Beispiel für einen Likelihood-Quotienten Test. Höhepunkt des Kapitels ist der Satz von Scheffé mit einem Konfidenzband für die Prognose in der linearen Regression.

Bemerkung 18.1 (LQ-Test: Das Konzept) Seien $X_1, \ldots, X_n$ ZVAs mit gemeinsamer Dichte $f_{\theta_0}(x)$ und $\theta_0 \in \Theta \subset \mathbb{R}^d$. Man will die

Hypothese	$H_0 : \theta_0 \in \Theta_0 \subset \Theta$	gegen die
Alternativhypothese	$H_A : \theta_0 \in \Theta \setminus \Theta_0$	testen.

Likelihood-Quotienten Test:

$$\phi(X) = \begin{cases} 1\,, & \Lambda(X) < c^* \\ 0\,, & \Lambda(X) \geq c^* \end{cases} \quad , \text{ wobei } \Lambda(X) := \frac{\sup\limits_{\theta \in \Theta_0} f_\theta(X)}{\sup\limits_{\theta \in \Theta} f_\theta(X)}.$$

Man wählt c^* möglichst groß derart, dass das Niveau α unter der Hypothese eingehalten wird, d.h. dass $\mathbf{P}_\theta\big(\Lambda(X) < c^*\big) \leq \alpha \quad \forall \theta \in \Theta_0$ gilt. Um c^* zu bestimmen benötigt man die Verteilung von $\Lambda(X)$ oder von monotonen Transformationen von $\Lambda(X)$.

[In der asymptotischen Statistik zeigt man, dass $-\log \Lambda(X)$ in vielen Fällen in Verteilung gegen eine χ^2-Verteilung konvergiert, d.h. man kann dann diese Grenzverteilung zur näherungsweisen Bestimmung von c^* verwenden. Man beachte auch die Analogie zum Neyman-Pearson-Test, bei dem einfache Hypothesen anhand des Likelihood-Quotienten getestet werden! Aber: der LQ-Test ist nicht immer optimal!]

18.2 (Der F-Test bei linearen Hypothesen) Wir betrachten wieder das Modell $Y = X\beta + \varepsilon$ und wollen Hypothesen der Form $H : H'\beta = 0$ testen. Dabei kann man

nur solche Hypothesen testen, für die $H'\beta$ identifizierbar ist, d.h. wir nehmen an, dass es ein H_0 gibt mit $H' = H_0'X$. [d.h. $\mathrm{Lin}(H) \subset \mathrm{Lin}(X')$].

Beispiel: Im Modell der zweifachen Varianzanalyse $Y_{ijk} = \alpha_i + \gamma_j + \varepsilon_{ijk}$ aus Bemerkung 17.5 war $\beta = (\alpha_1, \alpha_2, \gamma_1, \gamma_2)'$. Das bedeutet, dass man mit

$$H = \begin{pmatrix} 1 \\ -1 \\ 0 \\ 0 \end{pmatrix} \qquad H = \begin{pmatrix} 0 \\ 0 \\ 1 \\ -1 \end{pmatrix} \qquad H = \begin{pmatrix} 1 \\ -1 \\ 1 \\ -1 \end{pmatrix}$$

jeweils die Hypothesen $\alpha_1 = \alpha_2$, $\gamma_1 = \gamma_2$ bzw. $\big(\alpha_1 = \alpha_2, \gamma_1 = \gamma_2\big)$ testet. Die Bedingung $\mathrm{Lin}(H) \subset \mathrm{Lin}(X')$ ist offensichtlich erfüllt.

Um den LQ-Test anwenden zu können benötigen wir die Dichte der Beobachtungen. Aus diesem Grund (und um die Verteilung der Test-Statistik wirklich ausrechnen zu können) nehmen wir zusätzlich an, dass die Fehler ε_i normalverteilt sind, d.h. $Y \sim \mathcal{N}(X\beta, \sigma^2 I_n)$ mit Dichte φ_{β,σ^2}. Die LQ - Teststatistik ist dann wie oben

$$\Lambda(Y) = \frac{\sup\limits_{H'\beta=0\,,\,\sigma^2} \varphi_{\beta,\sigma^2}(Y)}{\sup\limits_{\beta,\sigma^2} \varphi_{\beta,\sigma^2}(Y)}$$

und es gilt

$$\Lambda(Y) \geq c \;\Leftrightarrow\; \log \Lambda(Y) = \sup_{H'\beta=0,\sigma^2} \log \varphi_{\beta,\sigma^2}(Y) - \sup_{\beta,\sigma^2} \log \varphi_{\beta,\sigma^2}(Y) \geq c_1$$

wobei

$$\begin{aligned} \log \varphi_{\beta,\sigma^2}(Y) &= \log\left[\frac{1}{(2\pi)^{n/2}\sigma^n} \exp(-\frac{1}{2\sigma^2}(Y - X\beta)'(Y - X\beta))\right] \\ &= -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log\sigma^2 - \frac{1}{2\sigma^2}(Y - X\beta)'(Y - X\beta). \end{aligned}$$

Die Ableitungen nach β und σ^2 ergeben

$$\frac{\partial}{\partial\beta} \log \varphi_{\beta,\sigma^2}(Y) = -\frac{1}{2\sigma^2}\frac{\partial}{\partial\beta}(Y - X\beta)'(Y - X\beta) = 0 \;\Leftrightarrow\; \beta = \widehat{\beta}$$

$$\frac{\partial}{\partial\sigma^2}\log\varphi_{\beta,\sigma^2}(Y) = -\frac{n}{2}\frac{1}{\sigma^2} + \frac{1}{2\sigma^4}(Y-X\beta)'(Y-X\beta) = 0 \;\Leftrightarrow\; \sigma^2 = \widehat{\sigma}^2 := \frac{1}{n}R_0^2,$$

wobei

$$R_0^2 := \min_{\beta}(Y-X\beta)'(Y-X\beta) = (Y-X\widehat{\beta})'(Y-X\widehat{\beta})$$

die minimale Fehlerquadratsumme ist. Damit gilt

$$\sup_{\beta,\sigma^2}\log\varphi_{\beta,\sigma^2}(Y) = \log\varphi_{\widehat{\beta},\widehat{\sigma}^2}(Y) = -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log\frac{1}{n}R_0^2 - \frac{n}{2}.$$

Analog erhält man

$$\sup_{H'\beta=0,\sigma^2}\log\varphi_{\beta,\sigma^2}(Y) = -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log\frac{1}{n}R_1^2 - \frac{n}{2},$$

wobei

$$R_1^2 := \min_{\beta\,:\,H'\beta=0}(Y-X\beta)'(Y-X\beta)$$

die minimale Fehlerquadratsumme unter der Restriktion $H'\beta = 0$ ist. Damit gilt für den LQ-Test

$$\log\Lambda(Y) = \frac{n}{2}\log\frac{R_0^2}{R_1^2} \geq c_1 \;\Leftrightarrow\; \frac{R_1^2}{R_0^2} \leq c_2 \;\Leftrightarrow\; \frac{R_1^2-R_0^2}{\ell}\Big/\frac{R_0^2}{n-r} \leq c_3$$

mit geeigneten c_2 und c_3 (monotone Transformation). Hierbei ist $r = \text{ Rang } X$ und $\ell = \text{ Rang } H$ [ℓ ist damit die Anzahl der 'linear unabhängigen Hypothesen'].

Wir zeigen im nachfolgenden Satz, dass $R_1^2 - R_0^2$ und R_0^2 unter der Hypothese stochastisch unabhängig und χ_ℓ^2 - bzw. χ_{n-r}^2 - verteilt sind, d.h. es gilt für alle $\theta \in \Theta_0$

$$F := \frac{R_1^2-R_0^2}{\ell}\Big/\frac{R_0^2}{n-r} \sim F_{\ell,n-r} := \frac{\frac{1}{\ell}\chi_\ell^2}{\frac{1}{n-r}\chi_{n-r}^2} \tag{9}$$

(**Definition** der Fisher-Verteilung mit ℓ und $n-r$ Freiheitsgraden). Da wir das größte c bzw. das kleinste c_3 suchen mit

$$\mathbf{P}_\theta\big(\Lambda(Y) < c\big) = \mathbf{P}_\theta\big(F > c_3\big) \leq \alpha \quad \forall\theta \in \Theta_0,$$

folgt für die Lösung $c_3^* = F_{\ell,n-r;1-\alpha}$. Tafeln der Quantile $F_{\ell,n-r;1-\alpha}$ sind im Anhang vieler Bücher zu finden.

[Man beachte, dass diese Konstruktion deshalb funktioniert, weil die Verteilung $\mathbf{P}_\theta^F$ für alle $\theta \in \Theta_0$ nicht von θ abhängt - eine solche Statistik F nennt man Pivot - Statistik.]

Beispiel 18.3 (Testen auf Gleichheit zweier Regressionsgeraden)
Angenommen wir haben Beobachtungen zu zwei Regressionsmodellen $Y_{1i} = a_1 + b_1 X_{1i} + \varepsilon_{1i}\,(i = 1, \ldots, n_1)$ und $Y_{2i} = a_2 + b_2 X_{2i} + \varepsilon_{2i}\,(i = 1, \ldots, n_2)$ und möchten die Hypothese testen, dass die Geraden gleich sind, d.h. dass $a_1 = a_2$ und $b_1 = b_2$ gilt. Dazu fassen wir die Modelle zu einem gemeinsamen Modell zusammen:

$$Y = X \begin{pmatrix} a_1 \\ b_1 \\ a_2 \\ b_2 \end{pmatrix} + \varepsilon \quad \text{mit} \quad Y = \begin{pmatrix} Y_{11} \\ \vdots \\ Y_{1n_1} \\ Y_{21} \\ \vdots \\ Y_{2n_2} \end{pmatrix}, \quad X = \begin{pmatrix} 1 & X_{11} & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 1 & X_{1n_1} & 0 & 0 \\ 0 & 0 & 1 & X_{21} \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 1 & X_{2n_2} \end{pmatrix} \text{ und } \varepsilon = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \end{pmatrix}.$$

Die Hypothese lautet dann $H'\beta = 0$ mit $H' = \left(\begin{smallmatrix} 1 & 0 & -1 & 0 \\ 0 & 1 & 0 & -1 \end{smallmatrix}\right)$, d.h. $\ell = \text{ Rang } H = 2$. Wir berechnen die F-Statistik für diese Hypothese: Die minimale Fehlerquadratsumme R_0^2 ergibt sich durch Regression in diesem Modell (das ist offensichtlich identisch zu der Summe der minimalen Fehlerquadratsummen in den einzelnen Modellen). Die minimale Fehlerquadratsumme R_1^2 unter der Restriktion $H'\beta = 0$ ist manchmal schwer zu berechnen. In vielen Fällen kann man sie aber durch ein modifiziertes Regressionsproblem (ohne Restriktion) erhalten. In diesem Fall kann man hierfür offensichtlich die minimale Fehlerquadratsumme in folgendem Regressionsmodell verwenden:

$$Y = \tilde{X} \begin{pmatrix} a_1 \\ b_1 \end{pmatrix} + \tilde{\varepsilon} \quad \text{mit} \quad \tilde{X} = \begin{pmatrix} 1 & X_{11} \\ \vdots & \vdots \\ 1 & X_{1n_1} \\ 1 & X_{21} \\ \vdots & \vdots \\ 1 & X_{2n_2} \end{pmatrix}.$$

Man verwendet jetzt die F-Statistik wie in (9) mit $n = n_1 + n_2$ und $r = \text{ Rang } X = 4$. □

Beispiel 18.4 (Zweifache Varianzanalyse / Test auf Additivität) Wir betrachten das allgemeine Modell der zweifachen Varianzanalyse

$$Y_{ijk} = \mu_{ij} + \varepsilon_{ijk} \qquad i = 1, \ldots, I;\ j = 1, \ldots, J;\ k = 1, \ldots, m.$$

und testen, ob die Effekte additiv sind, d.h. die Hypothese $\mu_{ij} = \alpha_i + \gamma_j$. Für $I = J = 2$ kann man z.B. zeigen, dass

$$\mu_{ij} = \alpha_i + \gamma_j \quad \Leftrightarrow \quad \mu_{11} - \mu_{12} - \mu_{21} + \mu_{22} = 0$$

(die Richtung "⇒" kann man einfach nachrechnen; für die Richtung "⇐" kann man z.B. $\alpha_i := \mu_{i\cdot}$ und $\beta_j := \mu_{\cdot j} - \mu_{\cdot\cdot}$ definieren und diese dann ebenfalls nachrechnen - hierbei haben wir wieder die Notation verwenden, dass ein Punkt die Mittelung über die entsprechende Komponente bedeutet). Man kann nun den F-Test aus Satz 18.5 anwenden. R_0^2 ist dabei die normale Fehlerquadratsumme. Die Fehlerquadratsumme R_1^2 unter der Restriktion der Hypothese ergibt sich als Fehlerquadratsumme (ohne Restriktion) im Modell aus Beispiel 17.5. Ferner ist $n = IJm = 4m$, $r = IJ = 4$ und $\ell = 1$. □

Wir wollen nun die Aussage $F \sim F_{\ell,n-r}$ beweisen.

Satz 18.5 (F-Test) *Sei $Y \sim \mathcal{N}(X\beta, \sigma^2 I_n)$ und H eine Matrix vom Rang ℓ mit $\text{Lin}(H) \subset \text{Lin}(X')$. Seien R_0^2 und R_1^2 wie oben (d.h. R_1^2 unter $H'\beta = 0$). Dann gilt unter der Hypothese*

(i) R_0^2 *und* $R_1^2 - R_0^2$ *sind stochastisch unabhängig.*

(ii) $R_0^2/\sigma^2 \sim \chi^2_{n-r}$ *und* $(R_1^2 - R_0^2)/\sigma^2 \sim \chi^2_\ell$ *und damit*

$$F := \frac{R_1^2 - R_0^2}{\ell} \Big/ \frac{R_0^2}{n-r} \sim F_{\ell,n-r}.$$

Beweis. Sei $V_H := \{X\gamma \mid H'\gamma = 0\}$. Wegen $H'\gamma = H_0'X\gamma = 0$ ist V_H ein Vektorraum $\subseteq \mathcal{X}$ mit $\dim V_H = \dim \mathcal{X} - \dim \text{Bild } H_0'|_{\mathcal{X}} = \dim \mathcal{X} - \text{Rang}H = r - \ell$. Sei nun (ONB = Orthonormalbasis)

$$\underbrace{\underbrace{Q_1, \ldots, Q_{r-\ell}}_{\text{ONB von } V_H}, Q_{r-\ell+1}, \ldots, Q_r}_{\text{ONB von } \mathcal{X}}, Q_{r+1}, \ldots, Q_n \quad \text{ONB von } \mathbb{R}^n.$$

Dann ist $I = \sum_{j=1}^n Q_j Q_j'$,

$$P_H := \sum_{j=1}^{r-\ell} Q_j Q_j' \quad \text{der Projektionsoperator auf } V_H,$$

und

$$P_{\mathcal{X}} := \sum_{j=1}^{r} Q_j Q_j' \quad \text{der Projektionsoperator auf } \mathcal{X}$$

(unmittelbar klar, da $P_H^2 = P_H$, P_H symmetrisch und $P_H\, x = x\ \forall x \in V_H$; für $P_{\mathcal{X}}$ analog).

Es gilt

$$Z := \begin{pmatrix} Q_1' \\ \vdots \\ Q_n' \end{pmatrix} (Y - X\beta) \sim \mathcal{N}\left(0, \sigma^2 \begin{pmatrix} Q_1' \\ \vdots \\ Q_n' \end{pmatrix} (Q_1, \ldots, Q_n)\right) = \mathcal{N}(0, \sigma^2 I_n),$$

d.h. die Z_j und damit auch die $Q_j'Y$ sind stoch. unabhängig. Damit gilt

$$\begin{aligned} R_1^2 &= \min_{\beta\,:\,H'\beta=0} (Y - X\beta)'(Y - X\beta) \\ &= Y'(I - P_H)(I - P_H)Y = Y'(I - P_H)Y, \end{aligned}$$

und analog

$$R_0^2 = Y'(I - P_{\mathcal{X}})Y \underset{\substack{\uparrow \\ \text{da } X\beta \in \mathcal{X}}}{=} (Y - X\beta)'(I - P_{\mathcal{X}})(Y - X\beta) = \sum_{j=r+1}^{n} Z_j^2 \sim \sigma^2 \chi_{n-r}^2$$

sowie

$$R_1^2 - R_0^2 = Y'(P_{\mathcal{X}} - P_H)Y = \sum_{j=r-\ell+1}^{r} Y'Q_jQ_j'Y \sim \sigma^2 \chi_\ell^2.$$

Damit folgen (i) und (ii). □

Korollar 18.6 *Sei*

$$\widehat{\sigma}^2 := \frac{1}{n-r} R_0^2.$$

Dann gilt

$$\mathbf{E}\widehat{\sigma}^2 = \sigma^2 \qquad \textit{und} \qquad \operatorname{Var} \widehat{\sigma}^2 = \frac{2\sigma^4}{n-r}.$$

Bemerkung: Ohne die Normalverteilungsannahme gilt ebenfalls $\mathbf{E}\widehat{\sigma}^2 = \sigma^2$. Die Varianz hat idR die gleiche Konvergenzrate.

Beweis. Aus Proposition 13.5 folgt für $Y \sim \chi_{n-r}^2$ $\mathbf{E}Y = n-r$ und $\operatorname{Var}Y = 2(n-r)$, d.h.

$$\mathbf{E}\widehat{\sigma}^2 = \mathbf{E}\Big(\frac{R_0^2}{\sigma^2}\frac{\sigma^2}{n-r}\Big) = \sigma^2$$

und

$$\operatorname{Var} \widehat{\sigma}^2 = \operatorname{Var}\Big(\frac{R_0^2}{\sigma^2}\frac{\sigma^2}{n-r}\Big) = \frac{2\sigma^4}{n-r}.$$

□

Auf analoge Art und Weise konstruieren wir nun Konfidenzellipsoide für identifizierbare $\psi = C\beta$ basierend auf dem Gauß-Markov-Schätzer $\widehat{\psi} = C\widehat{\beta} = C_0 P_{\mathcal{X}} Y$ (wobei $C = C_0 X$). Aus dem Gauß-Markov-Theorem folgt

$$\operatorname{Var}(\widehat{\psi}) = \sigma^2 B \quad \text{mit} \quad B := C_0 P_{\mathcal{X}} C_0' \ \big(= C(X'X)^{-1}C' \text{ falls } \operatorname{Rang} X = k\big).$$

Satz 18.7 (Konfidenzellipsoide für Gauß-Markov-Schätzer)
(i) Sei $Y \sim \mathcal{N}(X\beta, \sigma^2 I_n)$ *und* R_0^2 *wie oben. Ferner sei* $\psi = C\beta$ *identifizierbar mit Rang* $C = \ell$ *und* $\widehat{\psi}$ *der Gauß-Markov-Schätzer. Dann sind* $\widehat{\psi}$ *und* R_0^2 *stochastisch unabhängig mit* $\widehat{\psi} - \psi \sim \mathcal{N}(0, \sigma^2 B)$ *und* $R_0^2 \sim \sigma^2 \chi^2(n-r)$*, d.h.*

$$\frac{(\widehat{\psi} - \psi)' B^{-1} (\widehat{\psi} - \psi)}{\ell} \Big/ \frac{R_0^2}{n-r} \sim F_{\ell, n-r}.$$

Damit ist

$$K(Y) := \left\{ \psi \;\middle|\; (\widehat{\psi} - \psi)' B^{-1} (\widehat{\psi} - \psi) \le \ell\, \widehat{\sigma}^2\, F_{\ell, n-r; 1-\alpha} \right\}$$

ein $(1-\alpha)$*-Konfidenzellipsoid für* $\psi = C\beta$ $\big($*d.h.* $\mathbf{P}_{(\beta,\sigma^2)}\big(C\beta \in K(Y)\big) = 1-\alpha \;\forall\, (\beta, \sigma^2)\big)$.

(ii) Ist $\mathrm{Rang} X = k$*, so ist*

$$\left\{ \beta \;\middle|\; (\widehat{\beta} - \beta)'\, (X'X) (\widehat{\beta} - \beta) \le k\, \widehat{\sigma}^2\, F_{k, n-k; 1-\alpha} \right\}$$

ein $(1-\alpha)$ *- Konfidenzellipsoid für* β.

Bemerkung: Da die Matrix B^{-1} bzw. $X'X$ symmetrisch und positiv definit ist, handelt es sich wirklich um eine Ellipse (genauer um die von einer Ellipse begrenzte Fläche) mit Mittelpunkt $\widehat{\psi}$ bzw. $\widehat{\beta}$ und gedrehten Hauptachsen (analog zur Hauptachsentransformation in Bemerkung **??**).

Beweis. (i) Es gilt

$$\widehat{\psi} - \psi = C_0 P_{\mathcal{X}} Y - C_0 X \beta = C_0 P_{\mathcal{X}} (Y - X\beta) \sim \mathcal{N}\big(0, \sigma^2 C_0 P_{\mathcal{X}} C_0'\big)$$

und wie im Beweis vom obigen Satz 18.5

$$R_0^2 = (Y - X\beta)'(I - P_{\mathcal{X}})(Y - X\beta) \sim \sigma^2 \chi^2(n-r)$$

d.h. $\widehat{\psi} - C\beta$ und R_0^2 sind stochastisch unabhängig. Damit folgt die Behauptung.
(ii) folgt mit $C = I_k$ und $\ell = r = k$ unmittelbar aus (i). □

Beispiel 18.8 (Regression / Fortsetzung von Beispiel 17.4)
Wir betrachten wieder das Modell $Y_i = a + bX_i + \varepsilon_i$ $(i = 1, \ldots, n)$, d.h.

$$Y = X\beta + \varepsilon \quad \text{mit} \quad X = \begin{pmatrix} 1 & X_1 \\ \vdots & \vdots \\ 1 & X_n \end{pmatrix} \quad \text{und} \quad \beta = \begin{pmatrix} a \\ b \end{pmatrix}.$$

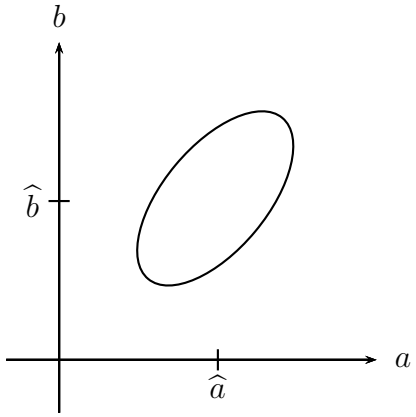
Sei $\widehat{\beta} = \binom{\widehat{a}}{\widehat{b}}$ der KQ-Schätzer. Wir suchen ein Konfidenzellipsoid für β und wenden dafür obigen Satz mit $C = I_2$ an. Es gilt $B = (X'X)^{-1}$ mit

$$(X'X) = \begin{pmatrix} n & \sum X_i \\ \sum X_i & \sum X_i^2 \end{pmatrix},$$

d.h. wir erhalten

$$K(Y) = \left\{ \binom{a}{b} \;\middle|\; \binom{\widehat{a} - a}{\widehat{b} - b}' (X'X) \binom{\widehat{a} - a}{\widehat{b} - b} \leq 2\,\widehat{\sigma}^2\, F_{2,n-2;1-\alpha} \right\}.$$

Man kann das Konfidenzellipsoid alternativ auch aus dem zentrierten Schätzer $\widehat{\beta}_z = \binom{\widehat{a'}}{\widehat{b}} = (X_{\mathrm{z}}'X_{\mathrm{z}})^{-1}X_{\mathrm{z}}'Y_{\mathrm{z}}$ und der Transformation $a' = a - \overline{Y}_n + b\overline{X}_n$ (s. Beispiel 17.4) konstruieren. Aufgrund der Eindeutigkeit der Größen $\widehat{\psi}, \psi, B$ und R_0^2 in Satz 18.7 folgt, dass dieses zu dem gleichen Konfidenzellipsoid führt. Man kann das aber auch durch Transformation des Konfidenzellipsoids explizit nachrechnen (s. (18.14) im Anhang 18.3).



Beispiel 18.9 (Prognose bei der linearen Regression)
Wir betrachten wieder die lineare Regression aus Beispiel 17.4

$$Y = X\beta + \varepsilon \quad \text{mit} \quad X = \begin{pmatrix} 1 & X_1 - \overline{X}_n \\ \vdots & \vdots \\ 1 & X_n - \overline{X}_n \end{pmatrix} \quad \text{und} \quad \beta = \begin{pmatrix} a \\ b \end{pmatrix}$$

($r = \text{Rang}X = 2$). Gegeben sei ein neuer Designpunkt $C = C_\xi = (1, \xi - \overline{X}_n)$. Schätze $\psi_\xi = C\beta = \beta_1 + \beta_2(\xi - \overline{X}_n) = \mathbf{E}Y_\xi$ durch den Gauß-Markov-Schätzer $\widehat{\psi}_\xi = \widehat{\beta}_1 + \widehat{\beta}_2(\xi - \overline{X}_n)$ [dieses ist der Punkt auf der Geraden an der Stelle ξ bzw. $\xi - \overline{X}_n$]. Wir wollen ein Konfidenzintervall [eindimensionales Konfidenzellipsoid] herleiten: Es gilt

$$B = B_{\psi_\xi} := C(X'X)^{-1}C' = C \begin{pmatrix} n & 0 \\ 0 & \sum_i (X_i - \overline{X}_n)^2 \end{pmatrix}^{-1} C' = \frac{1}{n} + \frac{(\xi - \overline{X}_n)^2}{\sum_i (X_i - \overline{X}_n)^2}$$

Damit ergibt Satz 18.7, dass

$$K(Y) := \left\{ \psi_\xi \;\middle|\; |\widehat{\psi_\xi} - \psi_\xi| \le \widehat{\sigma} \sqrt{\frac{1}{n} + \frac{(\xi - \overline{X}_n)^2}{\sum_i (X_i - \overline{X}_n)^2}} \sqrt{F_{1,n-2;1-\alpha}} \right\}$$

ein $(1-\alpha)$ - Konfidenzintervall [eindimensionales Konfidenzellipsoid] für ψ_ξ ist. Die Breite des Konfidenzintervalls hängt dabei von der Entfernung von ξ von $\overline{X}_n$ ab [heuristisch

klar, wenn man sich überlegt, wie sich Schwankungen von $\widehat{\beta}$ auf die Regressionsgerade auswirken].

Bemerkung: Das zugehörige Konfidenzintervall für den Wert $Y_\xi = C_\xi\beta + \varepsilon_\xi$ (anstelle von $\mathbf{E}Y_\xi$) ist wegen der zusätzlichen Streuung von ε_ξ natürlich größer (Ü-Aufgabe).

In vielen Fällen interessiert man sich anstelle eines Konfidenzintervalls an einem Designpunkt ξ eher für ein "Konfidenzband" für die gesamte Kurve. Es ist erstaunlich, dass man in der vorliegenden Situation ein solches exaktes Konfidenzband angeben kann [exakt in dem Sinne, dass das Konfidenzband wirklich die Wahrscheinlichkeit $(1-\alpha)$ annimmt und nicht nur die Wahrscheinlichkeit $\leq (1-\alpha)$ hat]. Dieses ist Aussage des folgenden Satzes, der im Anhang (zunächst als allgemeiner Satz über Konfidenzbänder für Vektorräume) bewiesen wird.

Satz 18.10 (Scheffé) *In der Situation von Beispiel 18.9 gilt*

$$\mathbf{P}\left(|\widehat{\psi}_\xi - \psi_\xi| \;\leq\; \widehat{\sigma}\sqrt{\frac{1}{n} + \frac{(\xi - \overline{X}_n)^2}{\sum_i (X_i - \overline{X}_n)^2}}\;\sqrt{2F_{2,n-2;1-\alpha}} \quad \forall \xi \in \mathbb{R}\right) = 1-\alpha$$

Zum Vergleich: Für $n = 100$ und $\alpha = 0,05$ erhalten wir $F_{1,98;0,95} = 3,94$ und $F_{2,98;0,95} = 3,09$ d.h.

$$\sqrt{F_{1,n-2;1-\alpha}} = 1,98 \quad \text{und} \quad \sqrt{2F_{2,n-2;1-\alpha}} = 2,49$$

Diese Zahlen verdeutlichen beispielhaft die Vergrößerung des Konfidenzintervalls beim Übergang vom einfachen Konfidenzintervall zum gleichmäßigen Konfidenzintervall.

18.1 Anhang: Der Satz von Scheffé

Wir zeigen zunächst eine allgemeinere Version des Satzes von Scheffé - nämlich über ein gleichmäßiges Konfidenzellipsoid für Gauß-Markov-Schätzer in ℓ - dimensionalen Vektorräumen.

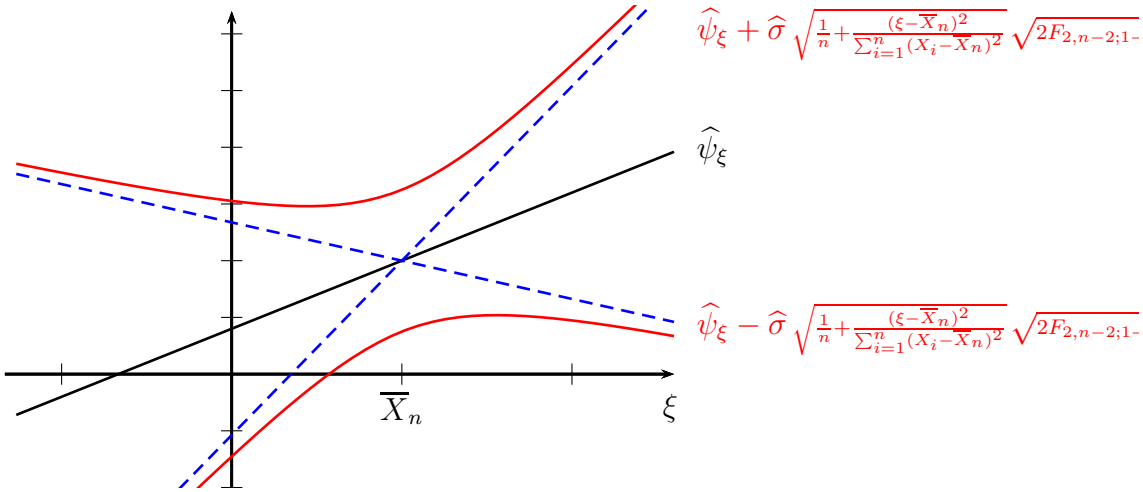


Abbildung 1: Scheffé-Band mit Asymptoten

Sei $\psi = C\beta$ identifizierbar (d.h. $C = C_0 X$) und $\widehat{\psi} = C\widehat{\beta} = C_0 P_{\mathcal{X}} Y$ der Gauß-Markov-Schätzer. Es gilt

$$\operatorname{Var}\widehat{\psi} = \sigma^2 C_0 P_{\mathcal{X}} C_0' \quad \big(= \sigma^2 C(X'X)^{-1}C' \ \text{ falls } r = \operatorname{Rang} X = k\big).$$

Setze $B_\psi := C_0 P_{\mathcal{X}} C_0'$ und $\widehat{\sigma}^2_{\widehat{\psi}} := \widehat{\sigma}^2 B_\psi = \frac{R_0^2}{n-r} B_\psi$.

Satz 18.11 (Scheffé) *Sei $\mathcal{C} \subset \operatorname{Lin}(X')$ ein ℓ-dimensionaler Vektorraum und $L := \{\psi \,|\, \psi = c\beta$ mit $c' \in \mathcal{C}\}$ der zugehörige ℓ-dimensionale Raum identifizierbarer Funktionen. Ferner sei $\widehat{\psi}$ der zu ψ gehörige Gauß-Markov-Schätzer. Dann gilt*

$$\mathbf{P}\big(|\widehat{\psi} - \psi| \le \widehat{\sigma}_{\widehat{\psi}} \sqrt{\ell\, F_{\ell,n-r;1-\alpha}} \quad \forall \psi \in L\big) = 1 - \alpha.$$

Beweis. Sei $\{c_1', \ldots, c_\ell'\}$ eine Basis von $\mathcal{C}$, $C' := (c_1', \ldots, c_\ell')$ und $\underline{\psi} = (\psi_1, \ldots, \psi_\ell)' = C\beta = C_0 X\beta$. Damit gilt $\psi \in L \Rightarrow \psi = \lambda' \underline{\psi}$ mit $\lambda \in \mathbb{R}^\ell$ und $\widehat{\psi} = \lambda' \widehat{\underline{\psi}}$. Satz 18.7 ergibt

$$\mathbf{P}\big((\widehat{\underline{\psi}} - \underline{\psi})' B_{\underline{\psi}}^{-1} (\widehat{\underline{\psi}} - \underline{\psi}) \le \underbrace{\widehat{\sigma}^2\, \ell\, F_{\ell,n-r;1-\alpha}}_{=:\gamma^2}\big) = 1 - \alpha.$$

Mit der Cauchy-Schwarz-Ungleichung folgt

$$(\underline{\widehat{\psi}}-\underline{\psi})B_{\underline{\psi}}^{-1}(\underline{\widehat{\psi}}-\underline{\psi}) \le \gamma^2$$

$$\Leftrightarrow |\lambda'(\underline{\widehat{\psi}}-\underline{\psi})| = |\lambda' B_{\underline{\psi}}^{1/2} B_{\underline{\psi}}^{-1/2}(\underline{\widehat{\psi}}-\underline{\psi})| \le \gamma \underbrace{\sqrt{\lambda' B_{\underline{\psi}} \lambda}}_{=\sqrt{B_{\lambda'\underline{\psi}}}} \quad \forall \lambda \in \mathbb{R}^\ell$$

(die Rückrichtung folgt durch setzen von $\lambda = B_{\underline{\psi}}^{-1}(\widehat{\psi}-\psi)$), und damit

$$\mathbf{P}\big(|\widehat{\psi}-\psi| \le \widehat{\sigma}_{\widehat{\psi}} \sqrt{\ell\, F_{\ell,n-r;1-\alpha}} \quad \forall \psi \in L\big) = 1-\alpha.$$

□

18.12 (Beweis von Satz 18.10 (Scheffé-Konfidenzband))
In der Notation des obigen Satzes gilt

$$\psi_\xi = (1, \xi - \overline{X}_n) \begin{pmatrix} a \\ b \end{pmatrix}$$

und

$$B_{\psi_\xi} = \frac{1}{n} + \frac{(\xi - \overline{X}_n)^2}{\sum_i (X_i - \overline{X}_n)^2}$$

d.h. zu zeigen ist

$$\mathbf{P}\left(|\widehat{\psi}-\psi| \le \widehat{\sigma}_{\widehat{\psi}} \sqrt{2\, F_{2,n-2;1-\alpha}} \quad \forall \psi = (1, \xi - \overline{X}_n)\begin{pmatrix} a \\ b \end{pmatrix} \text{ mit } \xi \in \mathbb{R}\right) = 1-\alpha$$

während Satz 18.11 ergibt

$$\mathbf{P}\left(|\widehat{\psi}-\psi| \le \widehat{\sigma}_{\widehat{\psi}} \sqrt{2\, F_{2,n-2;1-\alpha}} \quad \forall \psi = (\lambda_1, \lambda_2)\begin{pmatrix} a \\ b \end{pmatrix} \text{ mit } \lambda_1, \lambda_2 \in \mathbb{R}\right) = 1-\alpha.$$

Wir verwenden nun die Notation wie im Beweis des obigen Satzes mit $C = I_2$, d.h.

$\underline{\psi} = \beta = \binom{a}{b}$, und $\gamma = \widehat{\sigma}\,\sqrt{2\,F_{2,n-2;1-\alpha}}$. Damit gilt

$$
\begin{aligned}
&|\widehat{\psi}-\psi| \le \widehat{\sigma}_{\widehat{\psi}}\,\sqrt{2\,F_{2,n-2;1-\alpha}} \quad \forall \psi = (\lambda_1,\lambda_2)\binom{a}{b} \text{ mit } \lambda \in \mathbb{R}^2 \\
&\Leftrightarrow \left|\lambda'(\underline{\widehat{\psi}} - \underline{\psi})\right| \le \gamma\,\sqrt{\lambda' B_{\underline{\psi}} \lambda} \quad \forall \lambda \in \mathbb{R}^2 \\
&\Leftrightarrow \left|\lambda' B_{\underline{\psi}}^{-1/2}(\underline{\widehat{\psi}} - \underline{\psi})\right| \le \gamma\,\|\lambda\|_2 \quad \forall \lambda \in \mathbb{R}^2 \\
&\Leftrightarrow \left|\frac{\lambda'}{\|\lambda\|_2} B_{\underline{\psi}}^{-1/2}(\underline{\widehat{\psi}} - \underline{\psi})\right| \le \gamma \quad \forall \lambda_1,\lambda_2 \in \mathbb{R} \\
&\Leftrightarrow \left|\frac{\lambda'}{\|\lambda\|_2} B_{\underline{\psi}}^{-1/2}(\underline{\widehat{\psi}} - \underline{\psi})\right| \le \gamma \quad \text{mit } \lambda_1 = n^{-1/2},\ \forall \lambda_2 \in \mathbb{R} \\
&\Leftrightarrow \left|\lambda' B_{\underline{\psi}}^{-1/2}(\underline{\widehat{\psi}} - \underline{\psi})\right| \le \gamma\,\|\lambda\|_2 \quad \text{mit } \lambda_1 = n^{-1/2},\ \forall \lambda_2 \in \mathbb{R} \\
&\Leftrightarrow \left|\lambda'(\underline{\widehat{\psi}} - \underline{\psi})\right| \le \gamma\,\sqrt{\lambda' B_{\underline{\psi}} \lambda} \quad \text{mit } \lambda_1 = 1, \lambda_2 = \xi - \overline{X}_n \quad \forall \xi \in \mathbb{R} \\
&\Leftrightarrow |\widehat{\psi} - \psi| \le \widehat{\sigma}_{\widehat{\psi}}\,\sqrt{2\,F_{2,n-2;1-\alpha}} \quad \forall \psi = (1,\xi - \overline{X}_n)\binom{a}{b} \text{ mit } \xi \in \mathbb{R}.
\end{aligned}
$$

Damit ist die Aussage bewiesen. □

18.2 Anhang: Ein ausführliches Daten-Beispiel

Dieser Anhang ist eine Zusammenfassung von Kapitel 4g.3 in

Rao, C.R. (1973). *Linear Statistical Inference and its Applications, sec. ed.* John Wiley & Sons, New York.

welches wiederum auf folgenden Originalarbeiten basiert:

Hooke, B.G.E. (1926). A third study of the English skull with special reference to the Farrington Street crania. *Biometrika* **18**, 1-55.

Rao, C.R. and Shaw, D.C. (1948). On a formula for the prediction of cranial capacity. *Biometrics* 4, 247-253.

Beispiel 18.13 (Vorhersage von Schädelvolumina aus einzelnen Knochen)
Das Problem ist die Vorhersage von Schädelvolumina in der Anthropologie falls die aufgefundenen Schädel zerbrochen sind und nur einzelne Knochen vorhanden sind. In solchen Fällen kann man mit Hilfe der linearen Regression eine Vorhersage des Schädelvolumens erstellen.

Voraussetzung ist eine gewisse Anzahl von vollständig erhaltenen Schädeln, bei denen man das Volumen messen kann, und die man dann für die Bestimmung der Regressionsgeraden verwenden kann. Hierfür verwendet man neben dem Volumen V die folgenden Standardgrößen aus der Anthropologie, die sich in den meisten Fällen auch aus einzelnen Knochen bestimmen lassen:

- L ("occipitale Glabellalänge");
- B ("maximale parietale Breite");
- H ("basio-bregmatische Höhe").

Konkret versuchen wir eine Regressionsgerade aus den $n = 86$ männlichen Schädel der „Farringdon Street Reihe“ (Hooke, 1926) zu gewinnen.

(a) Modell:
Das bisher verwendete Modell $V = \beta_1 + \beta_2 L + \beta_3 B + \beta_4 H + \varepsilon$ eignet sich hier nicht, da wir ein Volumen betrachten. Besser ist:

$$V = \gamma L^{\beta_1} B^{\beta_2} H^{\beta_3}$$

d.h. mit $Y := \log_{10} V$

$$\begin{aligned} Y &= \log_{10}\gamma + \beta_1 \log_{10} L + \beta_2 \log_{10} B + \beta_3 \log_{10} H + \varepsilon \\ &= \alpha + \beta_1\big(\log L - \overline{\log L}\big) + \beta_2\big(\log B - \overline{\log B}\big) + \beta_3\big(\log H - \overline{\log H}\big) + \varepsilon \end{aligned} \tag{10}$$

wobei $\alpha = \log\gamma + \beta_1 \overline{\log L} + \beta_2 \overline{\log B} + \beta_3 \overline{\log H}$. Damit lineares Modell

$$Y = X \begin{pmatrix} \alpha \\ \beta_1 \\ \beta_2 \\ \beta_3 \end{pmatrix} + \varepsilon \quad \text{mit } X = (1, \log L - \overline{\log L}, \log B - \overline{\log B}, \log H - \overline{\log H}).$$

Im Fall der $n = 86$ Beobachtungen sind Y, ε, X Vektoren bzw. eine Matrix. Ferner gilt $r = k = 4$.

Die folgenden numerischen Werte kann man mit praktisch jedem Statistik-Programm erhalten:

(b) Gauß-Markov-Schätzer:
Es gilt mit $\beta = (\alpha, \beta_1, \beta_2, \beta_3)'$

$$\widehat{\beta} = (X'X)^{-1}X'Y = (3,1685; 0,878; 1,041; 0,733)'$$

sowie $R_0^2 = 0,0278$ und $\widehat{\sigma}^2 = R_0^2/(n-r) = 0,00034$, d.h. die geschätzte Gleichung lautet:

$$V = 0,0024 \cdot L^{0,878} B^{1,041} H^{0,733} \; [\times \exp\varepsilon]$$

[der Wert $\gamma = 0,00241$ erscheint etwas unplausibel, ist aber vermutlich auf die Wahl der Einheiten zurückzuführen].

(c) Test der Hypothese $H : \beta_1 = \beta_2 = \beta_3 = 0$:
(die Hypothese bedeutet, dass L, B, H gar keinen Erklärungsbeitrag liefern - z.B. weil man die falschen Knochen genommen hat). Es gilt

$$H : \beta_1 = \beta_2 = \beta_3 = 0 \quad \Leftrightarrow \quad \begin{pmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} \alpha \\ \beta_1 \\ \beta_2 \\ \beta_3 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}$$

Satz 18.5: Testgröße $F = \frac{R_1^2 - R_0^2}{l} \Big/ \frac{R_0^2}{n-r}$ mit $\ell = \operatorname{Rang} H = 3$ und $n - r = 86 - 4 = 82$. Ferner erhält man

$$R_1^2 = \min_{\beta_1=\beta_2=\beta_3=0} \|Y - X\beta\|^2 = 0,1269$$

[Berechnung als Fehlerquadratsumme mit „reduzierter“ Designmatrix X]. Die Ergebnisse werden traditionell in folgender „Regressionstabelle“ zusammengefasst:

Bezeichnung	Größe	FG	SS	MS	F
Regression	$R_1^2 - R_0^2$	$l = 3$	$0,0991$	$0,03303$	$97,2$
Rest	R_0^2	$n - r = 82$	$0,0278$	$0,00034$	
Summe	R_1^2	85	$0,1269$		

Dabei steht FG für Freiheitsgrade, SS für Sum of Squares und MS für Mean of Squares = SS/FG.

Nachschlagen in einer Tabelle ergibt $F_{3,82;0,01} = 4,03$, d.h. die Hypothese wird zum Niveau $\alpha = 0,01$ abgelehnt und wir erhalten einen signifikaten Erklärungsbeitrag von L, B, H.

(d) Test der Hypothese $H : \beta_1 = \beta_2 = \beta_3$:
Es gilt

$$H : \beta_1 = \beta_2 = \beta_3 \quad \Leftrightarrow \quad \begin{pmatrix} 0 & 1 & -1 & 0 \\ 0 & 0 & 1 & -1 \end{pmatrix} \begin{pmatrix} \alpha \\ \beta_1 \\ \beta_2 \\ \beta_3 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$$

Es gilt $\ell = 2$ und $R_1^2 = 0,0288$ [zur Berechnung von R_1^2 schreibt man das Modell unter der Hypothese am besten in der Form $Y = \log V = \alpha' + \beta_1(\log L + \log B + \log H)$]. Damit erhält man folgende „Regressionstabelle“:

Bezeichnung	Größe	FG	SS	MS	F
Regression	$R_1^2 - R_0^2$	$l = 2$	$0,0010$	$0,0005$	$1,47$
Rest	R_0^2	$n - r = 82$	$0,0278$	$0,00034$	
Summe	R_1^2	84	$0,0288$		

Es gilt $F_{2,82;0,01} = 4,88$, d.h. man lehnt die Hypothese nicht ab (die Unterschiede bei den β_i sind nicht signifikant).

(e) Test der Hypothese $H : \beta_1 = \beta_2 = \beta_3 = 1$:
(diese Hypothese passt am besten zu der Tatsache, dass es sich um ein Volumen handelt). Obwohl sich diese Hypothese nur in der Form $H'\beta = \xi$ mit $\xi \neq 0$ schreiben lässt, gilt die Aussage aus Satz18.5 zum F-Test trotzdem. Es gilt $\ell = 3$ und $R_1^2 = 0,0304$ und damit analog zu oben $F = 2,54$.

Andererseits gilt $F_{3,82;0,01} = 4,03$ (zum Niveau $0,01$ lehnt man nicht ab) und $F_{3,82;0,1} = 2,37$ (zum Niveau $0,1$ lehnt man ab).

(f) Prognose des mittleren Schädelvolumens:
Für neue Daten $L_0 = 198,5;\ B_0 = 147;\ H_0 = 131$ erhalten wir aus Gleichung (10) den zugehörigen Wert $\mathbf{E}Y_0 = 3,2069$ bzw. das Volumen $\tilde{V}_0 := 10^{\mathbf{E}Y_0} = 1610$. Wir berechnen nun analog zu Beispiel 18.9 dazu das Konfidenzintervall aus Satz 18.7 (wobei das hier wegen $r = \text{Rang}X = 4$ anstelle von $r = 2$ etwas komplizierter ist). Sei

$$\tilde{X} = \left(\log L_i - \overline{\log L}, \log B_i - \overline{\log B}, \log H_i - \overline{\log H}\right)_{i=1,\ldots,n}$$

und

$$\tilde{C} = \left(\log L_0 - \overline{\log L}, \log B_0 - \overline{\log B}, \log H_0 - \overline{\log H}\right)$$

sowie $X = (\mathbf{1}, \tilde{X})$ und $C = (1, \tilde{C})$. Das Konfidenzintervall für $\psi = C\beta = \mathbf{E}Y_0$ ist mit $\widehat{\psi} = C\widehat{\beta}$ nach Satz 18.7

$$K(Y) = \left\{\psi \;\middle|\; |\widehat{\psi} - \psi| \leq \widehat{\sigma}\sqrt{B}\sqrt{F_{1,n-4;1-\alpha}}\right\}$$

wobei

$$B = C(X'X)^{-1}C' = C \begin{pmatrix} n & 0 \\ 0 & \tilde{X}'\tilde{X} \end{pmatrix}^{-1} C' = \frac{1}{n} + \tilde{C}(\tilde{X}'\tilde{X})^{-1}\tilde{C}'.$$

Numerisch erhalten wir für $\mathbf{E}Y_0$ das $0,95$ - Konfidenzintervall $[3,1994; 3,2144]$ und (durch monotone Transformation!) für $\tilde{V}_0 := 10^{\mathbf{E}Y_0}$ das $0,95$ - Konfidenzintervall $[1583, 1638]$.

Das $0,95$ - Konfidenzintervall für Y_0 (und für $V_0 := 10^{Y_0}$) ist wegen der Streung von ε_0 natürlich größer. Man kann zeigen (Ü-Aufgabe), dass

$$K(Y) = \left\{ y \;\middle|\; |\widehat{\psi} - y| \leq \widehat{\sigma}\sqrt{B_{Y_0}}\sqrt{F_{1,n-4;1-\alpha}} \right\}$$

mit

$$B_{Y_0} = 1 + B = 1 + \frac{1}{n} + \tilde{C}(\tilde{X}'\tilde{X})^{-1}\tilde{C}'$$

ein $(1-\alpha)$ - Konfidenzintervall für Y_0 ist. Numerisch erhält man für Y_0 das $0,95$ - Konfidenzintervall $[3,1693; 3,2445]$ und für $V_0 = 10^{Y_0}$ das $0,95$ - Konfidenzintervall $[1476, 1755]$ (hier ist an sich die Bezeichnung Prognoseintervall anstelle von Konfidenzintervall üblich, da es sich nicht mehr um ein Intervall für einen Parameter, sondern für eine Zufallsvariable handelt).

(g) Aspekte der anthropologischen Studie:

An 2 Fundstellen hat man jeweils die empirischen Mittel des Volumens der erhaltenen Schädel und des prognostizierten Volumens $\widehat{V}$ gebildet:

Farringdon-Street-Serie:

$n = 86$ heile Schädel	$\bar{V}$= 1481,3
$n = 29$ zerbrochene Schädel	$\bar{\widehat{V}}$= 1498,3

Moorfields-Serie:

$n = 22$ heile Schädel	$\bar{V}$= 1473,8
$n \approx 40$ zerbrochene Schädel	$\bar{\widehat{V}}$= 1490,7

(genauer wurde bei der Moorfields-Serie die Prädiktionsformel auf die Mittelwerte

der Größen L, B, H angewendet).

Damit waren die zerbrochenen Schädel im Mittel größer (es ist plausibel, dass die größeren Schädel eher zerbrechen). Für die publizierten Schädelgrößen aus älteren anthropologischen Studien (basierend auf nicht zerbrochenen Schädeln) stellt sich damit die Frage, ob die Schätzer nicht zu klein sind und nach oben korrigiert werden müssen. Und es stellt sich natürlich die Frage, wie man die Schätzer korrigiert.

18.3 Anhang: Diverses

18.14 (Konfidenzellipsoide für zentrierte und nicht-zentrierte Schätzer)
Das zentrierte Modell aus Beispiel 17.4 lautet $Y_z = X_z \binom{a'}{b} + \varepsilon$. Mit Satz 18.7 erhalten wir daher als $(1-\alpha)$-Konfidenzellipsoid für $\binom{a'}{b}$ (man beachte, dass aufgrund der Transformation $\widehat{a'} = 0$ gilt)

$$K(Y) = \left\{ \binom{a'}{b} \;\middle|\; \binom{-a'}{\widehat{b}-b}' (X_z' X_z) \binom{-a'}{\widehat{b}-b} \leq 2\,\widehat{\sigma}^2\, F_{2,n-2;1-\alpha} \right\}$$

und wegen $a' = a - \overline{Y}_n + b\overline{X}_n$ damit als Konfidenzellipsoid für $\binom{a}{b}$

$$K(Y) = \left\{ \binom{a}{b} \;\middle|\; a = a' + \overline{Y}_n - b\overline{X}_n \text{ und } \binom{-a'}{\widehat{b}-b}' (X_z' X_z) \binom{-a'}{\widehat{b}-b} \leq 2\,\widehat{\sigma}^2\, F_{2,n-2;1-\alpha} \right\}.$$

Es gilt nun

$$\begin{aligned}(X'X) = \begin{pmatrix} n & \sum X_i \\ \sum X_i & \sum X_i^2 \end{pmatrix} &= \begin{pmatrix} 1 & 0 \\ \overline{X}_n & 1 \end{pmatrix} \begin{pmatrix} n & 0 \\ 0 & \sum_i (X_i - \overline{X}_n)^2 \end{pmatrix} \begin{pmatrix} 1 & \overline{X}_n \\ 0 & 1 \end{pmatrix} \\ &= \begin{pmatrix} 1 & 0 \\ \overline{X}_n & 1 \end{pmatrix} (X_z' X_z) \begin{pmatrix} 1 & \overline{X}_n \\ 0 & 1 \end{pmatrix}\end{aligned}$$

d.h.

$$(X_{\mathrm{z}}'X_{\mathrm{z}}) = \begin{pmatrix} 1 & 0 \\ -\overline{X}_n & 1 \end{pmatrix} (X'X) \begin{pmatrix} 1 & -\overline{X}_n \\ 0 & 1 \end{pmatrix}.$$

Wegen $a' = a - \overline{Y}_n + b\overline{X}_n$ und $\widehat{a} = \overline{Y}_n - \widehat{b}\,\overline{X}_n$ erhält man

$$\begin{aligned}
&\begin{pmatrix} -a' \\ \widehat{b} - b \end{pmatrix}' (X_{\mathrm{z}}'X_{\mathrm{z}}) \begin{pmatrix} -a' \\ \widehat{b} - b \end{pmatrix} \\
&\quad = \begin{pmatrix} -a + \overline{Y}_n - b\overline{X}_n \\ \widehat{b} - b \end{pmatrix}' \begin{pmatrix} 1 & 0 \\ -\overline{X}_n & 1 \end{pmatrix} (X'X) \begin{pmatrix} 1 & -\overline{X}_n \\ 0 & 1 \end{pmatrix} \begin{pmatrix} -a + \overline{Y}_n - b\overline{X}_n \\ \widehat{b} - b \end{pmatrix} \\
&\quad = \begin{pmatrix} \widehat{a} - a + (\widehat{b} - b)\overline{X}_n \\ \widehat{b} - b \end{pmatrix}' \begin{pmatrix} 1 & 0 \\ -\overline{X}_n & 1 \end{pmatrix} (X'X) \begin{pmatrix} 1 & -\overline{X}_n \\ 0 & 1 \end{pmatrix} \begin{pmatrix} \widehat{a} - a + (\widehat{b} - b)\overline{X}_n \\ \widehat{b} - b \end{pmatrix} \\
&\quad = \begin{pmatrix} \widehat{a} - a \\ \widehat{b} - b \end{pmatrix}' (X'X) \begin{pmatrix} \widehat{a} - a \\ \widehat{b} - b \end{pmatrix}
\end{aligned}$$

und damit

$$K(Y) = \left\{ \begin{pmatrix} a \\ b \end{pmatrix} \;\middle|\; \begin{pmatrix} \widehat{a} - a \\ \widehat{b} - b \end{pmatrix}' (X'X) \begin{pmatrix} \widehat{a} - a \\ \widehat{b} - b \end{pmatrix} \leq 2\,\widehat{\sigma}^2\, F_{2,n-2;1-\alpha} \right\},$$

d.h. das Konfidenzintervall aus Beispiel 18.8. □

Bayes-Statistik

Matthias Katzfuß[1]

8. Oktober 2012

[1]Institut für Angewandte Mathematik, Universität Heidelberg. Email: katzfuss@gmail.com.

Inhaltsverzeichnis

Kapitel 1

Einführung

Dieses Skript basiert auf den folgenden zwei Büchern:

- Held, L. 2008. *Methoden der statistischen Inferenz*. Spektrum.
- Gelman, A., Carlin, J.B., Stern, H.S., and Rubin, D.B. 2004. *Bayesian Data Analysis*, second edition. Chapman & Hall/CRC.

1.1 Bemerkung (Notation)
Seien X und Y Zufallsvariablen oder Zufallsvektoren. In diesem Skript bezeichnet $[Y]$ die Verteilung bzw. die Dichte von Y, und $[Y|X]$ die bedingte Verteilung oder Dichte von Y gegeben X. Bei Dichten werden für die Argumente häufig Kleinbuchstaben verwendet. Ob die Verteilung oder die Dichte gemeint ist, ergibt sich aber sowieso aus dem Kontext.

1.2 Bemerkung/Definition (Bayes-Inferenz)
Bisher haben wir klassische, frequentistische Inferenz betrachtet. Dort wird der Parameter(vektor) θ als unbekannte, aber feste Größe betrachtet.
In der Bayes-Inferenz ist θ eine Zufallsvariable (oder ein Zufallsvektor), mit einer *Priori-Verteilung* $[\theta]$. Nachdem Daten $X = x$ beobachtet wurden, basiert die Inferenz bezüglich θ dann ausschließlich auf der *Posteriori-Verteilung*, $[\theta|x]$, der bedingten Verteilung des Parameters gegeben die Daten.

1.3 Bemerkung (Posteriori-Verteilung)
Gegeben eine *Likelihood*, $[x|\theta]$, und eine Priori, $[\theta]$, ergibt sich für beobachtete Daten $X = x$ die Posteriori von θ aus dem Satz von Bayes:

$$[\theta|x] = \frac{[x|\theta][\theta]}{[x]}.$$

Die marginale Verteilung der Daten kann für stetig verteiltes θ durch $\int [x|\theta][\theta]d\theta$ berechnet werden, für diskret verteiltes θ durch $\sum_\theta [x|\theta][\theta]$. Da diese Verteilung nicht von θ abhängt, ist

die Posteriori proportional zum Produkt von Likelihood und Priori:

$$[\theta|x] \propto [x|\theta][\theta].$$

Wir nehmen im Folgenden an, dass die Priori eine stetige Verteilung ist (was sie in der Praxis meist ist). Alle Resultate gelten aber auch für diskret verteiltes θ, wobei Integrale über θ dann durch Summen ersetzt werden müssen.

1.4 Bemerkung (Bayesianische Punkt- und Intervallschätzung)
Da Bayesianische Inferenz ausschließlich auf der Posteriori beruht, sind Punktschätzer immer Zusammenfassungen der Posteriori, z.B. der Posteriori-Erwartungswert. Bayesianische Konfidenzintervalle (sogenannte *Kredibilitätsintervalle*) sind auch aus der Posteriori bestimmt. So hat ein $(1-\alpha)$-Kredibilitätsintervall K für θ Posteriori-Wahrscheinlichkeit $(1-\alpha)$, d.h. $P(\theta \in K|X=x) = 1-\alpha$. Mehr dazu in Kapitel 3.

1.5 Definition (Die Beta-Verteilung $Be(\alpha,\beta)$)
Schreibweise: $p \sim Be(\alpha,\beta)$, $\alpha,\beta > 0$.
Dichte: $[p] = \frac{1}{B(\alpha,\beta)} p^{\alpha-1}(1-p)^{\beta-1}$, für $p \in (0,1)$.
Die Normierungskonstante ist gegeben durch die Beta-Funktion $B(\alpha,\beta)$, definiert als

$$B(\alpha,\beta) = \int_0^1 p^{\alpha-1}(1-p)^{\beta-1}dp = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}.$$

Erwartungswert:

$$\begin{aligned} E(p) &= \int_0^1 p\frac{1}{B(\alpha,\beta)}p^{\alpha-1}(1-p)^{\beta-1}dp = \frac{1}{B(\alpha,\beta)}\int_0^1 p^{(\alpha+1)-1}(1-p)^{\beta-1}dp \\ &= \frac{B(\alpha+1,\beta)}{B(\alpha,\beta)} = \frac{\Gamma(\alpha+1)\Gamma(\beta)}{\Gamma(\alpha+1+\beta)}\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} = \frac{\alpha}{\alpha+\beta}, \end{aligned}$$

da $\Gamma(z+1) = z\Gamma(z)$ (kann man durch partielle Integration zeigen).
Der Modus ist gegeben durch $\frac{\alpha-1}{\alpha+\beta-2}$ für $\alpha,\beta > 1$.
Die Beta-Verteilung wird oft als Priori für Wahrscheinlichkeiten verwendet (siehe Beispiele 1.6 und 2.3).

1.6 Beispiel (Glühbirnen)
Ein Glühbirnenhersteller behauptet, dass höchstens 1% der von ihm hergestellten Glühbirnen defekt sind. Basierend auf einer Zufallsstichprobe von n Birnen, von der $X = x$ Birnen defekt waren, will er die Wahrscheinlichkeit berechnen, dass seine Behauptung falsch ist. Als Priori für p, den Anteil der defekten Glühbirnen, nehmen wir $p \sim U(0,1)$, welches der Beta-Verteilung mit Parametern $\alpha = 1$, $\beta = 1$ entspricht.
Annahme: Glühbirnen unabhängig, d.h. $X|p \sim Bin(n,p)$.

Posteriori für beliebige $\alpha, \beta > 0$:

$$[p|x] \propto [x|p][p] \propto p^x(1-p)^{n-x} \cdot p^{\alpha-1}(1-p)^{\beta-1} = p^{(\alpha+x)-1}(1-p)^{(\beta+n-x)-1},$$

d.h.

$$p|x \sim Be(\alpha + x, \beta + n - x).$$

Damit ist der Posteriori-Erwartungswert gegeben durch:

$$E(p|x) = \frac{\alpha + x}{\alpha + \beta + n} = \frac{\alpha + \beta}{\alpha + \beta + n} \cdot \frac{\alpha}{\alpha + \beta} + \frac{n}{\alpha + \beta + n} \cdot \frac{x}{n},$$

d.h. der Posteriori-Erwartungswert ist ein gewichtetes Mittel aus dem Priori-Erwartungswert $\frac{\alpha}{\alpha+\beta}$ und dem MLE $\bar{x} = x/n$, wobei des Gewicht des MLE mit steigendem Stichprobenumfang n zunimmt.
Für den konkreten Fall der Priori-Gleichverteilung gilt $E(p|x) = \frac{x+1}{n+2}$, und der Posteriori-Modus ist $\bar{x} = x/n$. Die gesuchte Wahrscheinlichkeit ist gegeben durch

$$P(p > 0.01|x) = \int_{0.01}^{1} [p|x]dp = \int_{0.01}^{1} \frac{1}{B(1+x, 1+n-x)} p^x(1-p)^{n-x} dp$$

Dieses Integral muss numerisch bestimmt werden, z.B. für $n = 100$, $x = 1$, mit Hilfe von R:

```
> 1-pbeta( q=.01 , shape1 = 1+1 , shape2 = 1+100-1 )
[1] 0.7320647
```

1.7 Bemerkung (Bayesian Updating)
Analyse von sequentiell verfügbar werdenden Daten kann in der Bayes-Inferenz sehr einfach durchgeführt werden. Dabei "wird die alte Posteriori zur neuen Priori".
Man betrachte zwei Beobachtungen X_1 und X_2, die sequentiell verfügbar werden und bedingt unabhängig sind gegeben θ. Dann gilt für die Posteriori von θ gegeben X_1 und X_2:

$$[\theta|x_1, x_2] \propto [x_1, x_2|\theta][\theta] = [x_2|\theta]\{[x_1|\theta][\theta]\} \propto [x_2|\theta][\theta|x_1].$$

Diese Posteriori ist also proportional zur Likelihood von X_2 multipliziert mit der Posteriori von θ gegeben x_1.
Man berechnet also nach Beobachtung von x_1 zunächst die Posteriori von θ gegeben x_1. Wenn dann x_2 beobachtet wird, ergibt sich die Posteriori von θ gegeben alle bis dahin verfügbaren Daten (also x_1 und x_2) also erneut aus dem Satz von Bayes, wobei nun die neue Priori die alte Posteriori ist.

1.8 Beispiel (Fortsetzung von Beispiel 1.6)
Der Glühbirnenhersteller aus Beispiel 1.6 ist nach der ersten Stichprobe noch nicht zufrieden mit der Präzision der Posteriori von p. Um die Genauigkeit der Schätzung zu verbessern erhebt

er unabhängig von der ersten Stichprobe noch eine zweite Stichprobe $X_2 = x_2$ vom Umfang n_2, d.h. $X_2|p \sim Bin(n_2, p)$.
Insgesamt ergibt sich damit die Posteriori:

$$[p|x, x_2] \propto [x_2|p][p|x] \propto p^{x_2}(1-p)^{n_2-x_2} \cdot p^{(\alpha+x)-1}(1-p)^{(\beta+n-x)-1},$$

d.h.

$$p|x, x_2 \sim Be(\alpha + x + x_2, \beta + n - x + n_2 - x_2).$$

Für $\alpha = \beta = 1$, $n = 100$, $x = 1$, $n_2 = 200$, und $x_2 = 0$ gilt (mit Hilfe von R):

$$P(p > 0.01|x, x_2) = \int_{0.01}^{1} [p|x, x_2]dp = \int_{0.01}^{1} \frac{1}{B(2, 300)} p(1-p)^{299} dp \approx 0.196.$$

Kapitel 2

Wahl der Priori

In vielen Situationen liegt Vorwissen über bestimmte Parameter vor, das dann in Form einer Wahrscheinlichkeitsverteilung als Priori verwendet werden kann. Dazu ist es oft hilfreich, bestimmte Verteilungsklassen zu wählen, die die spätere Inferenz erleichtern (s. Abschnitt 2.1). Falls kein Vorwissen vorliegt, kann eine sogenannte nicht-informative Priori verwendet werden (s. Abschnitt 2.3).

2.1 Konjugierte Prioris

2.1 Definition (Konjugierte Verteilungsklassen)
Eine Klasse $\mathcal{P}$ von Priori-Verteilungen heißt konjugiert bezüglich einer Familie von Likelihoods $\mathcal{F} = \{[x|\theta]|\theta \in \Theta\}$ bei Daten $X = x$, falls für alle $[\theta] \in \mathcal{P}$ auch $[\theta|x] \in \mathcal{P}$ gilt.

2.2 Bemerkung
Da die Posteriori proportional zum Produkt von Likelihood und Priori ist, ist die konjugierte Verteilungsklasse oft aus der Likelihood ersichtlich.
Zum Beispiel gilt für $X|p \sim Bin(n, p)$:

$$[x|p] = \binom{n}{x} p^x (1-p)^{n-x} \propto p^x (1-p)^{n-x},$$

wobei hier Proportionalität bezüglich p gemeint ist. Eine Dichte der Form, $[p] \propto p^a (1-p)^b$, ist damit konjugiert. Wie wir bereits in Beispiel 1.6 gesehen haben, ist also die Beta-Verteilung zur Binomialverteilung konjugiert.

2.3 Beispiel (Geometrische Verteilung)
Für $X|p \sim \mathcal{G}(p)$ gilt:

$$[x|p] = p(1-p)^{x-1},$$

daher ist auch hier die Beta-Verteilung konjugiert. Für $p \sim Be(\alpha, \beta)$ gilt:

$$[p|x] \propto [x|p][p] \propto p(1-p)^{x-1} \cdot p^{\alpha-1}(1-p)^{\beta-1} = p^{(\alpha+1)-1}(1-p)^{(\beta+x-1)-1},$$

d.h. die Posteriori ist:

$$p|x \sim Be(\alpha+1, \beta+x-1).$$

2.4 Beispiel (Multivariate Normalverteilung)
Seien $\mathbf{X}_i|\boldsymbol{\mu} \overset{iid}{\sim} N_p(\boldsymbol{\mu}, \Sigma)$, $i = 1, \ldots, n$, mit Σ bekannt, und sei $\mathbf{x}_{1:n} := (\mathbf{x}_1', \ldots, \mathbf{x}_n')'$. Dann gilt:

$$\begin{aligned}
[\mathbf{x}_{1:n}|\boldsymbol{\mu}] &\propto \exp\{-(1/2)\sum_{i=1}^{n}(\mathbf{x}_i - \boldsymbol{\mu})'\Sigma^{-1}(\mathbf{x}_i - \boldsymbol{\mu})\} \\
&= \exp\{-(1/2)\Big(\sum_{i=1}^{n}(\mathbf{x}_i - \bar{\mathbf{x}})'\Sigma^{-1}(\mathbf{x}_i - \bar{\mathbf{x}}) + 2\sum_{i=1}^{n}(\mathbf{x}_i - \bar{\mathbf{x}})'\Sigma^{-1}(\bar{\mathbf{x}} - \boldsymbol{\mu}) \\
&\quad + n(\bar{\mathbf{x}} - \boldsymbol{\mu})'\Sigma^{-1}(\bar{\mathbf{x}} - \boldsymbol{\mu})\Big)\} \\
&\propto \exp\{-(1/2)(\bar{\mathbf{x}} - \boldsymbol{\mu})'n\Sigma^{-1}(\bar{\mathbf{x}} - \boldsymbol{\mu})\},
\end{aligned}$$

da $\sum_{i=1}^{n}(\mathbf{x}_i - \bar{\mathbf{x}})'\Sigma^{-1}(\bar{\mathbf{x}} - \boldsymbol{\mu}) = 0$. Hier ist also die multivariate Normalverteilung konjugiert. Mit $\boldsymbol{\mu} \sim N_p(\boldsymbol{\nu}, \Lambda)$ gilt:

$$[\boldsymbol{\mu}|\mathbf{x}_{1:n}] \propto [\mathbf{x}_{1:n}|\boldsymbol{\mu}][\boldsymbol{\mu}] \propto \exp\{-(1/2)\big((\bar{\mathbf{x}} - \boldsymbol{\mu})'n\Sigma^{-1}(\bar{\mathbf{x}} - \boldsymbol{\mu}) + (\boldsymbol{\mu} - \boldsymbol{\nu})'\Lambda^{-1}(\boldsymbol{\mu} - \boldsymbol{\nu})\big)\}.$$

Definiere $A := \Lambda^{-1} + n\Sigma^{-1}$ und $\mathbf{b} := \Lambda^{-1}\boldsymbol{\nu} + n\Sigma^{-1}\bar{\mathbf{x}}$. Dann:

$$[\boldsymbol{\mu}|\mathbf{x}_{1:n}] \propto \exp\{-(1/2)\big(\boldsymbol{\mu}'A\boldsymbol{\mu} - 2\boldsymbol{\mu}'\mathbf{b}\big)\} \propto \exp\{-(1/2)(\boldsymbol{\mu} - A^{-1}\mathbf{b})'A(\boldsymbol{\mu} - A^{-1}\mathbf{b})\},$$

d.h. die Posteriori ist:

$$\boldsymbol{\mu}|\mathbf{x}_{1:n} \sim N_p(A^{-1}\mathbf{b}, A^{-1}).$$

2.2 Uneigentliche Prioris

2.5 Definition (Uneigentliche Priori-Verteilung)
Eine Priori-Verteilung $[\theta]$ für die $\int[\theta]d\theta = \infty$ bzw. $\sum_\theta[\theta] = \infty$ gilt, heißt *uneigentliche* Priori.

2.6 Bemerkung
Uneigentliche Prioris können dann verwendet werden, wenn die Posteriori von θ wieder integrierbar ist. Sie ergeben sich oft als Grenzfall von eigentlichen, konjugierten Prioris.

2.7 Beispiel (Beta-/Binomialverteilung)

Wenn im Fall von Beispiel 1.6 $\alpha = \beta = 0, d.h.$ "$p \sim Be(0,0)$", dann ist die resultierende Posteriori

$$p|x \sim Be(x, n-x)$$

eine integrierbare Verteilung falls $x > 0$ und $n - x > 0$. Der Posteriori-Erwartungswert, $E(p|x) = \bar{x}$, stimmt nun mit dem MLE von p überein.

2.8 Beispiel (Multivariate Normalverteilung)
Wenn im Fall von Beispiel 2.4 "$\Lambda = \infty I_p$", d.h. die Priori von $\boldsymbol{\mu}$ ist eine Gleichverteilung auf $\mathbb{R}^p$, dann ergibt sich $A = n\Sigma^{-1}$ und $\mathbf{b} = n\Sigma^{-1}\bar{\mathbf{x}}$, und damit

$$\boldsymbol{\mu}|\mathbf{x}_{1:n} \sim N_p(\bar{\mathbf{x}}, \Sigma/n).$$

2.9 Beispiel (Lineares Modell)
Das lineare Modell (für festes X und σ^2) kann geschrieben werden als

$$\mathbf{Y}|\boldsymbol{\beta} \sim N_n(X\boldsymbol{\beta}, \sigma^2 I_n).$$

Die konjugierte Verteilung für den Parametervektor ist erneut eine multivariate Normalverteilung. Verwendet man die uneigentliche Priori aus Beispiel 2.8, $[\mu] \propto 1$, ergibt sich analog zu Beispielen 2.4 und 2.8 die Posteriori

$$[\boldsymbol{\beta}|\mathbf{y}] \propto [\mathbf{y}|\boldsymbol{\beta}] \propto \exp\left\{-\frac{1}{2\sigma^2}(\mathbf{y} - X\boldsymbol{\beta})'(\mathbf{y} - X\boldsymbol{\beta})\right\} \propto \exp\left\{-\frac{1}{2}\left(\boldsymbol{\beta}'(X'X/\sigma^2)\boldsymbol{\beta} - 2\boldsymbol{\beta}'X'\mathbf{y}/\sigma^2\right)\right\},$$

d.h. falls X vollen Rang k hat, gilt:

$$\boldsymbol{\beta}|\mathbf{y} \sim N_k(\hat{\boldsymbol{\beta}}, \sigma^2(X'X)^{-1}),$$

mit $\hat{\boldsymbol{\beta}} := (X'X)^{-1}X'\mathbf{y}$.

2.3 Nicht-informative Prioris

2.10 Bemerkung (Problem der Priori-Gleichverteilung)
Eine (eventuell uneigentliche) Gleichverteilung auf Θ scheint auf den ersten Blick die beste Art, die Priori für θ möglichst nicht-informativ zu gestalten. Dabei ergibt sich allerdings folgendes Problem für andere Parametrisierungen (Transformationen) des unbekannten Parameters.
Sei $\psi := g(\theta)$, wobei $g(\cdot)$ eine bijektive Abbildung ist. Für die Dichte von ψ gilt nach dem Transformationssatz für Dichten (Satz 6.16 im Dahlhaus-Skript):

$$f_\psi(\psi) = f_\theta(g^{-1}(\psi)) \left|\frac{\partial g^{-1}(\psi)}{\partial \psi}\right|.$$

Im Fall einer Priori-Gleichverteilung für θ, d.h. $f_\theta(\theta) \propto 1$, ist die erste Größe auf der rechten Seite konstant. Die zweite Größe ist allerdings nur konstant, falls $g(\cdot)$ eine lineare Abbildung ist. Daher resultiert die Annahme einer Priori-Gleichverteilung für θ im Allgemeinen nicht in einer Priori-Gleichverteilung für ψ. Dies steht allerdings im Widerspruch zur Wahl einer Priori-Gleichverteilung für ψ, falls gleich von Anfang an die Parametrisierung mit ψ verwendet worden wäre.
Die sogenannte Jeffreys-Priori löst dieses Problem.

2.11 Definition (Jeffreys-Priori)
Sei X eine Zufallsvariable oder ein Zufallsvektor mit Dichte $[x|\theta]$ und θ der unbekannte, skalare Parameter. Die Jeffreys-Priori hat die Form,

$$f_\theta(\theta) \propto \sqrt{I(\theta)},$$

wobei $I(\theta)$ die Fisher-Information von θ bezeichnet.

2.12 Satz (Invarianz der Jeffreys-Priori)
Die Jeffreys-Priori ist invariant bezüglich bijektiven Transformationen von θ, d.h. bei $\psi := g(\theta)$ und $g(\cdot)$ bijektiv, folgt aus

$$f_\theta(\theta) \propto \sqrt{I_\theta(\theta)},$$

mit Hilfe des Transformationssatzes für Dichten, dass

$$f_\psi(\psi) \propto \sqrt{I_\psi(\psi)}.$$

Beweis. Bezeichne $f_x(x;\theta)$ die Likelihood. Die Fisher-Info von ψ kann mit Hilfe der Kettenregel geschrieben werden als:

$$\begin{aligned} I_\psi(\psi) &= E\left(\frac{\partial}{\partial\psi}\log f_x(X; g^{-1}(\psi))\right)^2 = E\left(\frac{\partial}{\partial\theta}\log f_x(X;\theta)\right)^2 \left(\frac{\partial g^{-1}(\psi)}{\partial\psi}\right)^2 \\ &= I_\theta(\theta)\left(\frac{\partial g^{-1}(\psi)}{\partial\psi}\right)^2 \end{aligned}$$

Bei der Wahl der Jeffreys-Priori für θ, d.h. $f_\theta(\theta) \propto \sqrt{I_\theta(\theta)}$, folgt daher mit dem Transformationssatz für Dichten:

$$f_\psi(\psi) = f_\theta(\theta)\left|\frac{\partial g^{-1}(\psi)}{\partial\psi}\right| \propto \left(I_\theta(\theta)\left|\frac{\partial g^{-1}(\psi)}{\partial\psi}\right|^2\right)^{1/2} = \sqrt{I_\psi(\psi)}.$$

□

2.13 Beispiel (Jeffreys-Priori für die Parameter einer Normalverteilung)
Sei $X|\mu \sim N(\mu, \sigma^2)$ mit σ^2 bekannt, und damit $\theta = \mu$ der unbekannte Parameter. Wir wissen aus Bsp. 15.10 im Dahlhaus-Skript, dass $I(\mu) = 1/\sigma^2$. Die Jeffreys-Priori für μ ist daher gegeben durch

$$[\mu] \propto 1,$$

was einer Gleichverteilung auf $\mathbb{R}$, d.h. der uneigentlichen Priori "$N(0, \infty)$", entspricht.
Ist $X|\sigma^2 \sim N(\mu, \sigma^2)$ mit μ bekannt und $\theta = \sigma^2$ unbekannt, folgt aus $I(\sigma^2) = 1/(2\sigma^4)$ (Dahlhaus Bsp. 15.10) die Jeffreys-Priori

$$[\sigma^2] \propto \sigma^{-2}.$$

2.14 Definition (Jeffreys-Priori für Parametervektoren)
Im Fall eines unbekannten Parametervektors $\boldsymbol{\theta}$ ist die Jeffreys-Priori definiert als

$$[\boldsymbol{\theta}] \propto |I(\boldsymbol{\theta})|^{1/2},$$

wobei $|I(\boldsymbol{\theta})|$ die Determinante der Fisher-Informations-Matrix von $\boldsymbol{\theta}$ bezeichnet.
Bemerkung: Für Parametervektoren (d.h. im mehrdimensionalen Fall) ist die Jeffreys-Priori umstritten und wird selten verwendet.

2.15 Beispiel (Fortsetzung von Beispiel 2.13)
Sei nun $X|\boldsymbol{\theta} \sim N(\mu, \sigma^2)$ mit $\boldsymbol{\theta} = (\mu, \sigma^2)'$ unbekannt.
Da $|I(\boldsymbol{\theta})| = 1/(2\sigma^6)$ (Dahlhaus Bsp. 15.10), ist die Jeffreys-Priori in diesem Fall gegeben durch

$$[\mu, \sigma^2] \propto \sigma^{-3}.$$

Wie bereits erwähnt, ist diese Priori allerdings umstritten. Viel häufiger wird die Priori

$$[\mu, \sigma^2] \propto \sigma^{-2}$$

verwendet, die sich bei Annahme der Priori-Unabhängigkeit von μ und σ^2 als Produkt der Jeffreys-Prioris der beiden Parameter (bei jeweils anderem Parameter bekannt) ergibt.

Kapitel 3

Wahl der Schätzer

3.1 Wahl der Punktschätzer

3.1 Definition (Verlustfunktion)
Eine Verlustfunktion ("loss function") $L(a, \theta)$ gibt den "Verlust" an, den man beim Schätzen von θ durch den Wert a erleidet.

3.2 Beispiel (Verlustfunktion)
Folgende Verlustfunktionen (VFs) werden häufig verwendet:

1. Quadratische VF: $L(a, \theta) = (a - \theta)^2$
2. Lineare VF: $L(a, \theta) = |a - \theta|$

3.3 Definition
Der *Bayes-Schätzer* eines Parameters θ bezüglich einer VF $L(a, \theta)$ bei beobachteten Daten $X = x$, ist der Wert von a, der den erwarteten Verlust,

$$E(L(a, \theta)|x) = \int_{\Theta} L(a, \theta)[\theta|x]d\theta,$$

minimiert.

3.4 Satz (Wichtige Bayes-Schätzer)
Aus den VFs in Beispiel 3.2 ergeben sich die folgenden Bayes-Schätzer:

1. Quadratische VF $\Rightarrow$ Posteriori-Erwartungswert
2. Lineare VF $\Rightarrow$ Posteriori-Median

Beweis.

1.

$$\frac{\partial}{\partial a}E(L(a,\theta)|x) = \frac{\partial}{\partial a}\int (a-\theta)^2[\theta|x]d\theta = 2\int (a-\theta)[\theta|x]d\theta \stackrel{!}{=} 0$$
$$\Leftrightarrow a\int [\theta|x]d\theta = \int \theta\,[\theta|x]d\theta$$
$$\Leftrightarrow a = E(\theta|x)$$

2. Es gilt:

$$E(L(a,\theta)|x) = \int |a-\theta|[\theta|x]d\theta$$
$$= \int_{\theta\leq a}(a-\theta)[\theta|x]d\theta + \int_{\theta\geq a}(\theta-a)[\theta|x]d\theta.$$

Und damit:

$$\frac{\partial}{\partial a}E(L(a,\theta)|x) = \int_{\theta\leq a}[\theta|x]d\theta - \int_{\theta\geq a}[\theta|x]d\theta \stackrel{!}{=} 0$$
$$\Leftrightarrow \int_{\theta\leq a}[\theta|x]d\theta = \int_{\theta\geq a}[\theta|x]d\theta.$$

Diese Gleichung gilt wenn a gleich dem Posteriori-Median ist.

□

3.5 Bemerkung

Wie bereits erwähnt, basiert Bayes-Inferenz bezüglich einer unbekannten Größe ausschließlich auf der Posteriori dieser Größe. Auch der Bayes-Schätzer von θ in Definition 3.3 ist eine Funktion der Posteriori von θ.

Ein Vorteil der Bayes-Inferenz ist aber, dass z.B. der Posteriori-Erwartungswert nur *eine* mögliche Zusammenfassung der Posteriori ist. Die Posteriori-Verteilung, die bei der Bayes-Inferenz berechnet wird (oft nur numerisch), enthält sehr viel mehr Informationen als eine solche skalare Zusammenfassung. Andere Funktionen der Posteriori, wie z.B. $P(\theta \in A|X = x) = \int_A [\theta|x]d\theta$, sind häufig von eigentlichem Interesse.

3.2 Wahl der Intervallschätzer

In diesem Abschnitt sei θ skalar.

3.6 Definition (Kredibilitätsregion)

Eine Menge $K \subset \Theta$ mit $\int_K [\theta|x]d\theta = P(\theta \in K|x) = 1-\alpha$ heißt $(1-\alpha)$-Kredibilitätsregion für θ bezüglich der Posteriori $[\theta|x]$. Ist K ein reelles Intervall, so nennt man K auch Kredibilitätsintervall.

3.7 Bemerkung

Analog zu den Konfidenzregionen in der frequentistischen Statistik, ist die Kredibilitätsregion für festes α nicht eindeutig. Als Ausweg kann man (wie für Punktschätzer) eine Verlustfunktion vorgeben, und anschließend die für diese Verlustfunktion optimale Kredibilitätsregion bestimmen.

Einfacher zu bestimmen ist häufig ein "symmetrisches" Kredibilitätsintervall der Form $(q_{\alpha/2}, q_{1-\alpha/2}]$, wobei mit q_γ das γ-Quantil der Posteriori $[\theta|x]$ bezeichnet wird. Zur Bestimmung dieses Intervalls wird also an beiden Enden der Posteriori $(\alpha/2)$ Wahrscheinlichkeit "abgeschnitten".

3.8 Beispiel (Fortsetzung von Beispiel 2.8)

Betrachte Beispiel 2.8 im Fall $p = 1$. Dann ist μ skalar mit Posteriori,

$$\mu|\mathbf{x}_{1:n} \sim N(\bar{x}, \sigma^2/n).$$

Ein $(1-\alpha)$-Kredibilitätsintervall ist daher durch $(q_{\alpha/2}, q_{1-\alpha/2}]$ gegeben, wobei hier q_γ das γ-Quantil der $N(\bar{x}, \sigma^2/n)$ bezeichnet. Das Kredibilitätsintervall kann auch mit Hilfe von $u_{1-\alpha/2}$, des $(1-\alpha/2)$-Quantils der Standardnormalverteilung, geschrieben werden als

$$(\bar{x} - \tfrac{\sigma}{\sqrt{n}} u_{1-\alpha/2}, \bar{x} - \tfrac{\sigma}{\sqrt{n}} u_{1-\alpha/2}].$$

Auf Grund der Verwendung der nicht-informativen Jeffreys-Priori, hat dieses Kredibilitätsintervall die gleiche Form wie das Konfidenzintervall in Bem. 9.1 des Dahlhaus-Skripts (aber es hat eine einfachere Interpretation, s. nachfolgende Bemerkung 3.9).

3.9 Bemerkung (Interpretation von Kredibilitätsregionen)

Die Interpretation einer $(1-\alpha)$-Kredibilitätsregion K nach Beobachtung von Daten $X = x$ ist denkbar einfach: Gegeben die Daten $X = x$, ist θ mit Wahrscheinlichkeit $(1-\alpha)$ in der Region K.

Vergleiche dies mit der Interpretation einer frequentistischen Konfidenzregion $K(x)$: Falls unendlich viele Beobachtungen X_i aus der Verteilung von x (gegeben θ) gezogen würden, und man für jedes X_i eine Konfidenzregion $K(X_i)$ nach der gleichen Formel berechnen würde, wäre das (als fest betrachtete) θ in $(1-\alpha) \cdot 100\%$ der Konfidenzregionen enthalten.

Kapitel 4

Komplexere Modelle

4.1 Bemerkung (Inferenz bei Nuisance-Parametern)
Die Likelihood habe die Form $[x|\theta, \psi]$, wobei θ der Parameter(-vektor) von Interesse ist, und ψ der Nuisance-Parameter(-vektor). Nach Beobachtung von $X = x$ ist also die (marginale) Posteriori von θ gesucht. Diese ist nach Bestimmung der gemeinsame Posteriori von θ und ψ,

$$[\theta, \psi|x] = \frac{[x|\theta, \psi][\theta, \psi]}{[x]},$$

durch "rausintegrieren" von ψ erhältlich:

$$[\theta|x] = \int [\theta, \psi|x] d\psi.$$

Die Behandlung von Nuisance-Parametern in der Bayes-Inferenz ist also kanonisch. Inferenz bezüglich des Parameters θ basiert wieder auf dessen Posteriori, $[\theta|x]$.

4.2 Beispiel (Fortsetzung von Beispiel 2.4)
Wir betrachten wieder die Situation aus Beispiel 2.4 im Fall $p = 1$. Allerdings nehmen wir nun an, dass auch der Nuisance-Parameter σ^2 unbekannt ist. Als gemeinsame Priori für μ und σ^2 nehmen wir die aus Beispiel 2.15 bekannte, nicht-informative Priori,

$$[\mu, \sigma^2] \propto \sigma^{-2}.$$

In Beispiel 3.8 haben wir gesehen, dass

$$\mu|\mathbf{x}_{1:n}, \sigma^2 \sim N(\bar{x}, \sigma^2/n).$$

Aus der Umformung in Beispiel 2.4 geht hervor, dass die gemeinsame Posteriori von μ und σ^2

folgende Form hat:

$$\begin{aligned}[\mu, \sigma^2 | \mathbf{x}_{1:n}] &\propto \sigma^{-2} \cdot \sigma^{-n} \exp\left\{ -\frac{1}{2\sigma^2} \sum_{i=1}^{n} (x_i - \mu)^2 \right\} \\ &= \sigma^{-n-2} \exp\left\{ -\frac{1}{2\sigma^2} \left(\sum_{i=1}^{n} (x_i - \bar{x})^2 + n(\bar{x} - \mu)^2 \right) \right\} \\ &= \sigma^{-n-2} \exp\left\{ -\frac{1}{2\sigma^2} \big((n-1)s^2 + n(\bar{x} - \mu)^2 \big) \right\},\end{aligned}$$

wobei $s^2 := \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2$ der übliche Varianzschätzer ist.
Die Posteriori von μ ergibt sich nun durch Integration über σ^2, mit Hilfe der Substitution

$$z = A/(2\sigma^2) \quad \Leftrightarrow \quad \sigma^2 = z^{-1} A/2 \quad \Rightarrow \quad d\sigma^2 = -z^{-2} A/2 \, dz,$$

wobei $A := (n-1)s^2 + n(\bar{x} - \mu)^2$:

$$\begin{aligned}[\mu | \mathbf{x}_{1:n}] &= \int_0^\infty [\mu, \sigma^2 | \mathbf{x}_{1:n}] d\sigma^2 \\ &\propto \int_0^\infty (\sigma^2)^{-n/2-1} \exp\{-A/(2\sigma^2)\} d\sigma^2 \\ &\propto \int_\infty^0 (z^{-1} A)^{-n/2-1} e^{-z} (-z^{-2} A) dz \\ &= A^{-n/2} \int_0^\infty z^{n/2-1} e^{-z} dz \\ &= \big((n-1)s^2 + n(\bar{x} - \mu)^2 \big)^{-n/2} \, \Gamma(n/2) \\ &\propto \left(1 + \frac{n(\bar{x} - \mu)^2}{(n-1)s^2} \right)^{-n/2}.\end{aligned}$$

Dies ist der Kern der Dichte von μ wenn $\mu := (s/\sqrt{n})Y + \bar{x}$ und $Y \sim t_{n-1}$ (siehe Prop. 13.6 und Bem. 6.15 im Dahlhaus-Skript), das heißt

$$\frac{\mu - \bar{x}}{s/\sqrt{n}} \, | \mathbf{x}_{1:n} \sim t_{n-1}.$$

Vergleiche dies mit Satz 13.4 (ii) im Dahlhaus-Skript:

$$\frac{\bar{X} - \mu}{S/\sqrt{n}} \, | \mu, \sigma^2 \sim t_{n-1}.$$

4.3 Bemerkung (Prognose)

Oft sind die Parameter selbst gar nicht von Interesse, d.h. alle Parameter sind Nuisance-Parameter. Stattdessen ist man an Inferenz bezüglich des beobachteten Prozesses selbst interessiert. Hier betrachten wir als Beispiel die Prognose von Y basierend auf Daten $X = x$, wobei bedingt unabhängig sind gegeben θ.

Vor Beobachtung von x ist die *priori-prädiktive* Verteilung von Y von Interesse:

$$[y] = \int [y|\theta][\theta]d\theta.$$

Nach Beobachtung von $X = x$ betrachtet man die *posteriori-prädiktive* Verteilung von Y gegeben x:

$$[y|x] = \int [y, \theta|x]d\theta = \int [y|\theta, x][\theta|x]d\theta = \int [y|\theta][\theta|x]d\theta,$$

auf Grund der bedingten Unabhängigkeit von X und Y gegeben θ.
Man beachte dass diese posteriori-prädiktive Verteilung ungleich der (in der frequentistischen Inferenz häufig verwendeten) "plug-in"-Prognose $[Y|\hat{\theta}]$ ist. Bei der plug-in-Prognose wird die Posteriori-Verteilung $[\theta|x]$ durch eine Punkt-Verteilung im Schätzer $\hat{\theta}$ ersetzt. Die Unsicherheit in der Schätzung von θ wird damit ignoriert, und die resultierenden Prognose-Konfidenzintervalle sind zu schmal.

4.4 Bemerkung (Prädiktive Verteilungen in konjugierten Verteilungsklassen)
Da bei der Verwendung einer konjugierten Priori die Posteriori $[\theta|x]$ und Priori $[\theta]$ zur selben Verteilungsklasse gehören, gibt es dann technisch gesehen keinen Unterschied zwischen der Berechnung der priori-prädiktiven Verteilung,

$$[y] = \int [y|\theta][\theta]d\theta, \tag{4.1}$$

und der posteriori-prädiktiven Verteilung,

$$[y|x] = \int [y|\theta][\theta|x]d\theta.$$

Daher reicht es die Berechnung der priori-prädiktiven Verteilung zu betrachten.
Bei Verwendung einer konjugierten Priori kann die Integration in (4.1) vermieden werden. Aus dem Satz von Bayes,

$$[\theta|y] = \frac{[y|\theta][\theta]}{[y]},$$

folgt dass

$$[y] = \frac{[y|\theta][\theta]}{[\theta|y]}.$$

Hier sind nun die Normalisierungskonstanten wichtig. Diese sind in konjugierten Verteilungsklassen aber bekannt.

4.5 Beispiel (Prognose im linearen Modell)
In Beispiel 2.9 haben wir Inferenz bezüglich des Parametervektors $\boldsymbol{\beta}$ im Fall von $rangX = k$ betrieben. Nun wollen wir für das gleiche Modell Inferenz bezüglich eines neuen Response-Vektors $\mathbf{Y}_N$ basierend auf einer zugehörigen Designmatrix X_N (und basierend auf den alten

Beobachtungen $\mathbf{Y} = \mathbf{y}$) beschreiben. Da die Designmatrizen (und σ^2) wie bisher als fest betrachtet werden, suchen wir also die posteriori-prädiktive Verteilung $[\mathbf{Y}_N|\mathbf{y}]$.
Auf Grund der bedingten Unabhängigkeit von $\mathbf{Y}_N$ und $\mathbf{Y}$ gegeben $\boldsymbol{\beta}$, gilt $[\mathbf{Y}_N|\boldsymbol{\beta}, \mathbf{y}] = [\mathbf{Y}_N|\boldsymbol{\beta}]$, und damit

$$[\boldsymbol{\beta}|\mathbf{y}_N, \mathbf{y}] = \frac{[\mathbf{y}_N|\boldsymbol{\beta}][\boldsymbol{\beta}|\mathbf{y}]}{[\mathbf{y}_N|\mathbf{y}]}.$$

Analog zu Bemerkung 4.4 ist die posteriori-prädiktive Verteilung also gegeben durch

$$[\mathbf{y}_N|\mathbf{y}] = \frac{[\mathbf{y}_N|\boldsymbol{\beta}][\boldsymbol{\beta}|\mathbf{y}]}{[\boldsymbol{\beta}|\mathbf{y}_N, \mathbf{y}]},$$

wobei $\mathbf{y}_N|\boldsymbol{\beta} \sim N(X_N\boldsymbol{\beta}, \sigma^2 I)$ und $\boldsymbol{\beta}|\mathbf{y} \sim N_k(\hat{\boldsymbol{\beta}}, \sigma^2(X'X)^{-1})$ (s. Beispiel 2.9). Analog dazu lässt sich zeigen, dass

$$\boldsymbol{\beta}|\mathbf{y}_N, \mathbf{y} \sim N\left((X'X + X_N'X_N)^{-1}(X'\mathbf{y} + X_N'\mathbf{y}_N), \sigma^2(X'X + X_N'X_N)^{-1}\right).$$

Ebenso ist aus dem obigen Ausdruck für $[\mathbf{y}_N|\mathbf{y}]$ leicht zu sehen, dass damit $[\mathbf{Y}_N|\mathbf{Y}]$ auch eine Normalverteilung ist. Zur Bestimmung der posteriori-prädiktiven Verteilung müssen wir also lediglich Erwartungswert und Varianz dieser Verteilung ausrechnen. Mit Satz 16.7 (i) im Dahlhaus-Skript erhalten wir

$$E(\mathbf{Y}_N|\mathbf{y}) = E(E(\mathbf{Y}_N|\mathbf{y}, \boldsymbol{\beta})|\mathbf{y}) = E(X_N\boldsymbol{\beta}|\mathbf{y}) = X_N\hat{\boldsymbol{\beta}},$$

und mit Teil (v) des gleichen Satzes gilt

$$\begin{aligned} Var(\mathbf{Y}_N|\mathbf{y}) &= E(Var(\mathbf{Y}_N|\mathbf{y}, \boldsymbol{\beta})|\mathbf{y}) + Var(E(\mathbf{Y}_N|\mathbf{y}, \boldsymbol{\beta})|\mathbf{y}) \\ &= E(\sigma^2 I|\mathbf{y}) + Var(X_N\boldsymbol{\beta}|\mathbf{y}) \\ &= \sigma^2 I + X_N\sigma^2(X'X)^{-1}X_N. \end{aligned}$$

Würde man (wie in der klassischen Statistik oft üblich) den Parameter (in diesem Fall $\boldsymbol{\beta}$) als fix betrachten und eine plug-in-Prognose vornehmen, so würde die Varianz der Prognose nur aus dem ersten Teil bestehen, da der zweite, auf der Unsicherheit in der Schätzung von $\boldsymbol{\beta}$ beruhende Teil ignoriert würde.
Bemerkung: Typischerweise ist σ^2 auch unbekannt, und man nimmt eine Priori für den Parameter an. Die posteriori-prädiktive Verteilung von $\mathbf{Y}_N$ wird dann oft mit Hilfe eines Computers berechnet (s. Kapitel 5).

4.6 Bemerkung (Hierarchische Modelle)
Bis jetzt haben wir bayesianische Modelle mit zwei "Ebenen" betrachtet: Die Likelihood, $[x|\theta]$, und die Priori, $[\theta]$. In einem bayesianischen hierarchischen Modell (BHM) sind theoretisch aber beliebig viele Ebenen möglich, d.h. die Parameter, ψ, der Priori können selbst wieder Priori-Verteilungen (sogenannte Hyperprioris) besitzen. Die Parameter der Hyperpriori ("Hyperparameter") können dann selbst wieder als zufällig modelliert werden, usw.

Die gemeinsame Posteriori aller Parameter ist im Fall eines BHM mit drei Ebenen gegeben durch:

$$[\theta, \psi|x] \propto [x|\theta, \psi][\theta, \psi] = [x|\theta][\theta|\psi][\psi].$$

Da die Parameter der Verteilung(en) auf einer Ebene oft eine gemeinsame Verteilung in der Ebene "darunter" besitzen, haben selbst Modelle mit mehr Parametern als Beobachtungen oft genug Struktur für vernünftige Inferenz.
Ein großer Vorteil hierarchischer Modelle besteht darin, dass aus relativ simplen (oft konjugierten) bedingten Verteilungen auf allen Ebenen, insgesamt ein sehr komplexes, flexibles Modell entstehen kann. Numerische Verfahren wie der Gibbs-Sampler (siehe Kapitel 5) ermöglichen einfache (Computer-gestützte) Inferenz.

4.7 Beispiel (Hierarchische Modelle)

Eine Stadt will die Mathe-Kenntnisse der Neuntklässler an den örtlichen Gymnasien testen. Dazu wird ein standardisierter Test entwickelt, dessen Ergebnisse annähernd normalverteilt sind. Das Testergebnis des k-ten Schülers der j-ten neunten Klasse im i-ten Gymnasium wird bezeichnet mit

$$T_{ijk}; \quad k = 1, \ldots, K_{ij};\ j = 1, \ldots, J_i;\ i = 1, \ldots, I.$$

Ein BHM kann nun z.B. durch die folgenden Annahmen bestimmt werden:

$$\begin{aligned} T_{ijk}|\theta_{ij}, \sigma_1^2 &\overset{iid}{\sim} N(\theta_{ij}, \sigma_1^2); \quad k = 1, \ldots, K_{ij}, \\ \theta_{ij}|\psi_j, \sigma_2^2 &\overset{iid}{\sim} N(\psi_j, \sigma_2^2); \quad j = 1, \ldots, J_i, \\ \psi_j|\mu, \sigma_3^2 &\overset{iid}{\sim} N(\mu, \sigma_3^2); \quad i = 1, \ldots, I, \\ \mu &\sim N(\nu, \lambda), \end{aligned}$$

wobei die θ_{ij} die Klassen-spezifischen Mittelwerte und ψ_j die Schul-spezifischen Mittelwerte sind. Die Parameter ν und λ sind fest. Die Varianz-Parameter werden als Nuisance-Parameter betrachtet, mit $[\sigma_1^2, \sigma_2^2, \sigma_3^2] = [\sigma_1^2][\sigma_2^2][\sigma_3^2]$, wobei $[\sigma_l^2]$ beliebige (idealerweise informative) Prioris sind.
Nach Beobachtung aller Testergebnisse $\mathbf{t}$, ist die Posteriori von Interesse also

$$[\boldsymbol{\theta}, \boldsymbol{\psi}, \mu|\mathbf{t}] \propto [\mu] \iiint \Big(\prod_{i=1}^{I} [\psi_j|\mu, \sigma_3^2] \prod_{j=1}^{J_i} [\theta_{ij}|\psi_j, \sigma_2^2] \prod_{k=1}^{K_{ij}} [T_{ijk}|\theta_{ij}, \sigma_1^2] \Big) \Big(\prod_{l=1}^{3} [\sigma_l^2] \Big) d\sigma_1^2 d\sigma_2^2 d\sigma_3^2.$$

Kapitel 5

Numerische Verfahren

Die explizite Berechnung der Normalisierungskonstante der Posteriori ist, außer für einfache Modelle mit konjugierten Prioris, oft nicht möglich. Bei einer kleinen Anzahl von unbekannten Parametern kann die nötige Integration numerisch durchgeführt werden.
Ist die Anzahl der unbekannten Parameter allerdings hoch, ist numerische Integration nicht mehr präzise genug bzw. zu computerintensiv. In solchen Fällen verwendet man normalerweise sogenannte Monte-Carlo-Methoden. Dabei werden Zufallszahlen (bzw. -vektoren) $\theta^{(1)}, \ldots, \theta^{(M)}$ aus der Posteriori erzeugt. Anschließend können beliebige Zusammenfassungen der Posteriori durch Zusammenfassungen der Werte $\theta^{(1)}, \ldots, \theta^{(M)}$ geschätzt werden, und zwar beliebig genau (da M theoretisch beliebig großgewählt werden kann). So gilt nach dem Gesetz der großen Zahlen,

$$E(g(\theta)|x) \approx \frac{1}{M} \sum_{m=1}^{M} g(\theta^{(m)}),$$

d.h. z.B.

$$E(\theta|x) \approx \frac{1}{M} \sum_{m=1}^{M} \theta^{(m)}$$

$$Var(\theta|x) = E(\theta^2|x) - (E(\theta|x))^2 \approx \frac{1}{M} \sum_{m=1}^{M} (\theta^{(m)})^2 - (\frac{1}{M} \sum_{m=1}^{M} \theta^{(m)})^2$$

$$P(\theta \in A|x) = E(I(\theta \in A)|x) \approx \frac{1}{M} \sum_{m=1}^{M} I(\theta^{(m)} \in A).$$

Ebenso können Quantile der Posteriori (z.B. der Posteriori-Median) aus $\theta^{(1)}, \ldots, \theta^{(M)}$ geschätzt werden.
Das wichtigste Verfahren zur Erzeugung von Zufallszahlen aus der Posteriori ist die sogenannte Markov chain Monte Carlo (MCMC) Methode. Hierbei wird eine Markov-Kette (s. weiterführende Vorlesungen) simuliert, die gegen die Posteriori konvergiert. Nachdem die Konvergenz erfolgt ist, können die so erzeugten Zufallszahlen $\theta^{(1)}, \ldots, \theta^{(M)}$ wie oben beschrieben zur Inferenz benutzt werden.

Den meisten solchen MCMC-Methoden liegen der Metropolis-Hastings-Algorithmus und der Gibbs Sampler zugrunde. Für weitere Informationen siehe z.B. Held (2008, Kap. 6).